Applied Deep Learning

Towards Conversational AI

June 2nd, 2020 http://adl.miulab.tw

國立臺灣大學
National Taiwan University

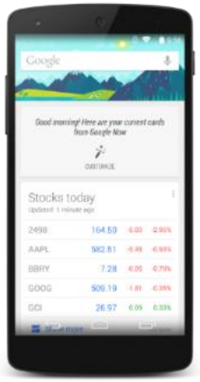
Iron Man (2008)

What can machines achieve now or in the future?

3

Language Empowering Intelligent Assistants

Apple Siri (2011)



Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

Amazon Alexa/Echo (2014)

Google Home (2016)

Apple HomePod (2017)

Facebook Portal (2019)

Why Natural Language?

- Global Digital Statistics (2018 January)

Total Population
7.59B

Internet Users
4.02B

Active Social Media Users
3.20B

Unique Mobile Users
5.14B

Active Mobile Social Users
2.96B

The more **natural** and **convenient** input of devices evolves towards speech.

5 Why and When We Need?

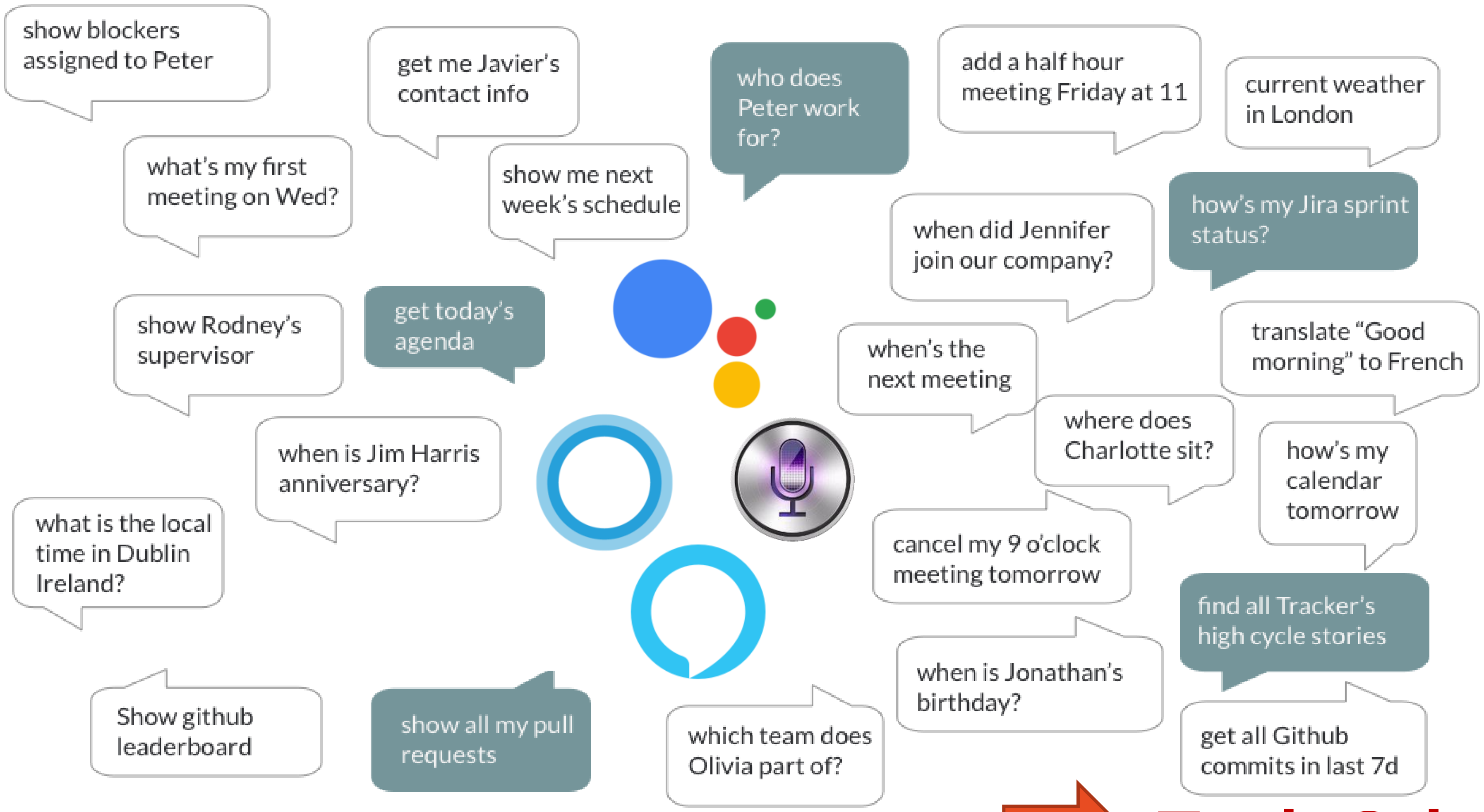
"I want to chat"	Turing Test (talk like a human)	Social Chit-Chat
"I have a question"	Information consumption	Task-Oriented Dialogues
"I need to get this done"	Task completion	
"What should I do?"	Decision support	

- *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?*

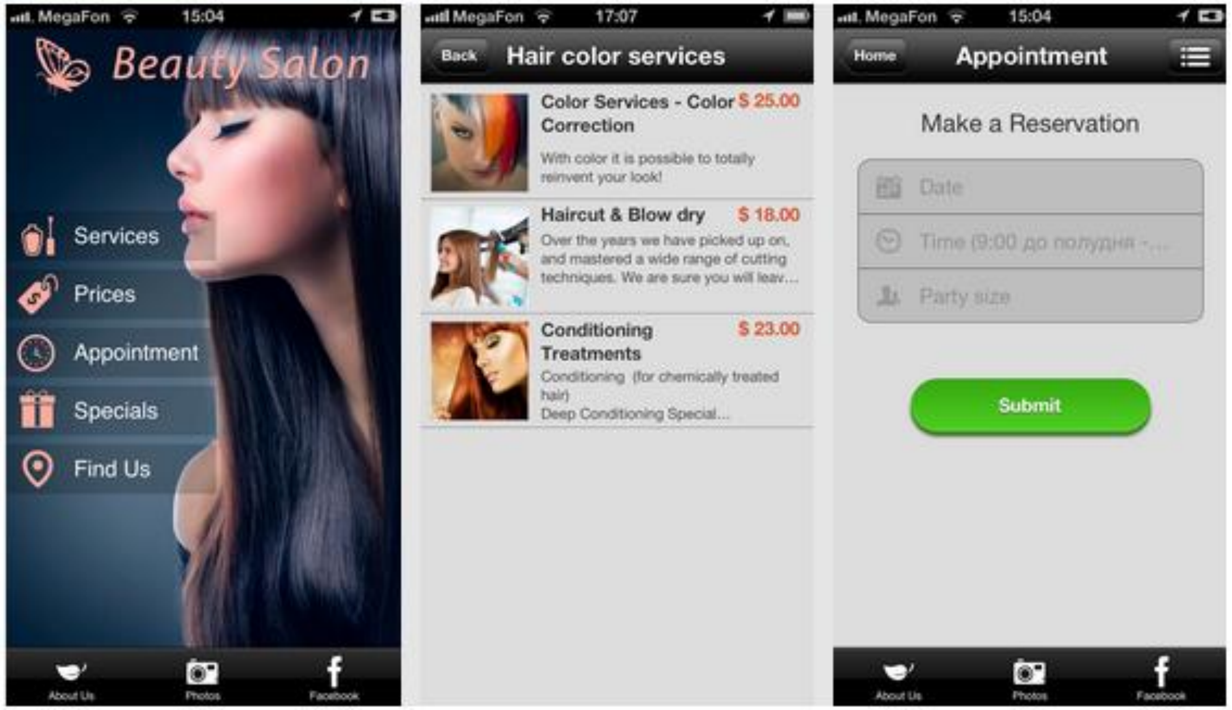
Intelligent Assistants

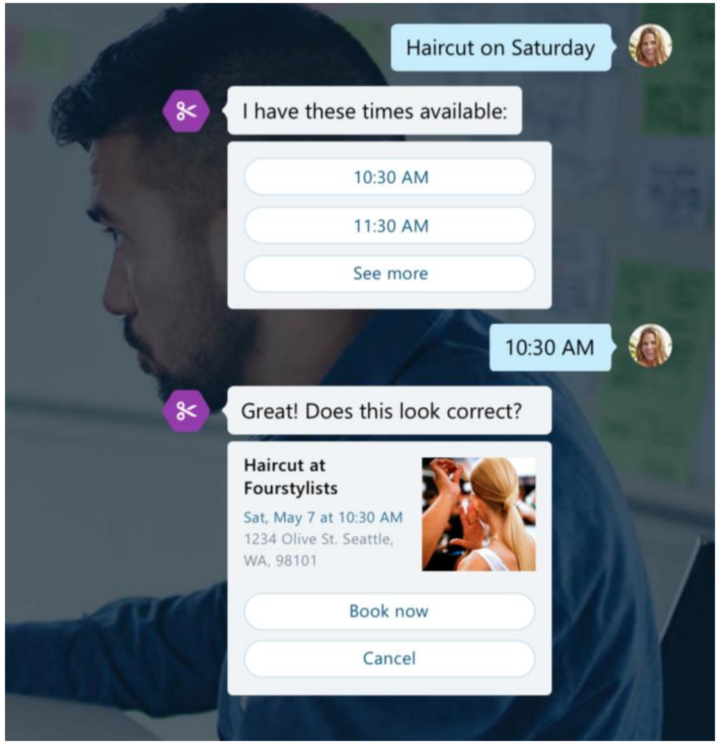


Task-Oriented

App → Bot

- A bot is responsible for a "single" domain, similar to an app





Users can initiate dialogues instead of following the GUI design

Two Branches of Conversational AI

Chit-Chat

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

Task-Oriented

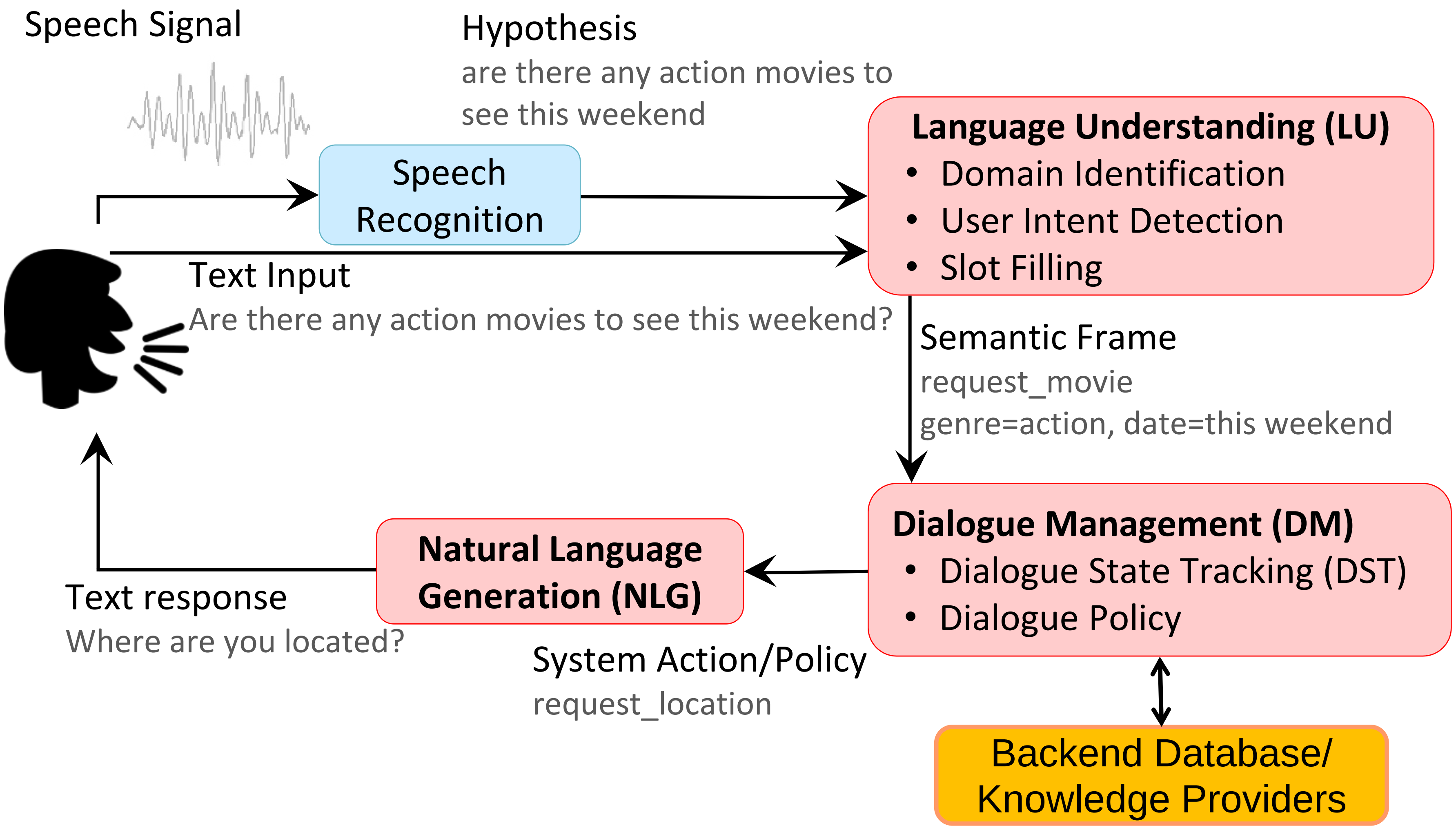
- Single-domain, system-initiative
- Multi-domain, contextual, mixed-initiative
- End-to-end learning, massively multi-domain

Task-Oriented Dialogues

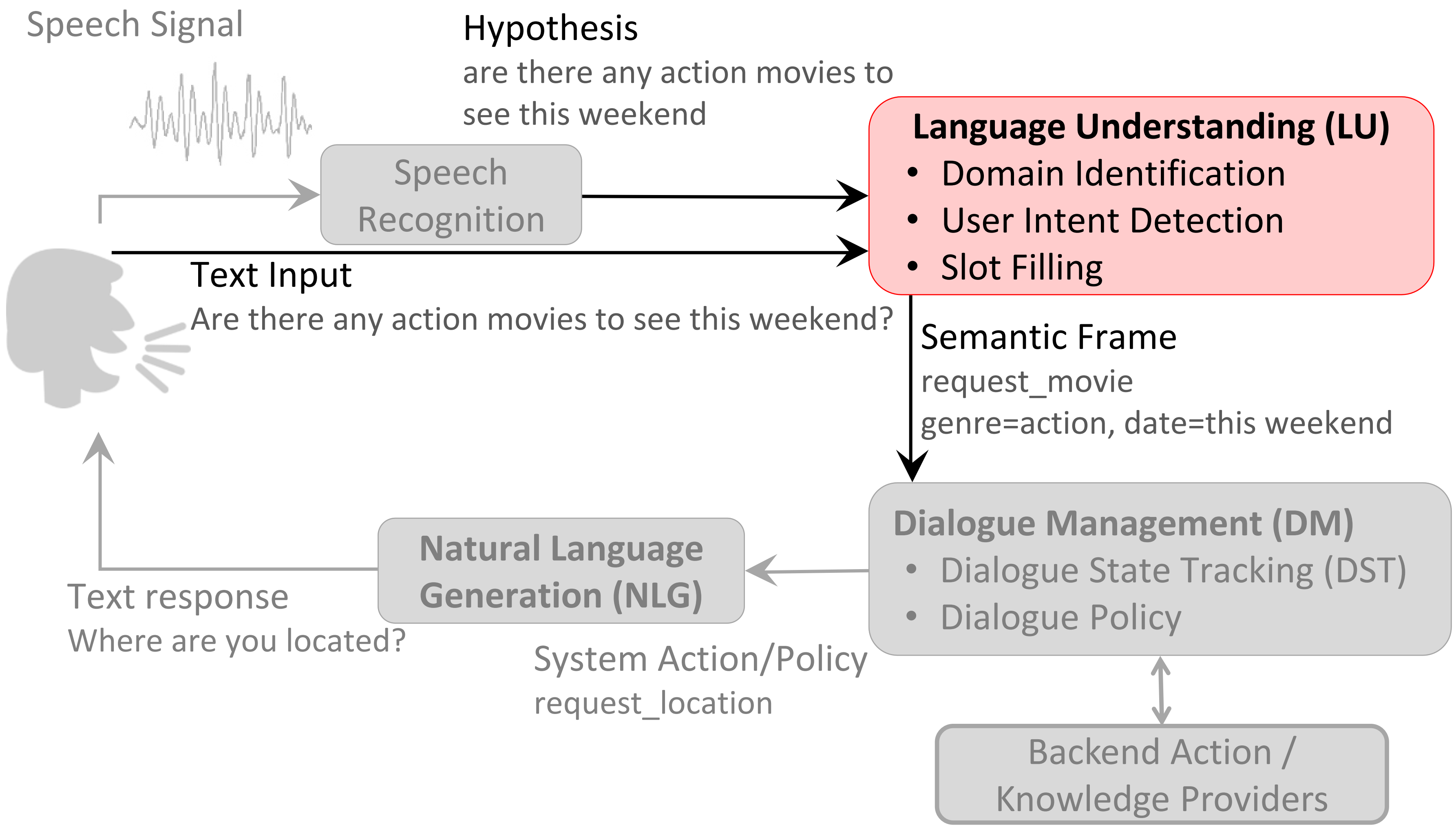
JARVIS – Iron Man's Personal Assistant

Baymax – Personal Healthcare Companion

Task-Oriented Dialogue Systems (Young, 2000)

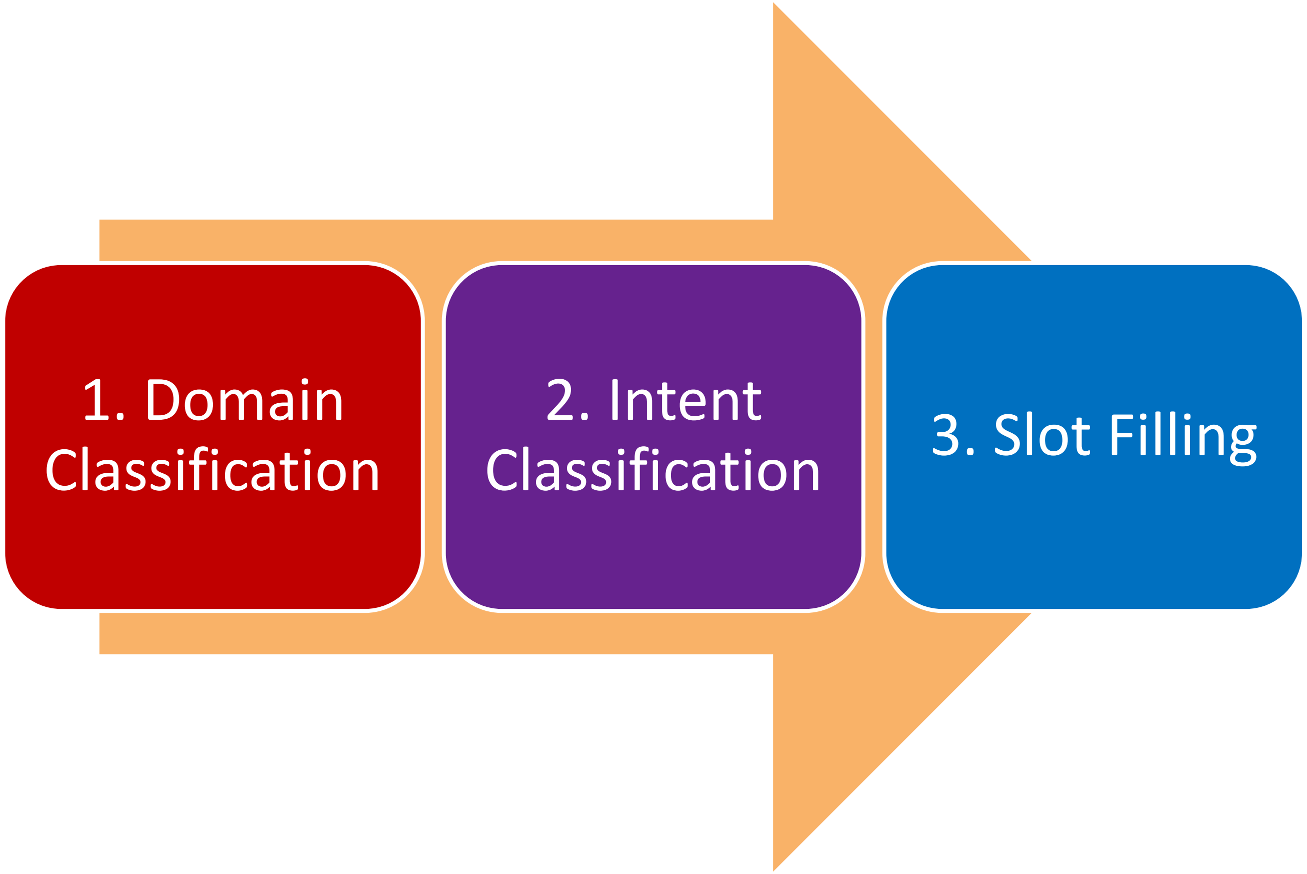


Task-Oriented Dialogue Systems (Young, 2000)



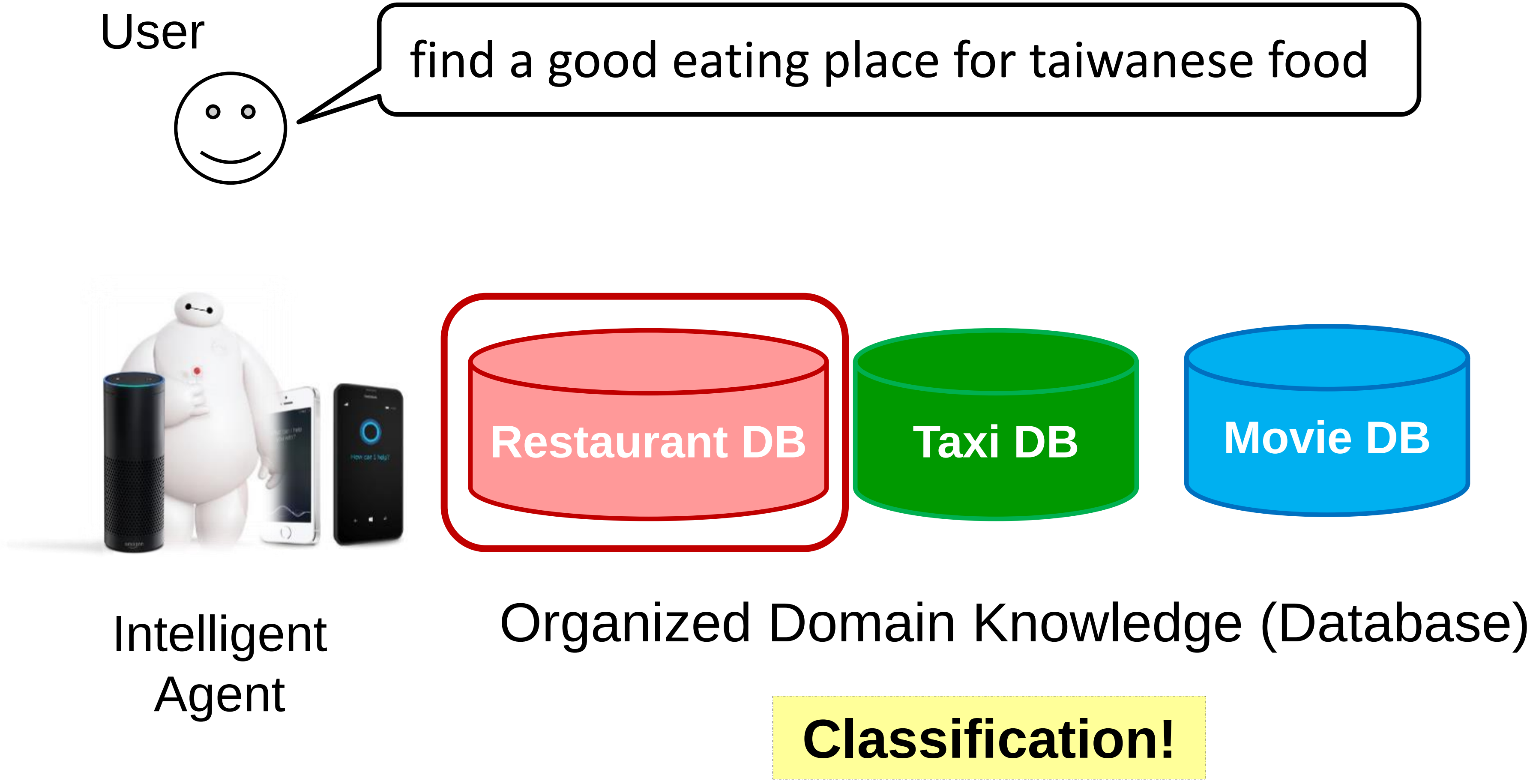
Language Understanding (LU)

- Pipelined



1. Domain Identification

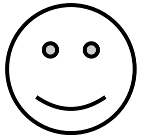
Requires Predefined Domain Ontology



2. Intent Detection

Requires Predefined Schema

User



find a good eating place for taiwanese food

Intelligent
Agent

Restaurant DB

FIND_RESTAURANT
FIND_PRICE
FIND_TYPE
:

Classification!

3. Slot Filling

Requires Predefined Schema

User

O O B-rating O O O B-type O

find a good eating place for taiwanese food

Restaurant	Rating	Type
Rest 1	good	Taiwanese
Rest 2	bad	Thai
:	:	:

Restaurant DB

Intelligent Agent

FIND_RESTAURANT
rating="good"
type="taiwanese"

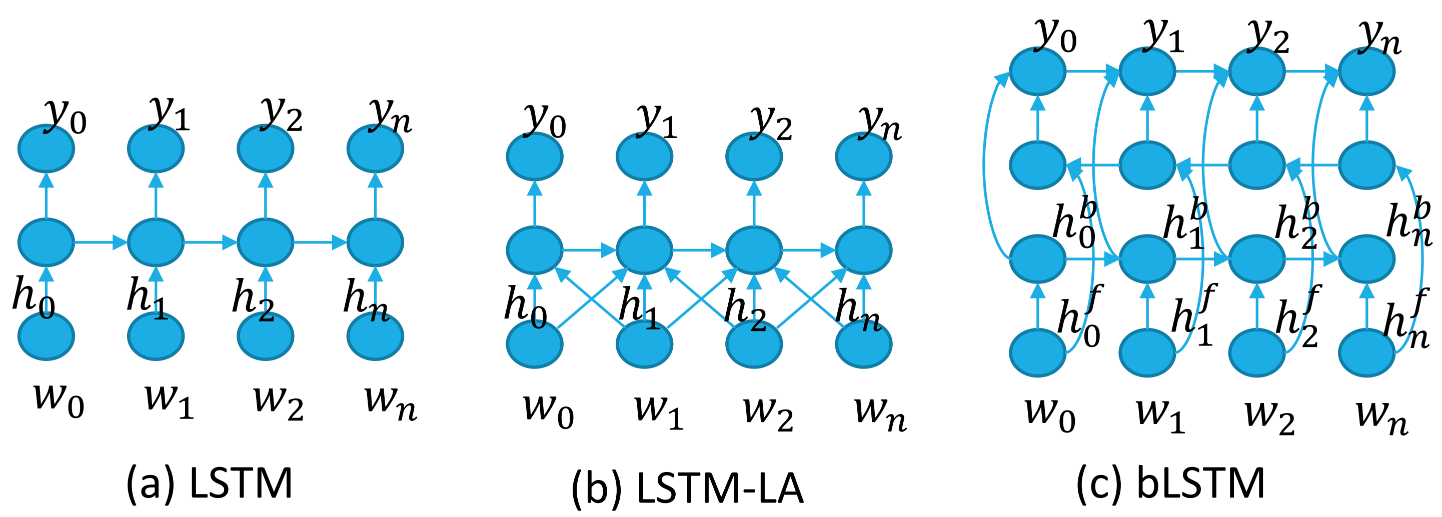
Semantic Frame

SELECT restaurant {
rest.rating="good"
rest.type="taiwanese"
}

Sequence Labeling

Slot Tagging (Yao et al, 2013; Mesnil et al, 2015)

- Variations:
  a. RNNs with LSTM cells
  b. Input, sliding window of n-grams
  c. Bi-directional LSTMs



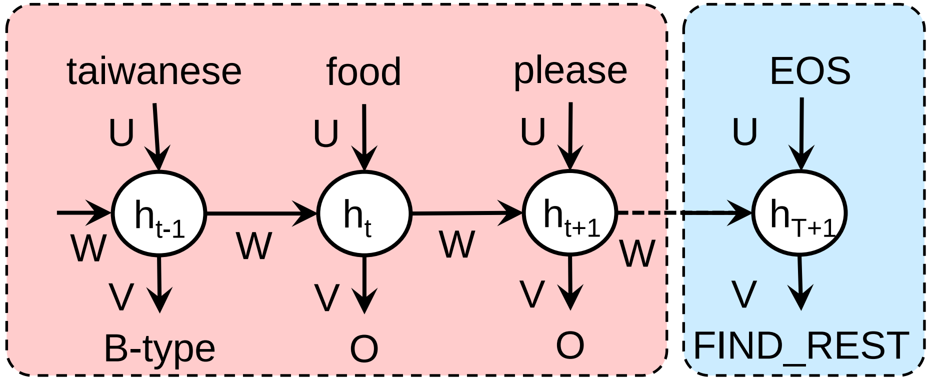
Joint Semantic Frame Parsing

Sequence-based (Hakkani-Tur et al., 2016)

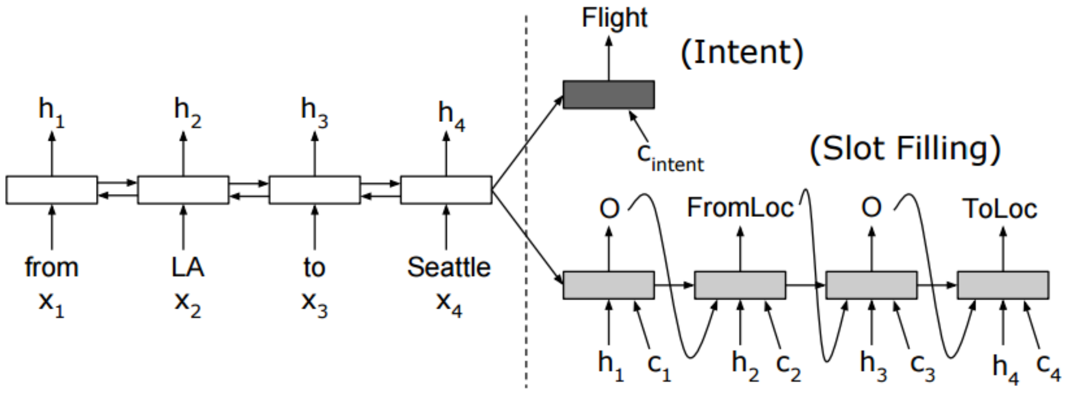
- Slot filling and intent prediction in the same output sequence

Parallel (Liu and Lane, 2016)

- Intent prediction and slot filling are performed in two branches



Slot Filling Intent Prediction



Contextual Language Understanding

- User utterances are highly ambiguous in isolation

Restaurant Booking

Book a table for 10 people tonight.

Which restaurant would you like to book a table for?

Cascal, for 6.

?

#people time

End-to-End Memory Networks (Sukhbaatar et al, 2015)

U: "i d like to purchase tickets to see deepwater horizon"

S: "for which theatre"

U: "angelika"

S: "you want them for angelika theatre?"

U: "yes angelika"

S: "how many tickets would you like ?"

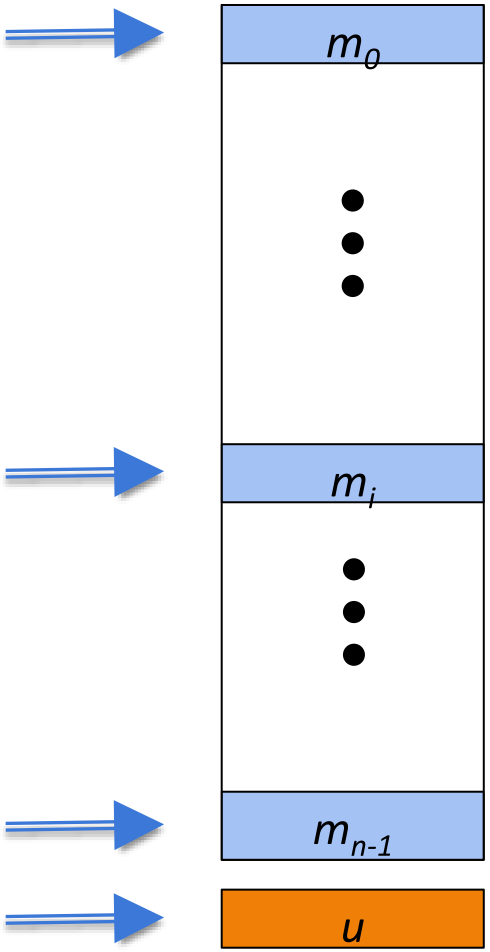
U: "3 tickets for saturday"

S: "What time would you like ?"

U: "Any time on saturday is fine"

S: "okay , there is 4:10 pm , 5:40 pm and 9:20 pm"

U: "Let's do 5:40"



E2E MemNN for Contextual LU (Chen et al., 2016)

U: *"i d like to purchase tickets to see deepwater horizon"* 0.69

S: *"for which theatre"*

U: *"angelika"*

S: *"you want them for angelika theatre?"*

U: *"yes angelika"*

S: *"how many tickets would you like ?"* 0.13

U: *"3 tickets for saturday"*

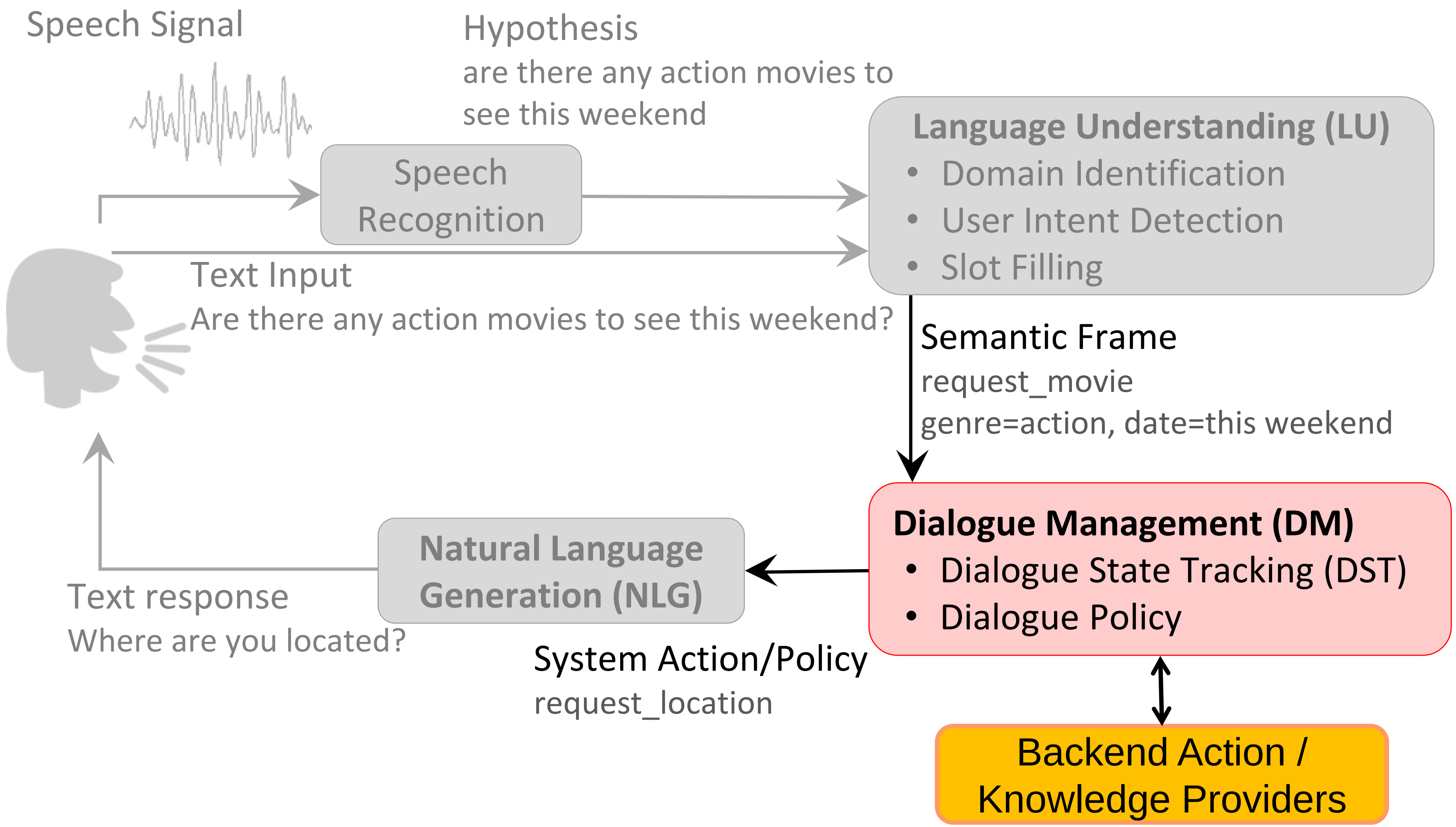
S: *"What time would you like ?"*

U: *"Any time on saturday is fine"*

S: *"okay , there is 4:10 pm , 5:40 pm and 9:20 pm"* 0.16

U: "Let's do 5:40"

Task-Oriented Dialogue Systems (Young, 2000)



Dialogue State Tracking

Hello, how may I help you?

I’m looking for a Thai restaurant. — request (restaurant; foodtype=Thai)

What part of town do you have in mind?

Something in the centre. — inform (area=centre)

Bangkok city is a nice place, it is in the centre of town and it serves Thai food.

What’s the address? — request (address)

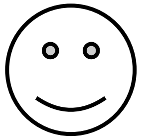
Bangkok city is a nice place, their address is 24 Green street.

Thank you, bye. — bye ()

Dialogue State Tracking

Requires Hand-Crafted States

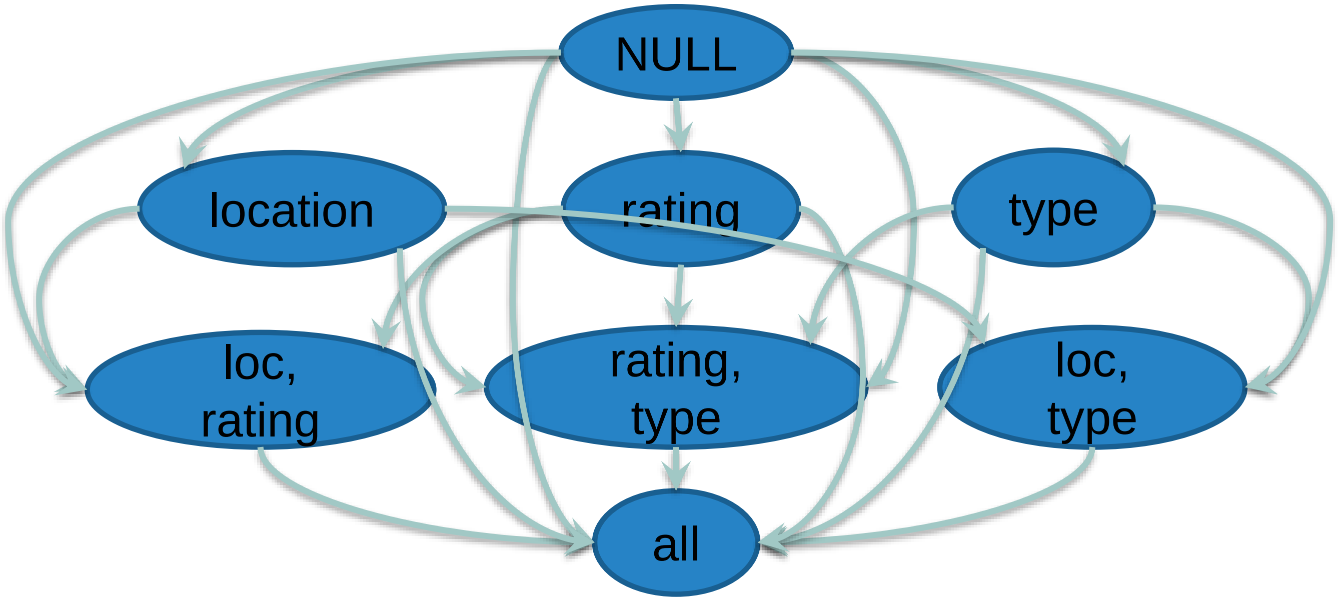
User



find a good eating place for taiwanese food

i want it near to my office

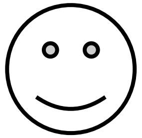
Intelligent Agent



Dialogue State Tracking

Requires Hand-Crafted States

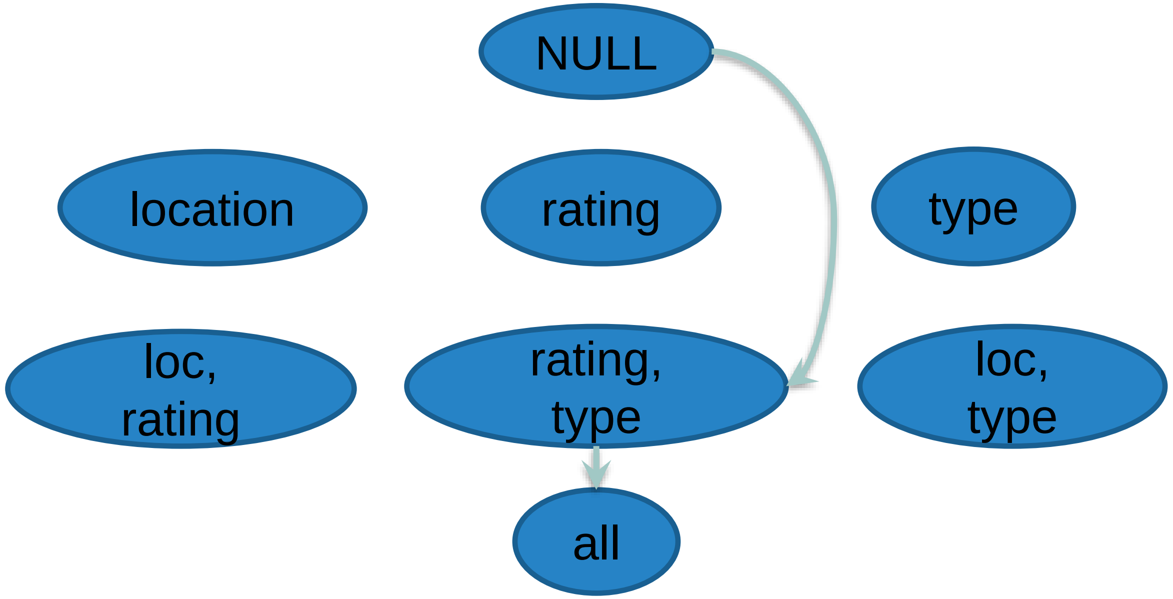
User



find a good eating place for taiwanese food

i want it near to my office

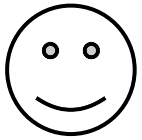
Intelligent Agent



Dialogue State Tracking

Handling Errors and Confidence

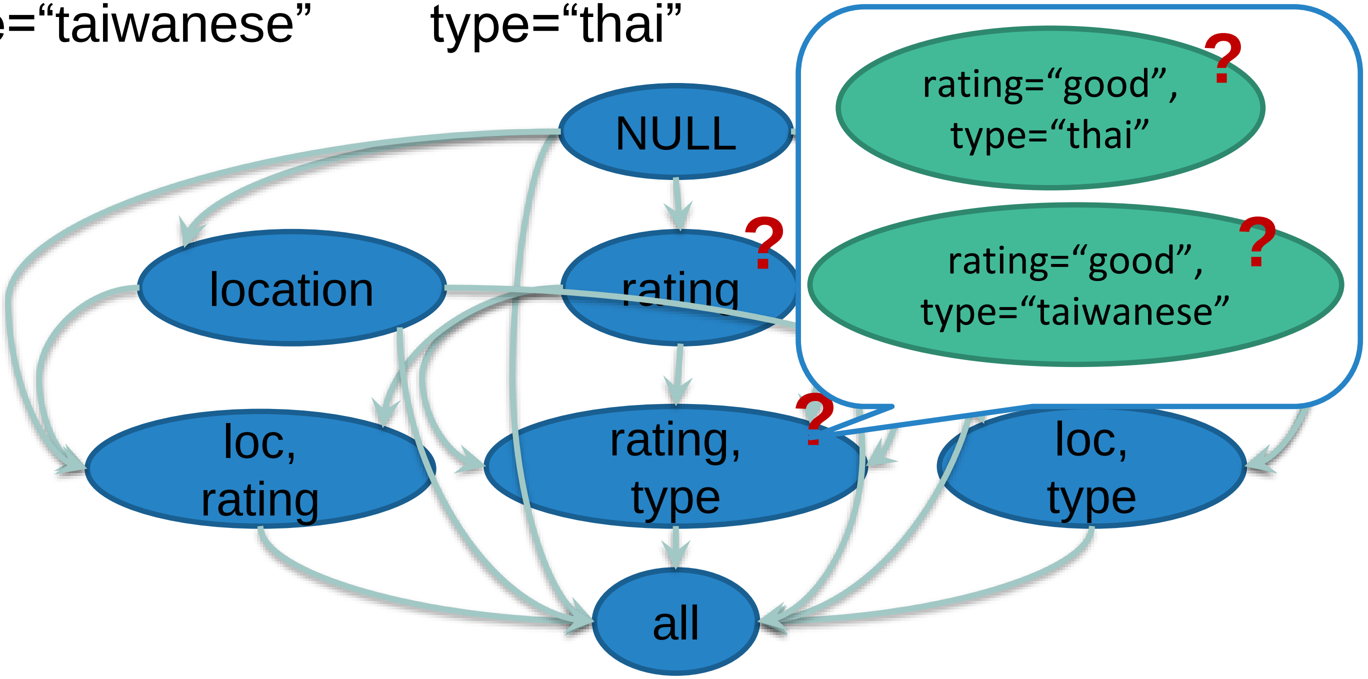
User



find a good eating place for taixxxx food

FIND_RESTAURANT	FIND_RESTAURANT	FIND_RESTAURANT
rating="good"	rating="good"	rating="good"
type="taiwanese"	type="thai"	

Intelligent Agent



Dialogue State Tracking (DST)

- Maintain a probabilistic distribution instead of a 1-best prediction for better robustness to SLU errors or ambiguous input

Slot	Value
# people	5 (0.5)
time	5 (0.5)

Slot	Value
# people	3 (0.8)
time	5 (0.8)

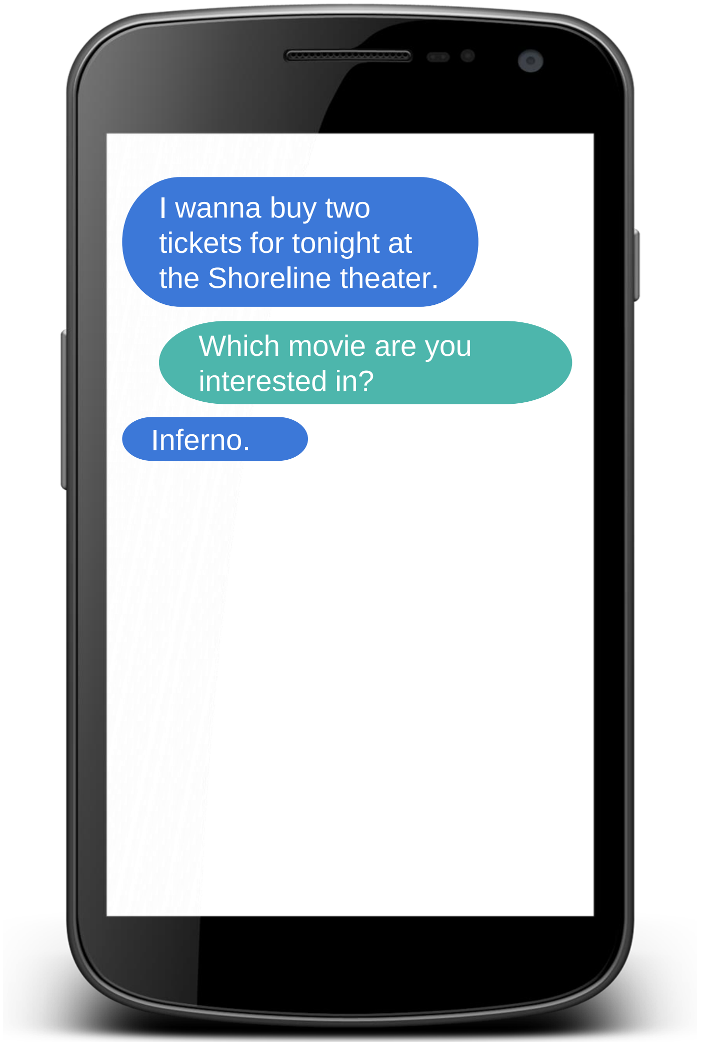
Multi-Domain Dialogue State Tracking

- A full representation of the system's belief of the user's goal at any point during the dialogue
- Used for making API calls

Movies

Date	11/15/17			
Time	6 pm	7 pm	8 pm	9 pm
#People	2			
Theater	Century 16 Shoreline			
Movie	Inferno			

Less Likely ... More Likely



Multi-Domain Dialogue State Tracking

- A full representation of the system's belief of the user's goal at any point during the dialogue
- Used for making API calls

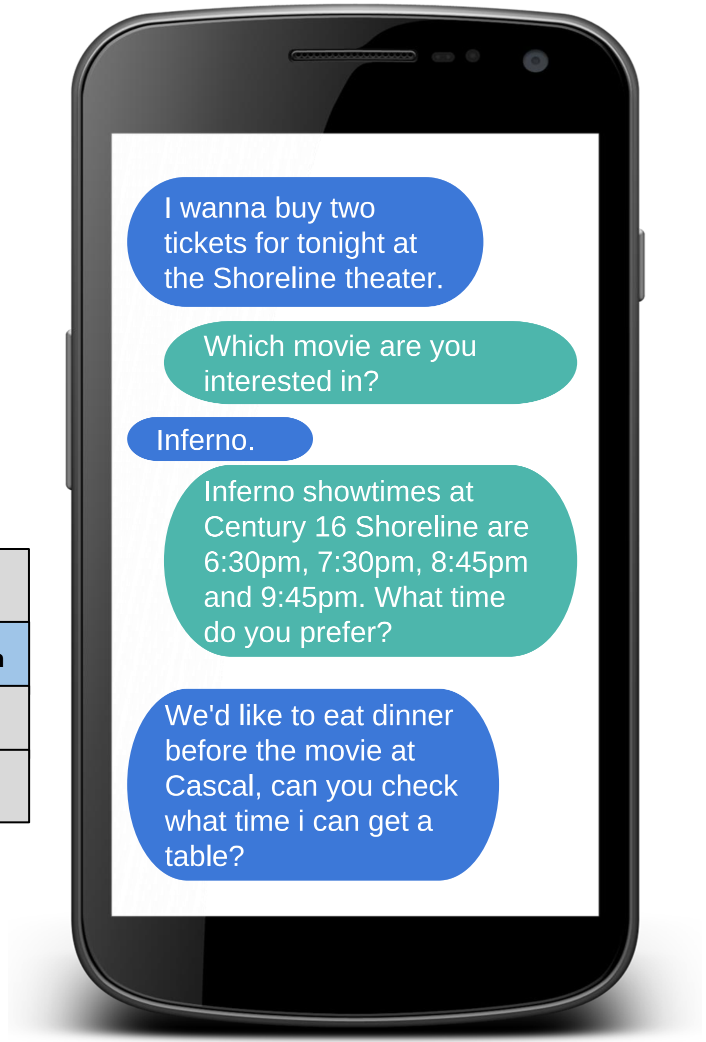
Movies

Date	11/15/17			
Time	6:30 pm	7:30 pm	8:45 pm	9:45 pm
#People	2			
Theater	Century 16 Shoreline			
Movie	Inferno			

Restaurants

Date	11/15/17		
Time	6:00 pm	6:30 pm	7:00 pm
Restaurant	Cascal		
#People	2		

Less Likely — More Likely



Multi-Domain Dialogue State Tracking

- A full representation of the system's belief of the user's goal at any point during the dialogue
- Used for making API calls

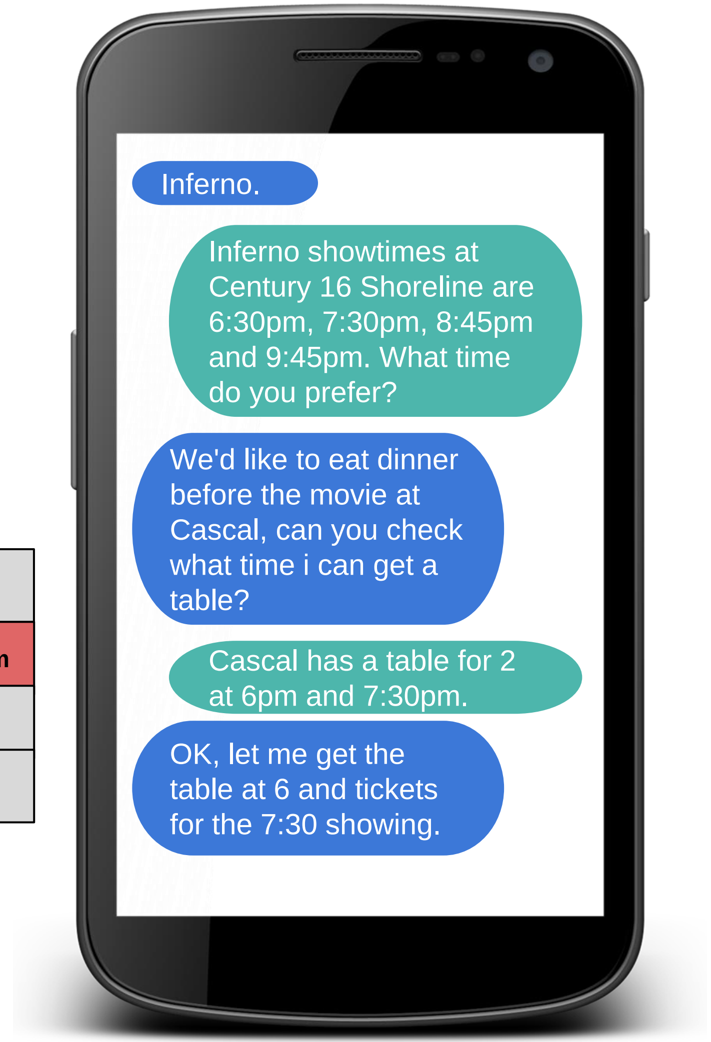
Movies

Date	11/15/17			
Time	6:30 pm	7:30 pm	8:45 pm	9:45 pm
#People	2			
Theater	Century 16 Shoreline			
Movie	Inferno			

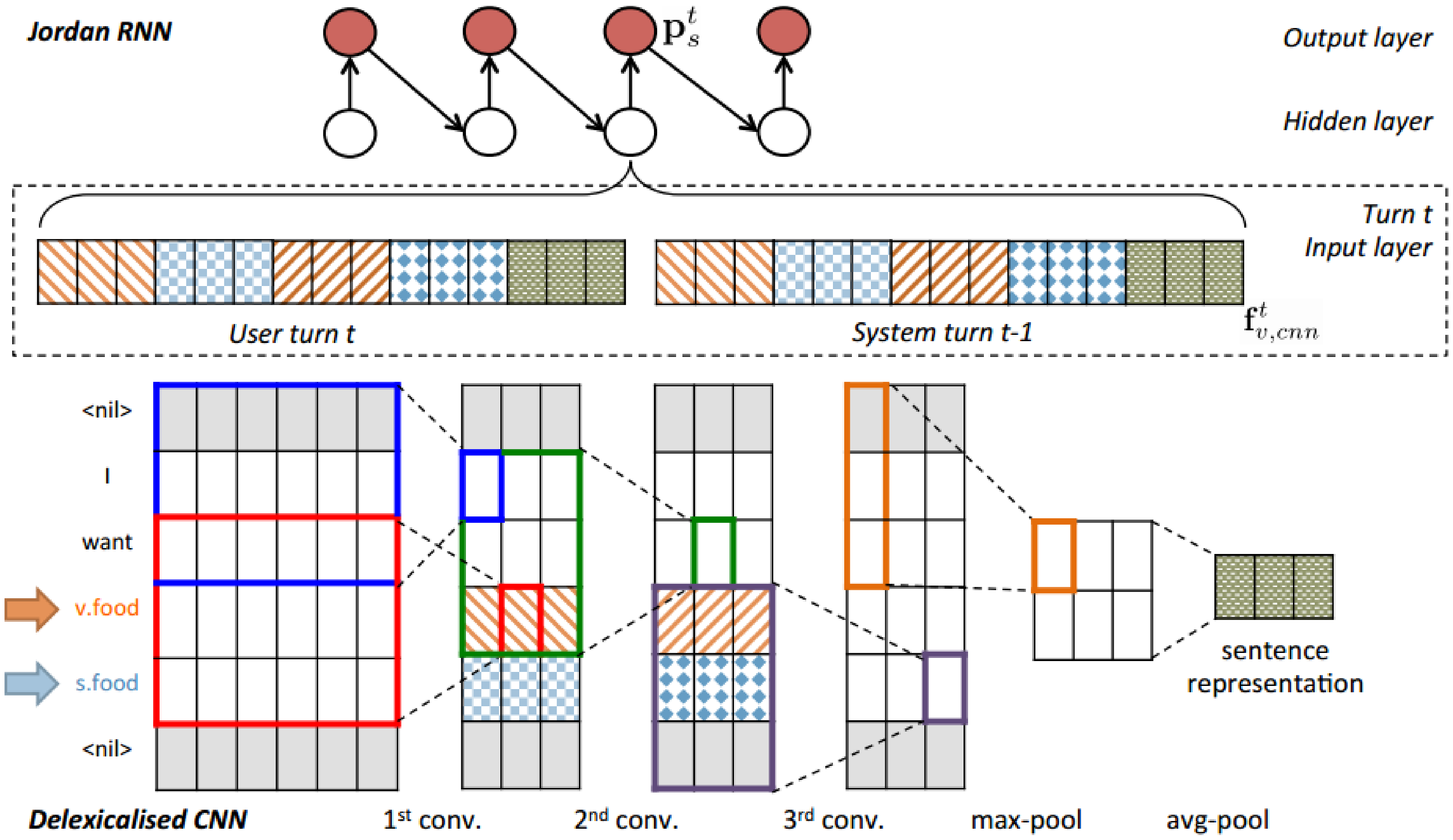
Restaurants

Date	11/15/17		
Time	6:00 pm	6:30 pm	7:00 pm
Restaurant	Cascal		
#People	2		

Less Likely — More Likely

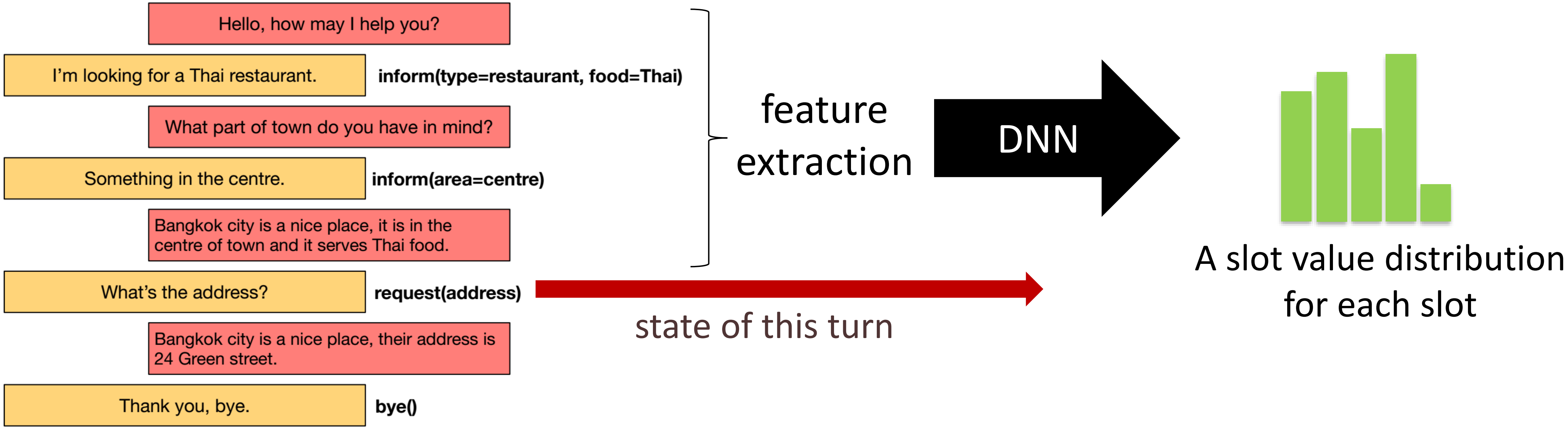


RNN-CNN DST (Mrkšić+, 2015)



(Figure from Wen et al, 2016)

DNN for DST



multi-turn conversation

Dialog State Tracking Challenge (DSTC)

(Williams et al. 2013, Henderson et al. 2014, Henderson et al. 2014, Kim et al. 2016, Kim et al. 2016)

Challenge	Type	Domain	Data Provider	Main Theme
DSTC1	Human-Machine	Bus Route	CMU	Evaluation Metrics
DSTC2	Human-Machine	Restaurant	U. Cambridge	User Goal Changes
DSTC3	Human-Machine	Tourist Information	U. Cambridge	Domain Adaptation
DSTC4	Human-Human	Tourist Information	I2R	Human Conversation
DSTC5	Human-Human	Tourist Information	I2R	Language Adaptation

DSTC4-5

- Type: Human-Human
- Domain: Tourist Information

{Topic: Accommodation; NAME: InnCrowd Backpackers Hostel; GuideAct: REC; TouristAct: ACK}

Guide: Let's try this one, okay?

Tourist: Okay.

Guide: It's InnCrowd Backpackers Hostel in Singapore. If you take a dorm bed per person only twenty dollars. If you take a room, it's two single beds at fifty nine dollars.

Tourist: Um. Wow, that's good.

Guide: Yah, the prices are based on per person per bed or dorm. But this one is room. So it should be fifty nine for the two room. So you're actually paying about ten dollars more per person only.

Tourist: Oh okay. That's- the price is reasonable actually. It's good.

{Topic: Accommodation; Type: Hostel; Pricerange: Cheap; GuideAct: ACK; TouristAct: REQ}

Tourist: Can you give me some uh- tell me some cheap rate hotels, because I'm planning just to leave my bags there and go somewhere take some pictures.

Guide: Okay. I'm going to recommend firstly you want to have a backpack type of hotel, right?

Tourist: Yes. I'm just gonna bring my backpack and my buddy with me. So I'm kinda looking for a hotel that is not that expensive. Just gonna leave our things there and, you know, stay out the whole day.

Guide: Okay. Let me get you hm hm. So you don't mind if it's a bit uh not so roomy like hotel because you just back to sleep.

Tourist: Yes. Yes. As we just gonna put our things there and then go out to take some pictures.

Guide: Okay, um-

Tourist: Hm.

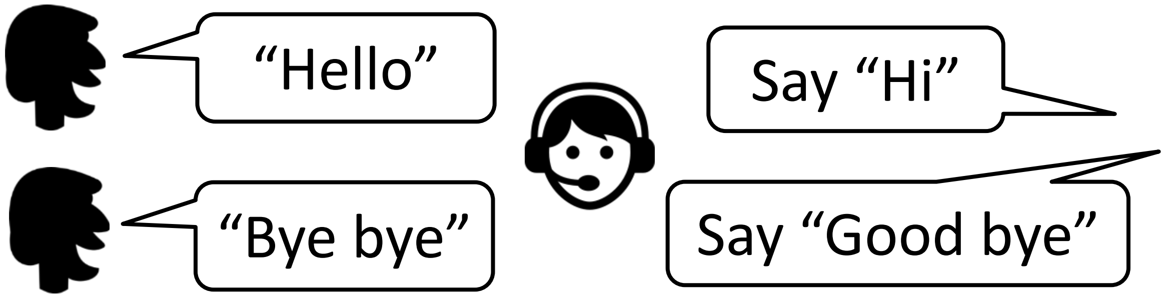
Dialogue Policy Optimization

Hello, how may I help you?	greeting ()
I’m looking for a Thai restaurant.	request (restaurant; foodtype=Thai)
What part of town do you have in mind?	request (area)
Something in the centre.	inform (area=centre)
Bangkok city is a nice place, it is in the centre of town and it serves Thai food.	inform (restaurant=Bangkok city, area=centre of town, foodtype=Thai)
What’s the address?	request (address)
Bangkok city is a nice place, their address is 24 Green street.	inform (address=24 Green street)
Thank you, bye.	bye ()

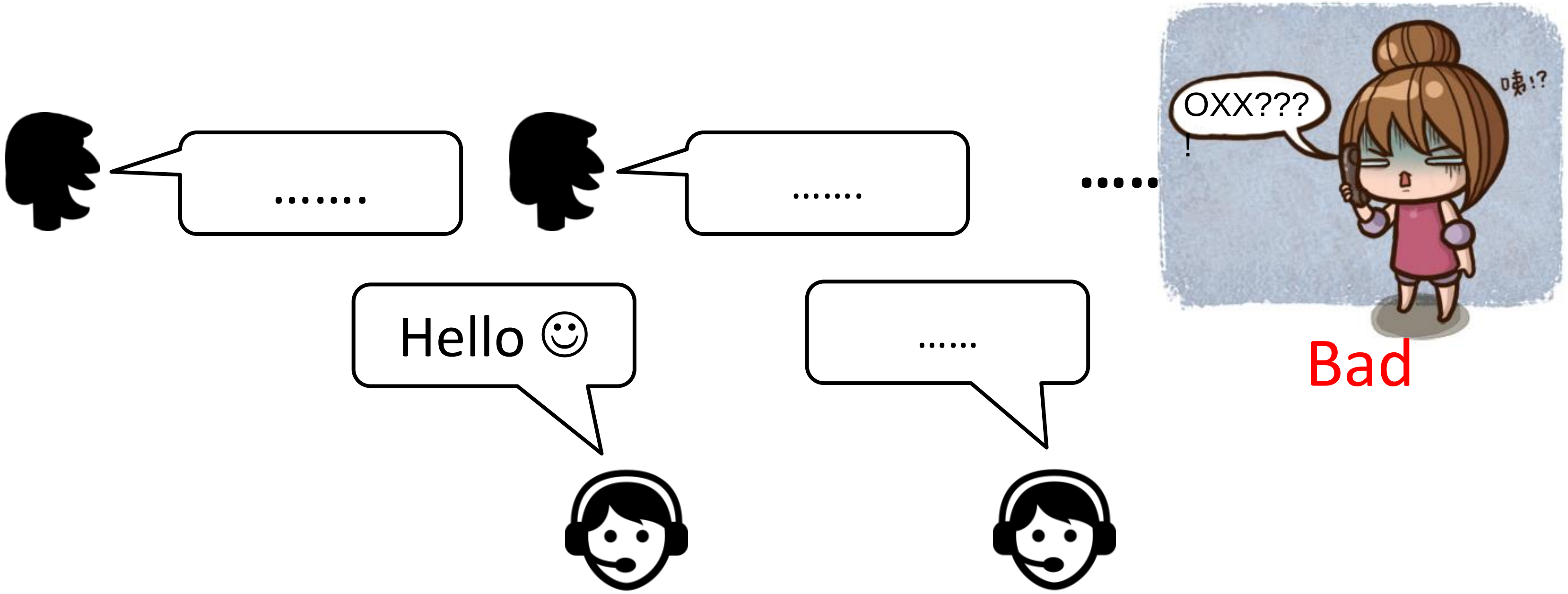
Supervised v.s. Reinforcement

- Supervised

Learning from teacher



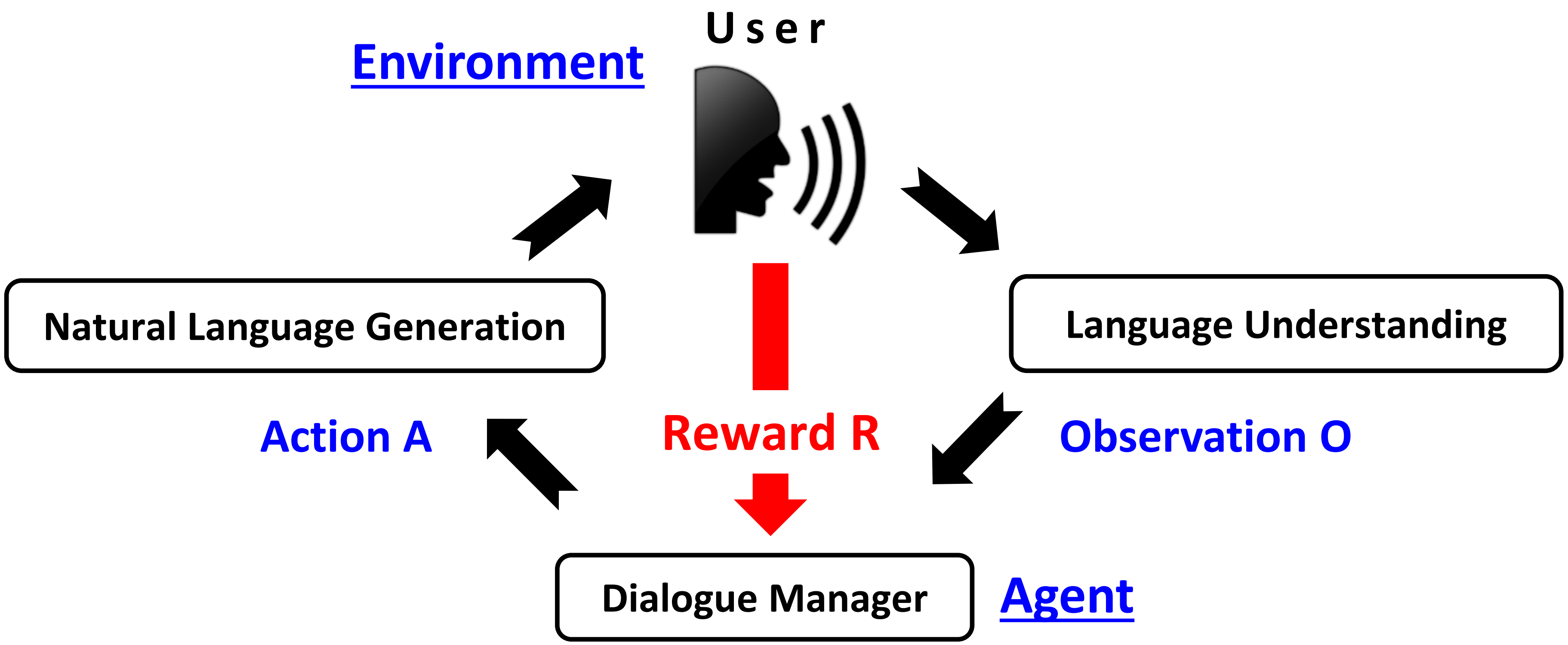
- Reinforcement



Learning from critics

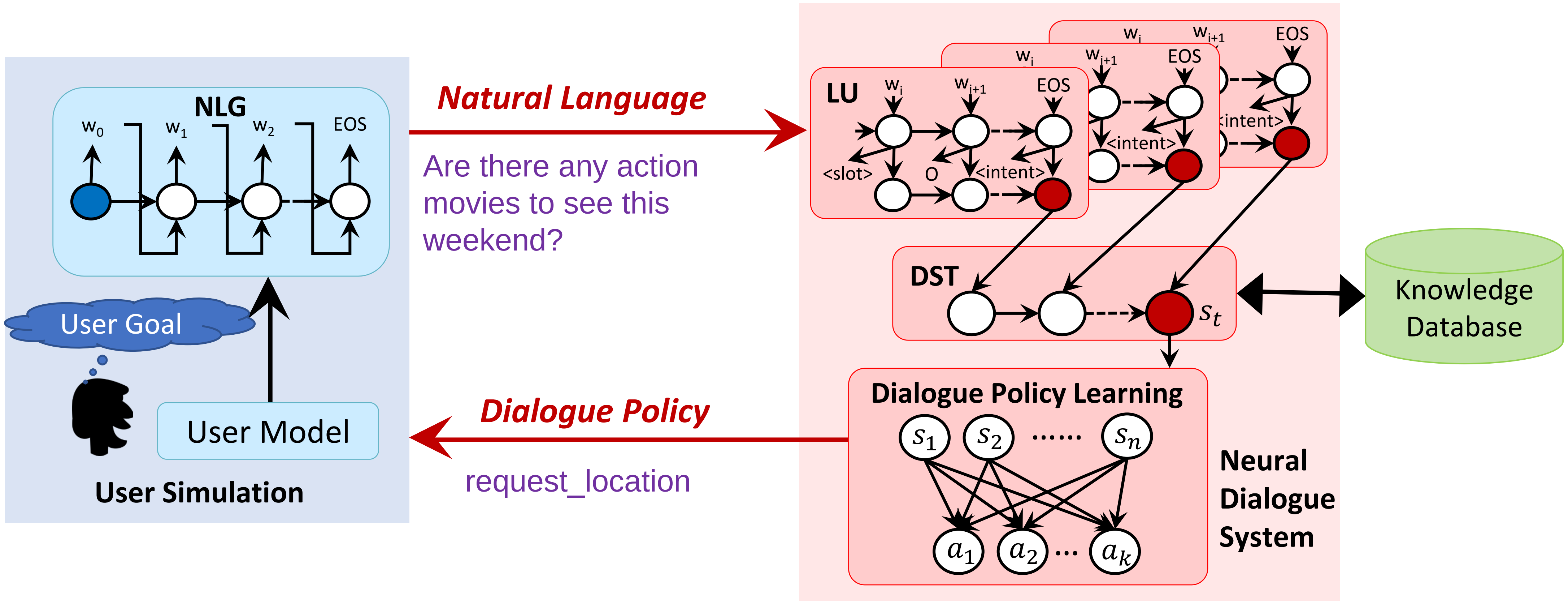
Dialogue Policy Optimization

- Dialogue management in a RL framework



Select the best action that **maximizes the future reward**

E2E Task-Completion Bot (TC-Bot) (Li et al., 2017)



Idea: SL for each component and RL for end-to-end training the neural dialogue system

Reinforcement Learning Agent (Li et al., 2017)

User goal – ***Two tickets for "Deadpool" tomorrow 9PM at AMC Pacific Place 11 theater, Seattle.***

RULE BASED SYSTEM

User: Do you have any ticket for movie 'Deadpool' at 9:00 pm?
Agent: Which theater do you want?
User: AMC pacific place 11 theater, please!
Agent: How many tickets do you want?
User: Two!
Agent: What date?
User: Tomorrow!
Agent: Which city do you want?
User: Seattle!
Agent: Do you have any specification for zip code?
User: No, I do not care.
Agent: Do you have any distance constraint?
User: No, I do not care.
Agent: How about video format? 2d? Imax 3d?
User: Any kind is okay.
Agent: Do you have any constraint to price?
User: No!
Agent: Okay, I will issue 2 tickets for you, tomorrow 9:00 pm at AMC pacific place 11 theater, Seattle, movie 'Deadpool'.
User: Okay, t
Agent: Thanks

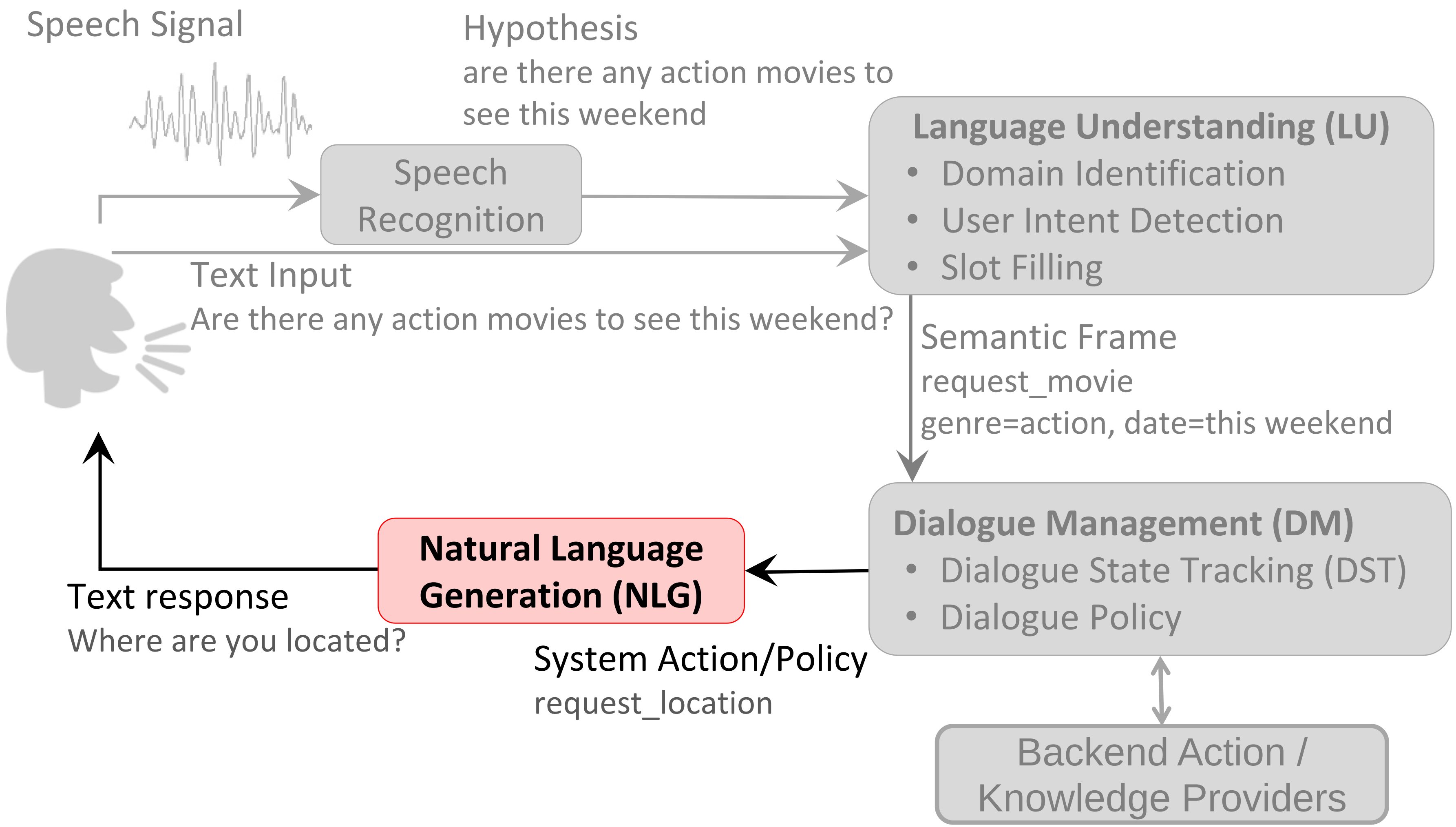
REINFORCEMENT LEARNING SYSTEM

User: I want two tickets for movie 'Deadpool'.
Agent: What is the starttime you prefer?
User: 9:00 PM
Agent: Which day?
User: Tomorrow
Agent: Which city?
User: Seattle
Agent: Which theater do you prefer?
User: AMC Pacific Place 11 theater.
Agent: Okay, I will issue 2 tickets for you, tomorrow 9:00 pm at AMC pacific place 11 theater, Seattle, movie 'Deadpool'.
User: Okay, thanks.
Agent: Thanks!

Issue: no notion about what requests can be skipped

Skip the requests the user may not care about to improve efficiency

Task-Oriented Dialogue Systems (Young, 2000)



Natural Language Generation (NLG)

- Mapping dialogue acts into natural language

inform(name=Seven_Days, foodtype=Chinese)

⇩

Seven Days is a nice Chinese restaurant

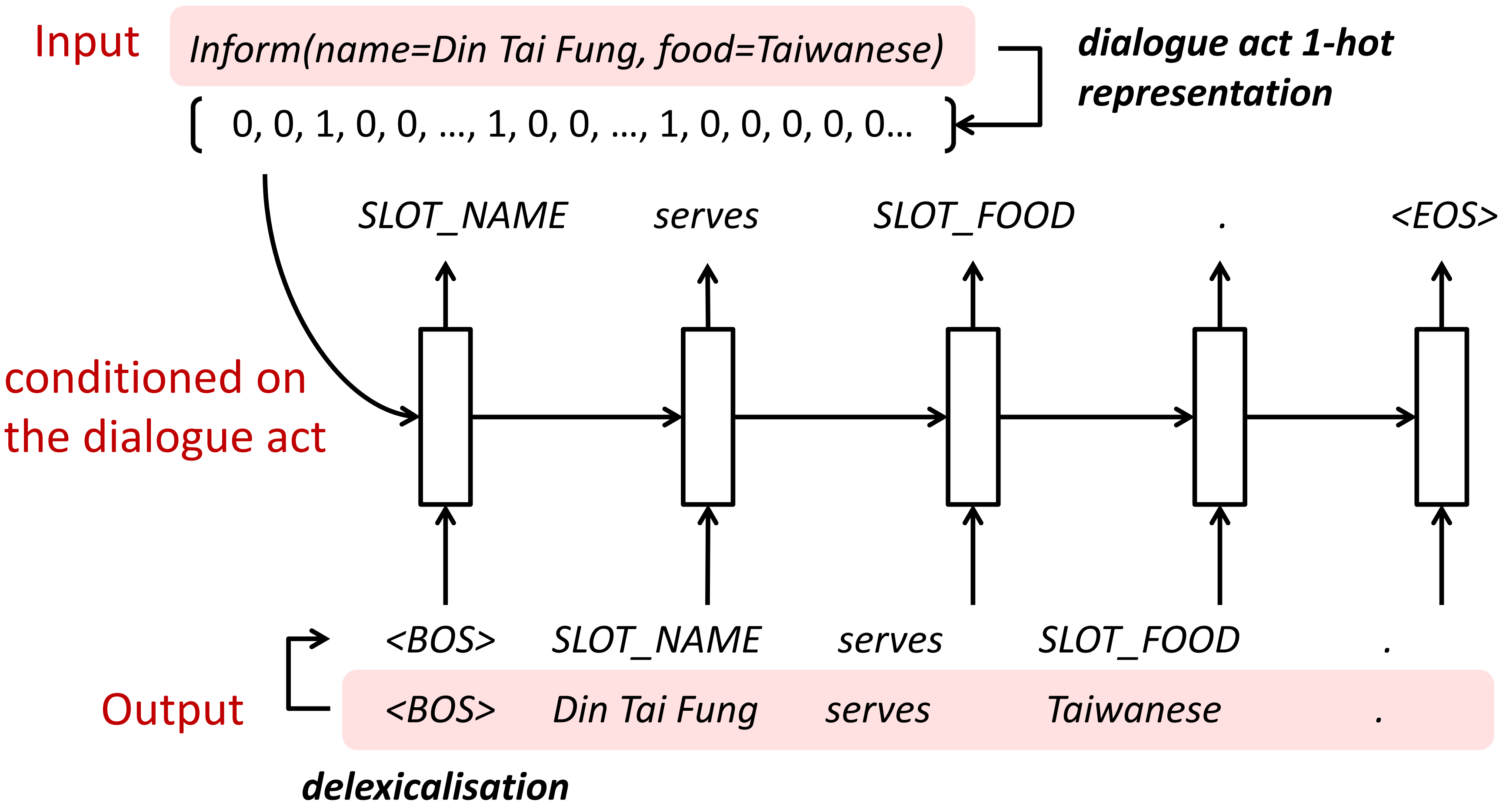
Template-Based NLG

- Define a set of rules to map frames to natural language

Semantic Frame	Natural Language
confirm()	“Please tell me more about the product you are looking for.”
confirm(area=$V)	“Do you want somewhere in the $V?”
confirm(food=$V)	“Do you want a $V restaurant?”
confirm(food=$V,area=$W)	“Do you want a $V restaurant in the $W.”

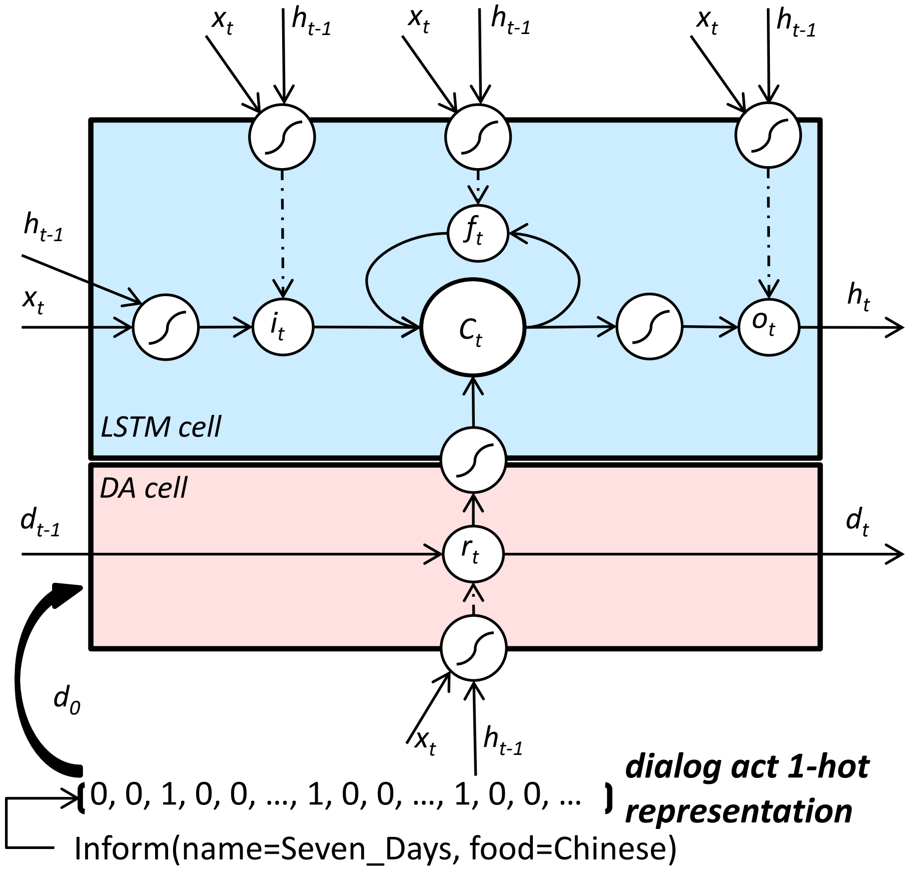
Pros: simple, error-free, easy to control
Cons: time-consuming, rigid, poor scalability

RNN-Based LM NLG (Wen et al., 2015)



Semantic Conditioned LSTM (Wen et al., 2015)

- Issue: semantic repetition
  - Din Tai Fung is a great Taiwanese restaurant that serves Taiwanese.
  - Din Tai Fung is a child friendly restaurant, and also allows kids.



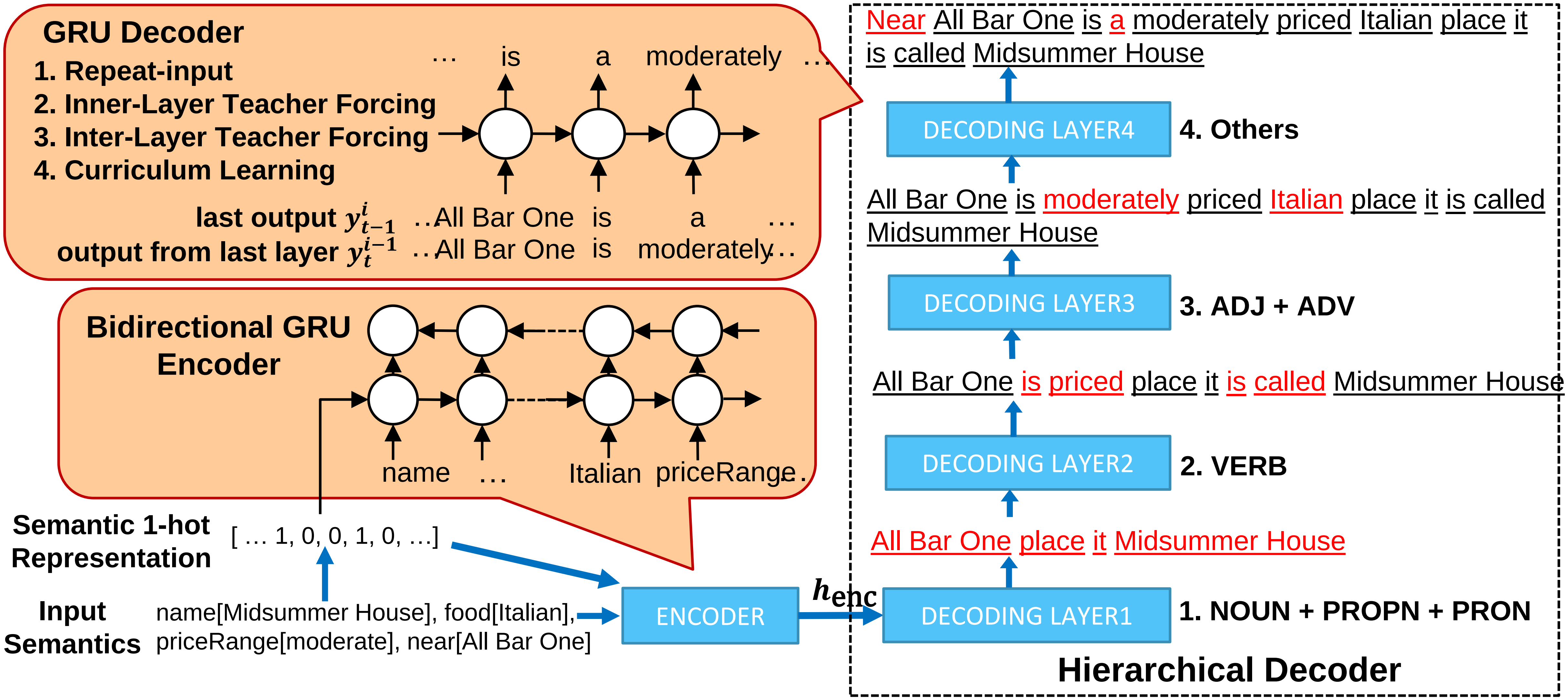
Idea: using gate mechanism to control the generated semantics (dialogue act/slots)

Issues in NLG

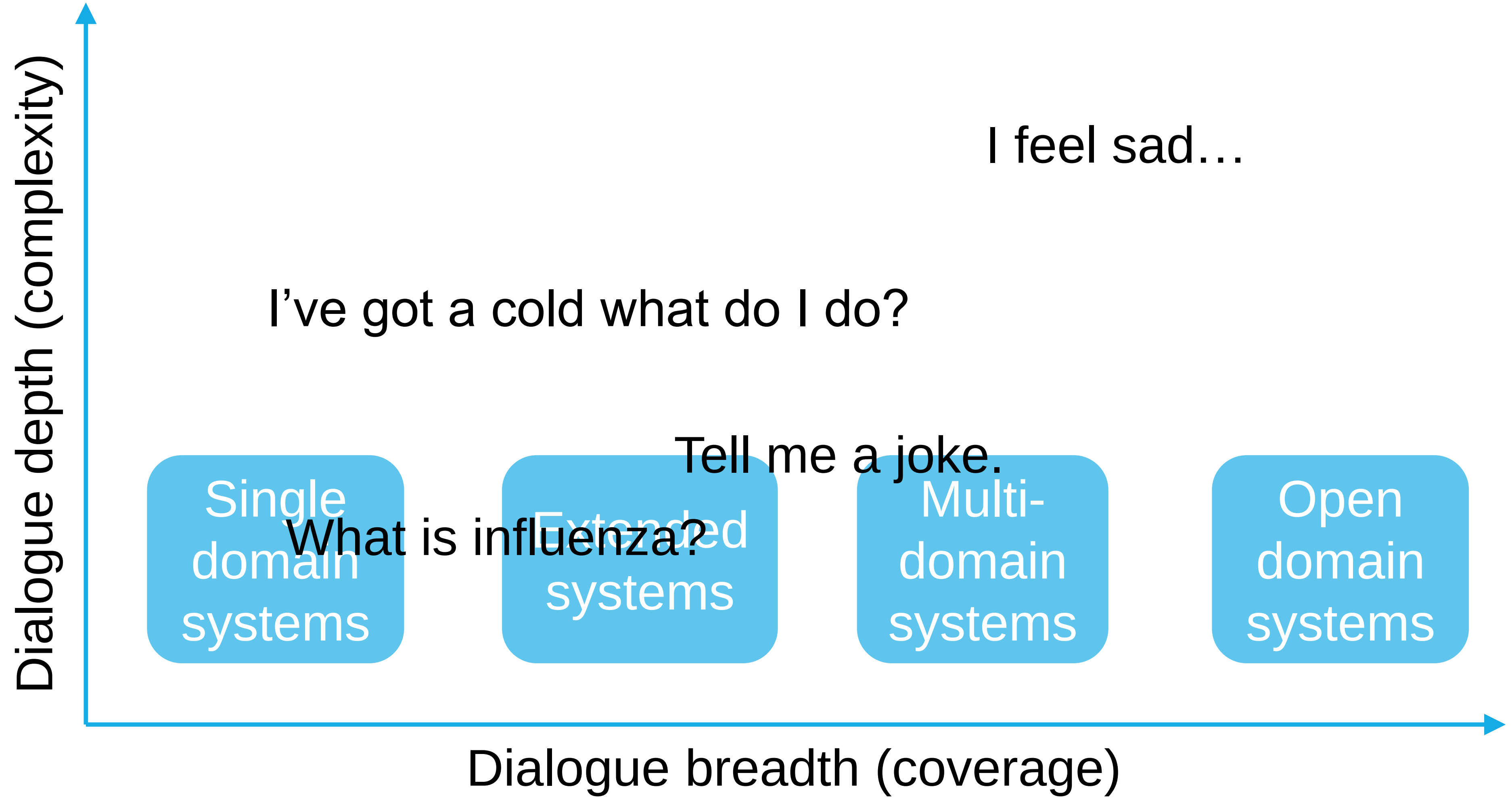
- Issue
  - NLG tends to generate shorter sentences
  - NLG may generate grammatically-incorrect sentences
- Solution
  - Generate word patterns in an order
  - Consider linguistic patterns

Hierarchical NLG w/ Linguistic Patterns

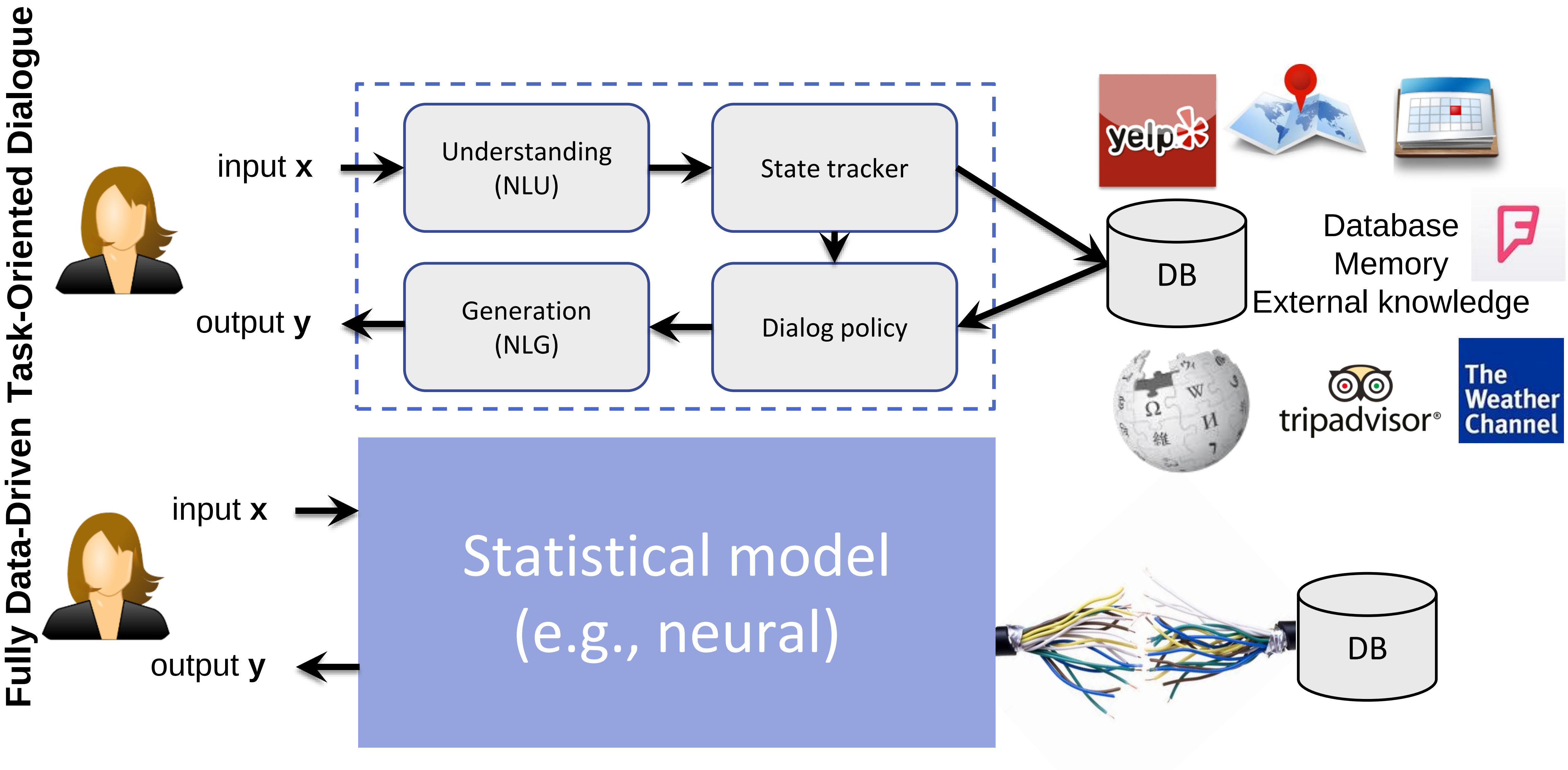
(Su et al., 2018)



Evolution Roadmap



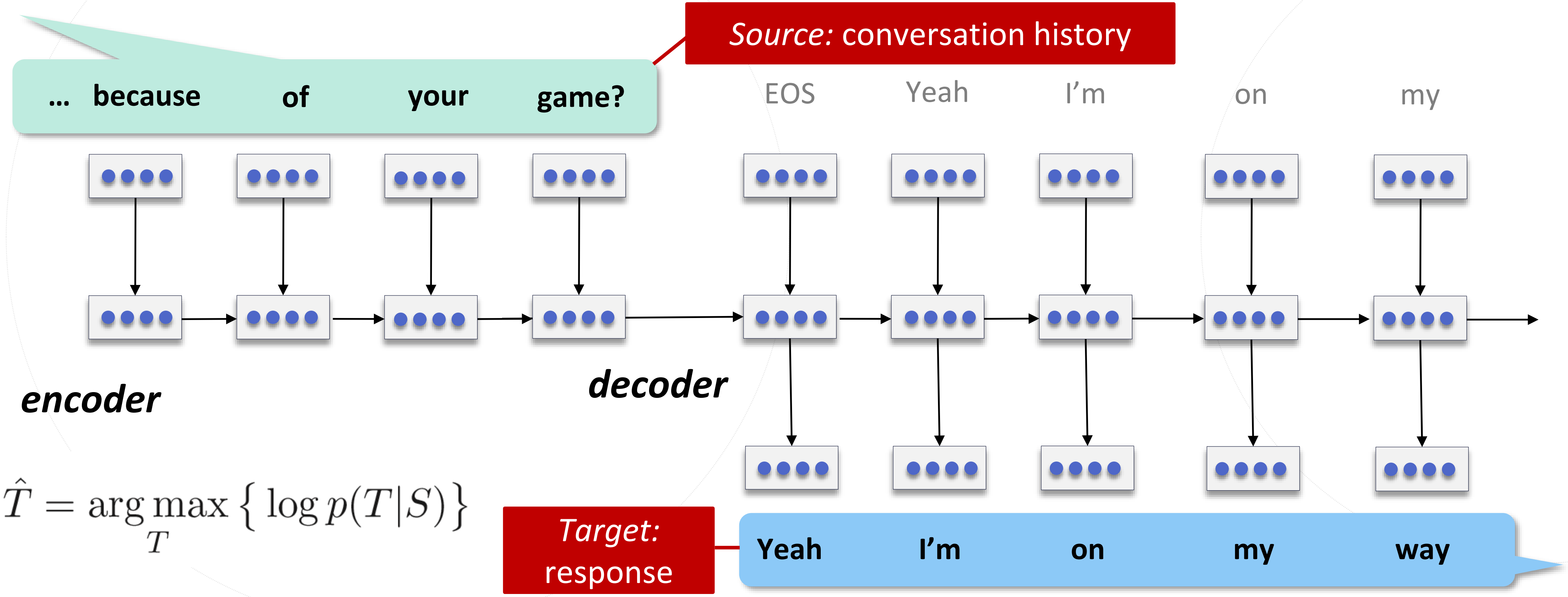
Dialogue Systems



48

Chit-Chat Social Bots

Neural Response Generation (Sordoni et al., 2015; Vinyals & Le, 2015)

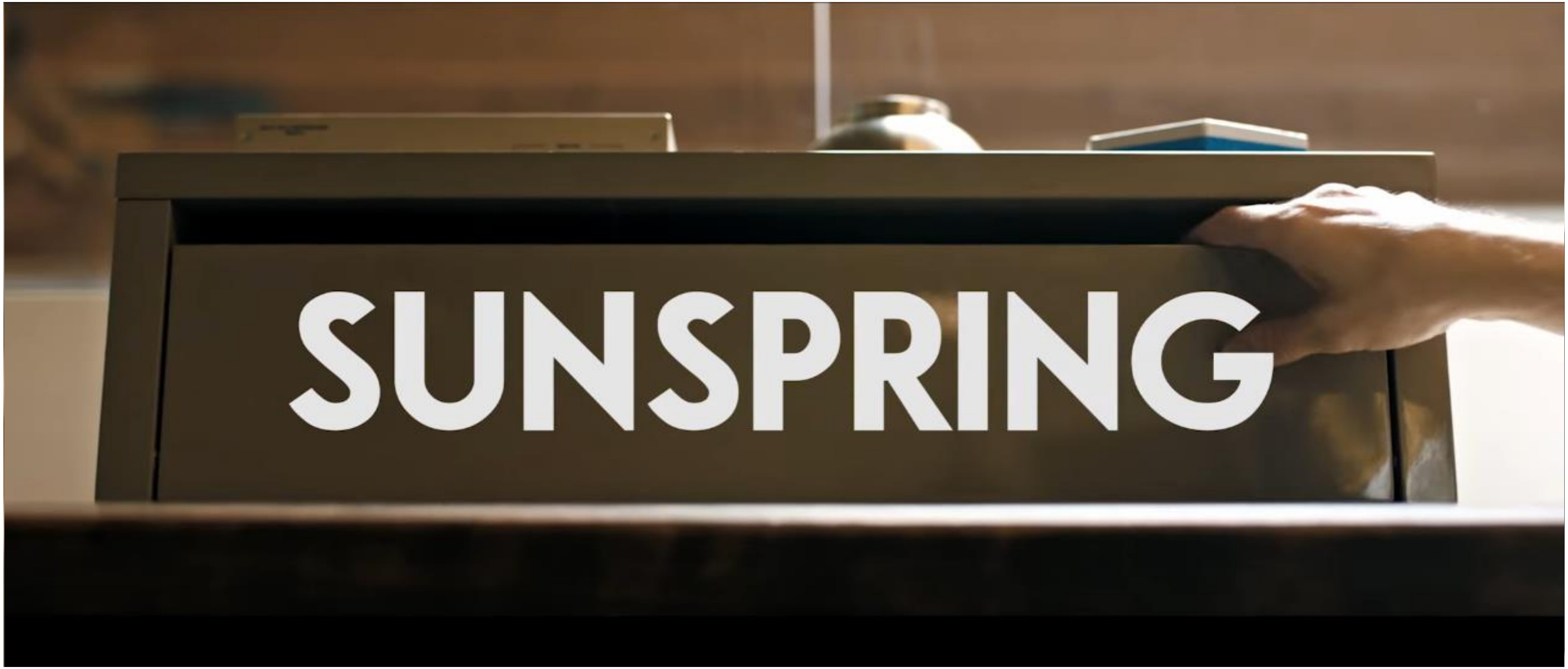


$$\hat{T} = \arg\max_{T} \left\{ \log p(T|S) \right\}$$

Learns to generate dialogues from offline data (no state, action, intent, slot, etc.)

Sci-Fi Short Film - SUNSPRING

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

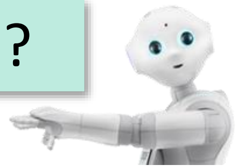


Issue 1: Blandness Problem

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”

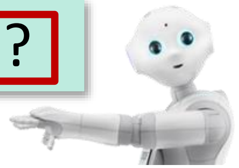
“Oh my god”

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?

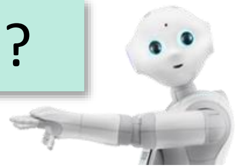
I think he'd be a good signing.

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

Depends on how much you drink!

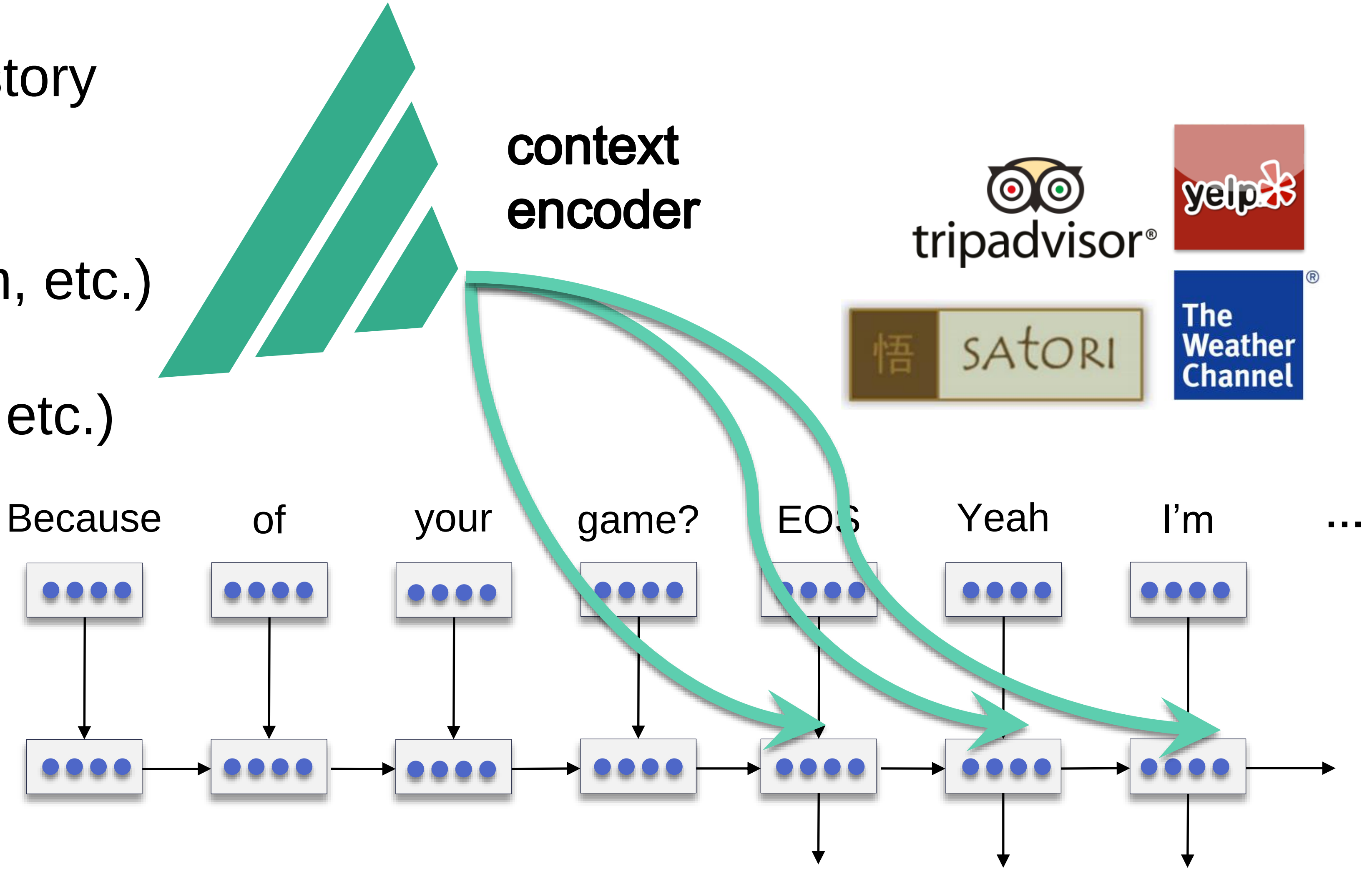
Well he was on in Bromley a while ago... still touring.

I've never seen him live.



Real-World Conversations

- Multimodality
  - Conversation history
  - Persona
  - User profile data (bio, social graph, etc.)
  - Visual signal (camera, picture etc.)
  - Knowledge base
  - Mood
  - Geolocation
  - Time

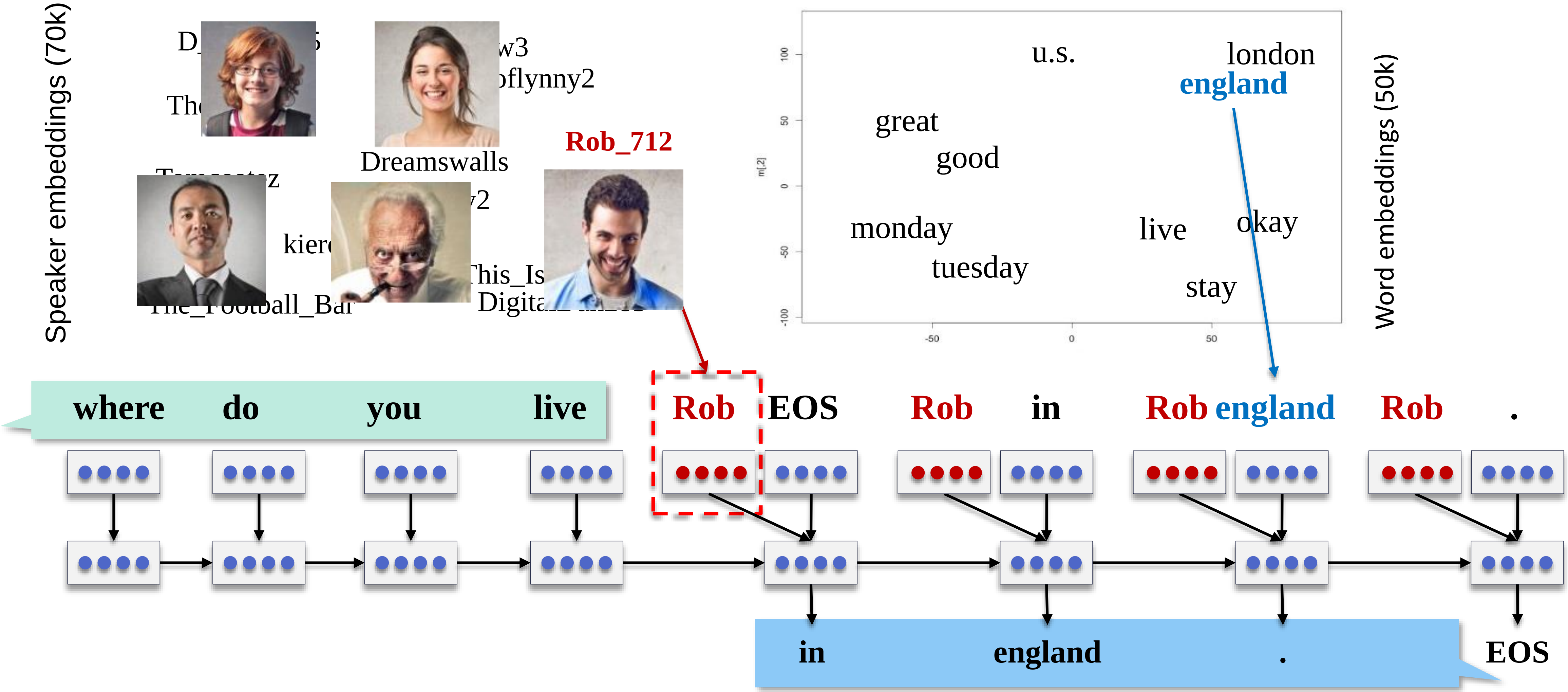


Issue 2: Response Inconsistency

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.

Personalized Response Generation (Li et al., 2016)



Persona Model for Speaker Consistency

(Li et al., 2016)

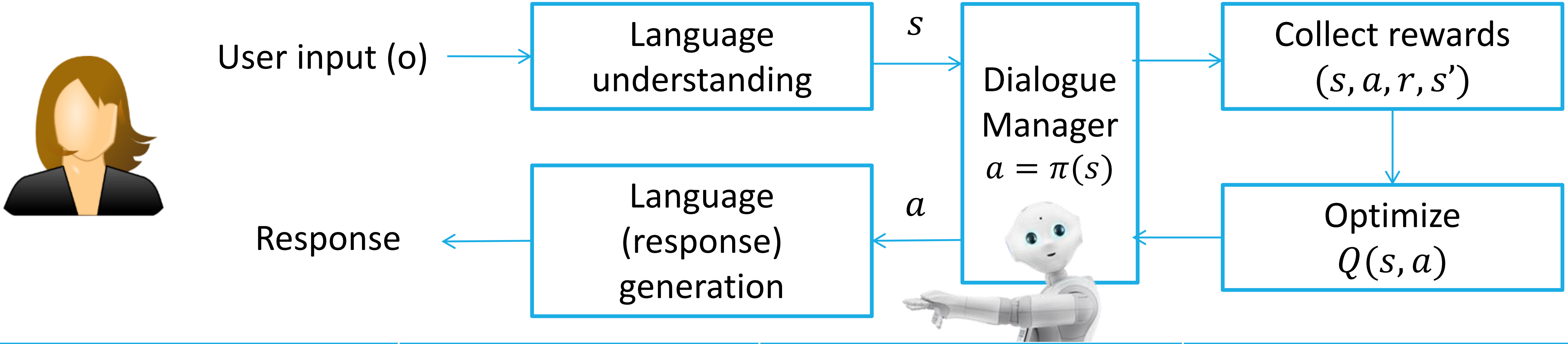
Baseline model → inconsistency

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.

Issue 3: Dialogue-Level Optimization via RL



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(?)

Deep RL for Response Generation (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

Issue 4: No Grounding (Sordoni et al., 2015; Li et al., 2016)

Social Chat
Engaging, Human-Like Interaction
(Ungrounded)

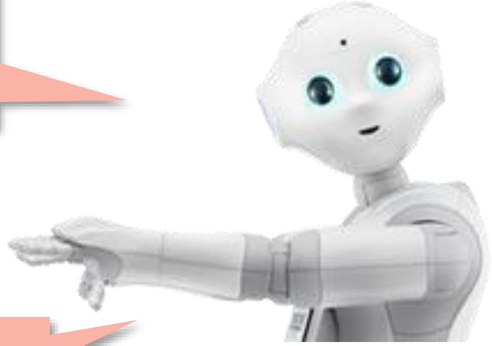
Task-Oriented
Task Completion, Decision Support
(Grounded)

The weather is so depressing these days.

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

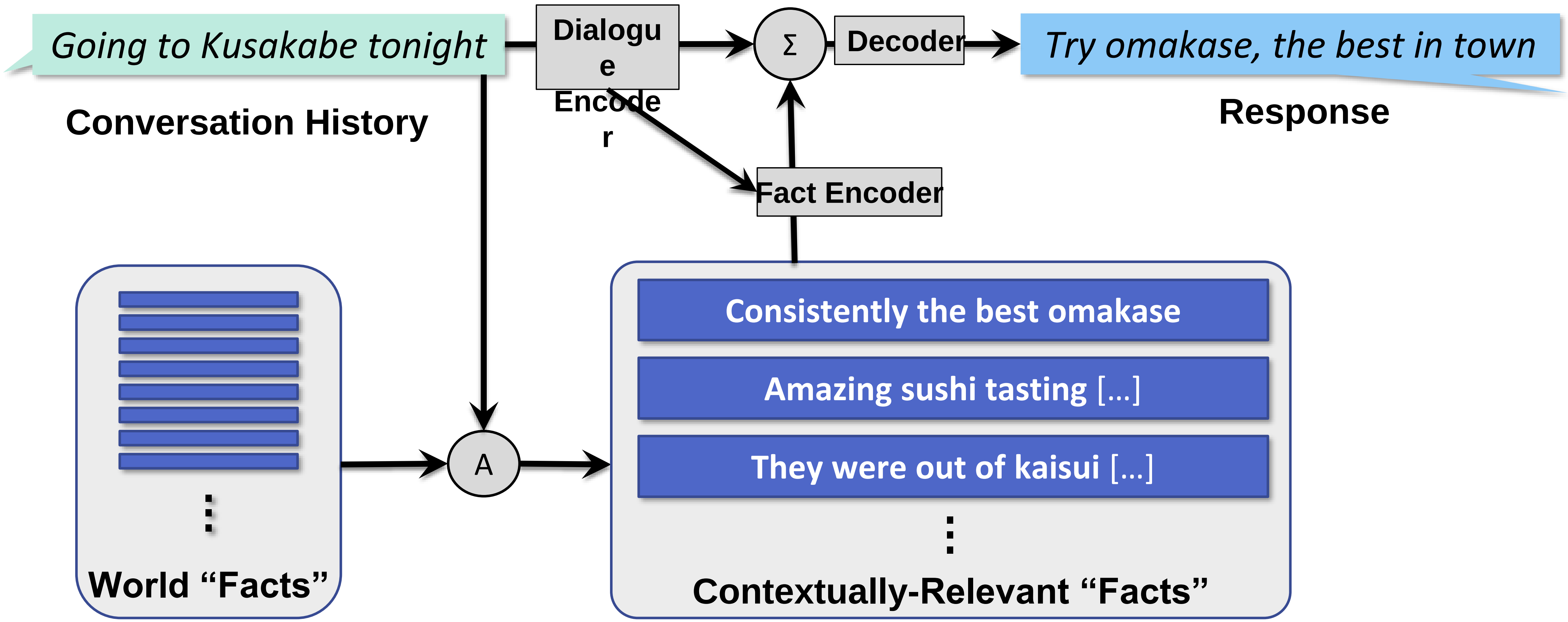
Any recommendation?

Try Dry Falls, it's spectacular!



Knowledge-Grounded Responses

(Ghazvininejad et al., 2017)



Conversation and Non-Conversation Data

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

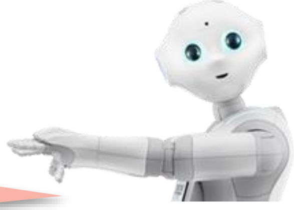
*Try **C**, one of the best **D** in the city.*

Conversation Data

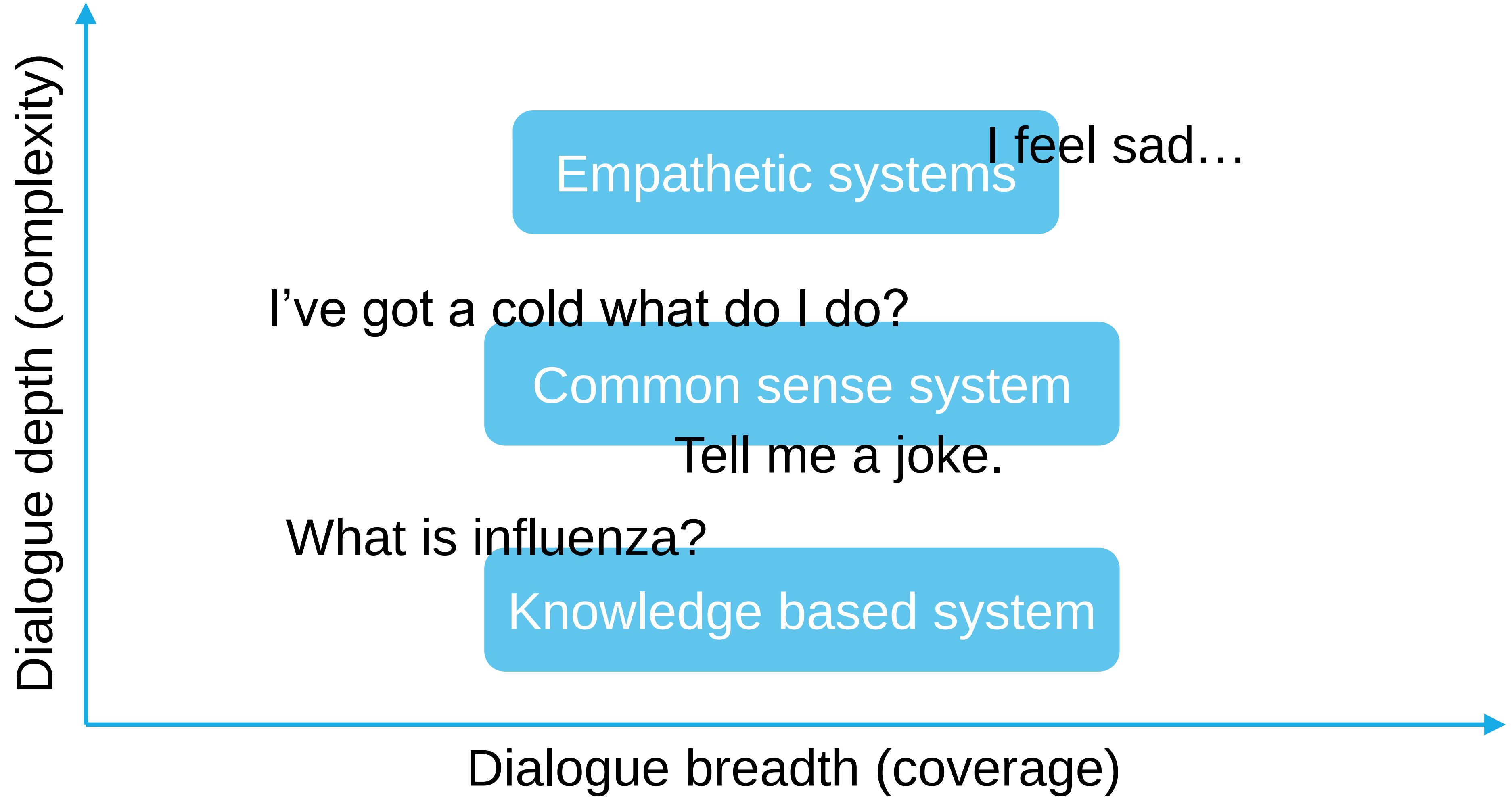
Knowledge Resource

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

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



Evolution Roadmap



Multimodality & Personalization (Chen et al., 2018)

- Task: user intent prediction
- Challenge: language ambiguity

send to vivian

Communication

Email? v.s. Message?

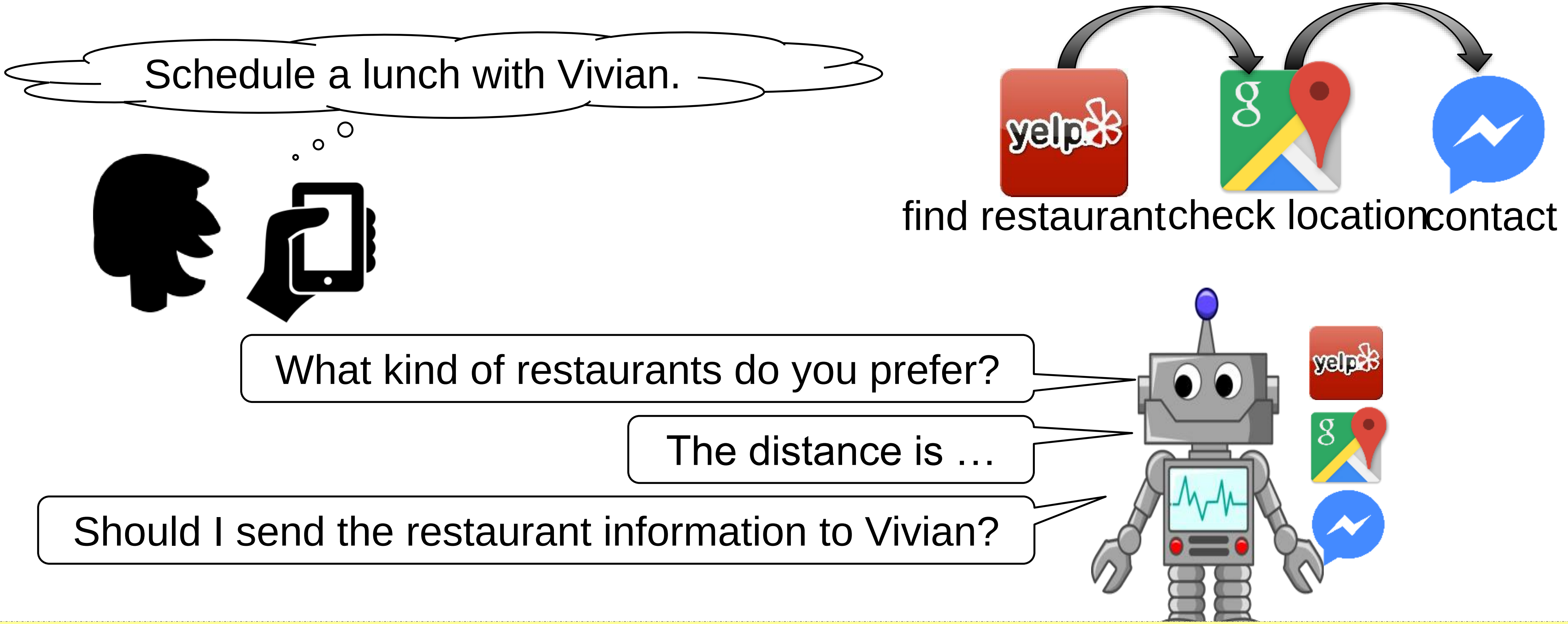
① **User preference**
  - ✓ Some people prefer “Message” to “Email”
  - ✓ Some people prefer “Ping” to “Text”

② **App-level contexts**
  - ✓ “Message” is more likely to follow “Camera”
  - ✓ “Email” is more likely to follow “Excel”

Behavioral patterns in history helps intent prediction.

High-Level Intention Learning (Sun et al., 2016)

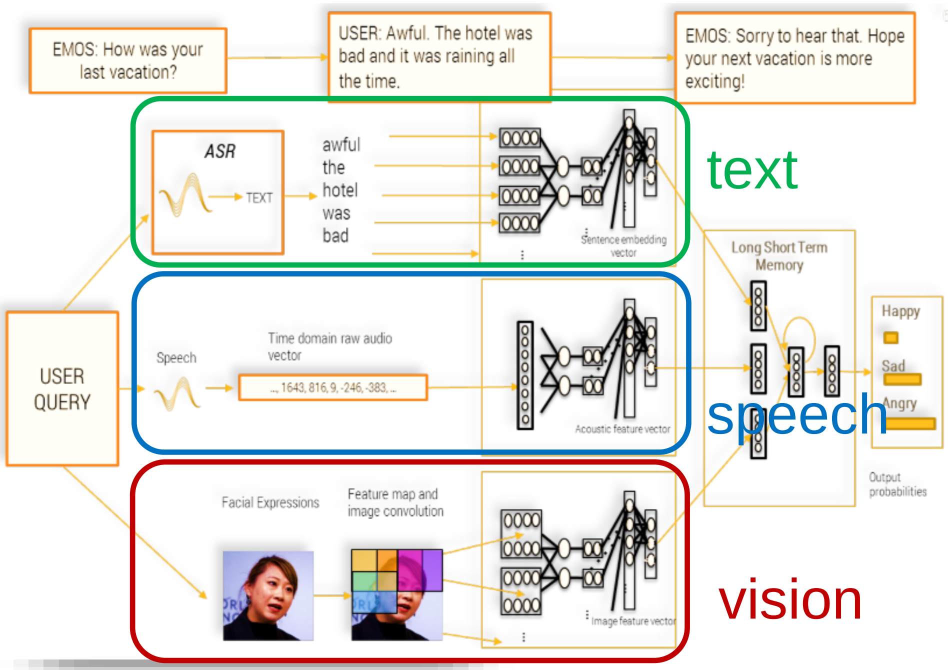
- High-level intention may span several domains



Users interact via <u>high-level descriptions</u> and the system learns <u>how to plan the dialogues</u>

Empathy in Dialogue System (Fung et al., 2016)

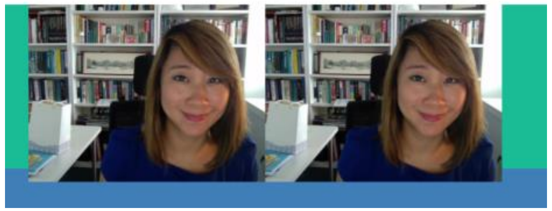
- Embed an empathy module
  - Recognize emotion using multimodality
  - Generate emotion-aware responses



Emotion Recognizer

Zara - The Empathetic Supergirl

Made with love by Ivo Technologies in collaboration with Hong Kong University of Science and Technology



Face recognition output

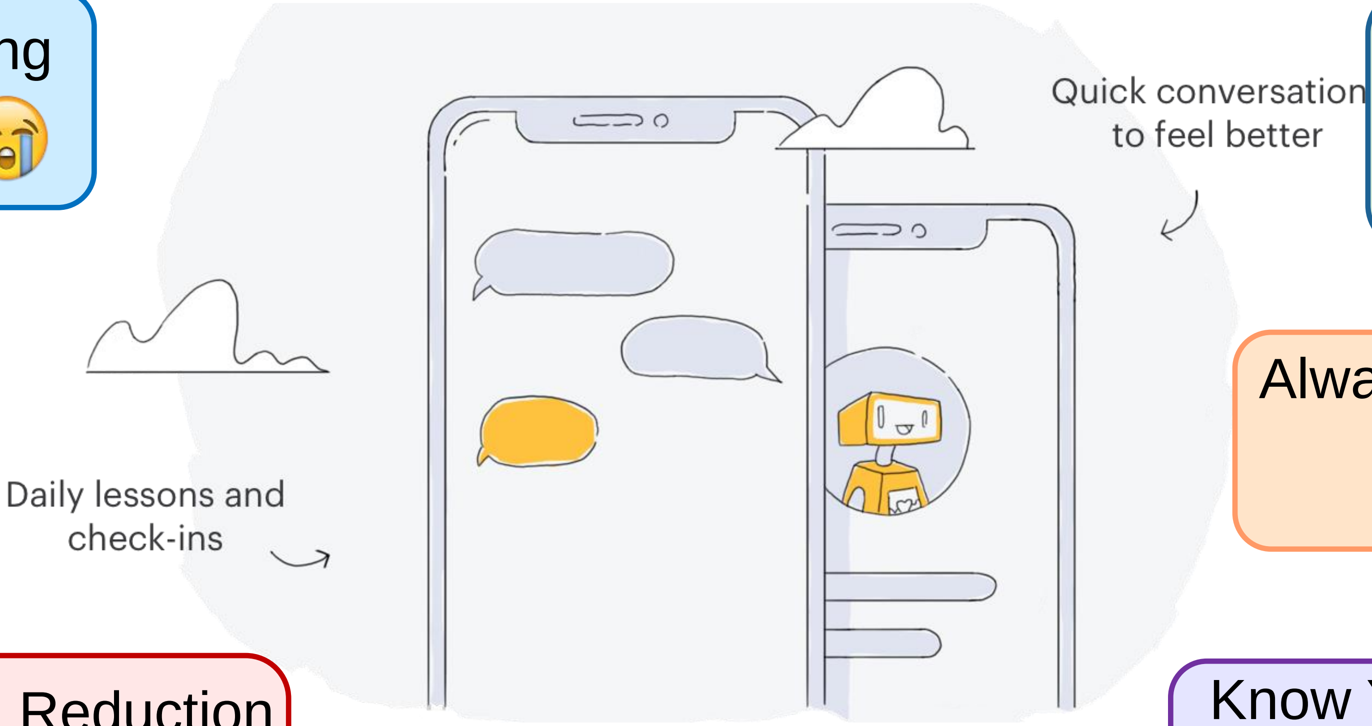
(index):1728
(index):1729
{
"recognition": "Race: Asian Confidence: 65.42750000000001 Smiling: 3.95896 Gender: Female Confidence: 88.9369",
"race": "Asian",
"race_confidence": "65.42750000000001",
"smiling": "3.95896",
"gender": "Female",
"gender_confidence": "88.9369"
}

Cognitive Behavioral Therapy (CBT)

Mood Tracking

Pattern Mining

Depression Reduction



Content Pr ing

Always Be There

Know You Well

Challenges & Conclusions

Challenge Summary

- The human-machine interface is a hot topic but several components must be integrated!
- Most state-of-the-art technologies are based on DNN
  - Requires huge amounts of labeled data
  - Several frameworks/models are available
- Fast domain adaptation with scarse data + re-use of rules/knowledge
- Handling reasoning and personalization
- Data collection and analysis from un-structured data
- Complex-cascade systems require high accuracy for working good as a whole


Her (2013)
What can machines achieve now or in the future?