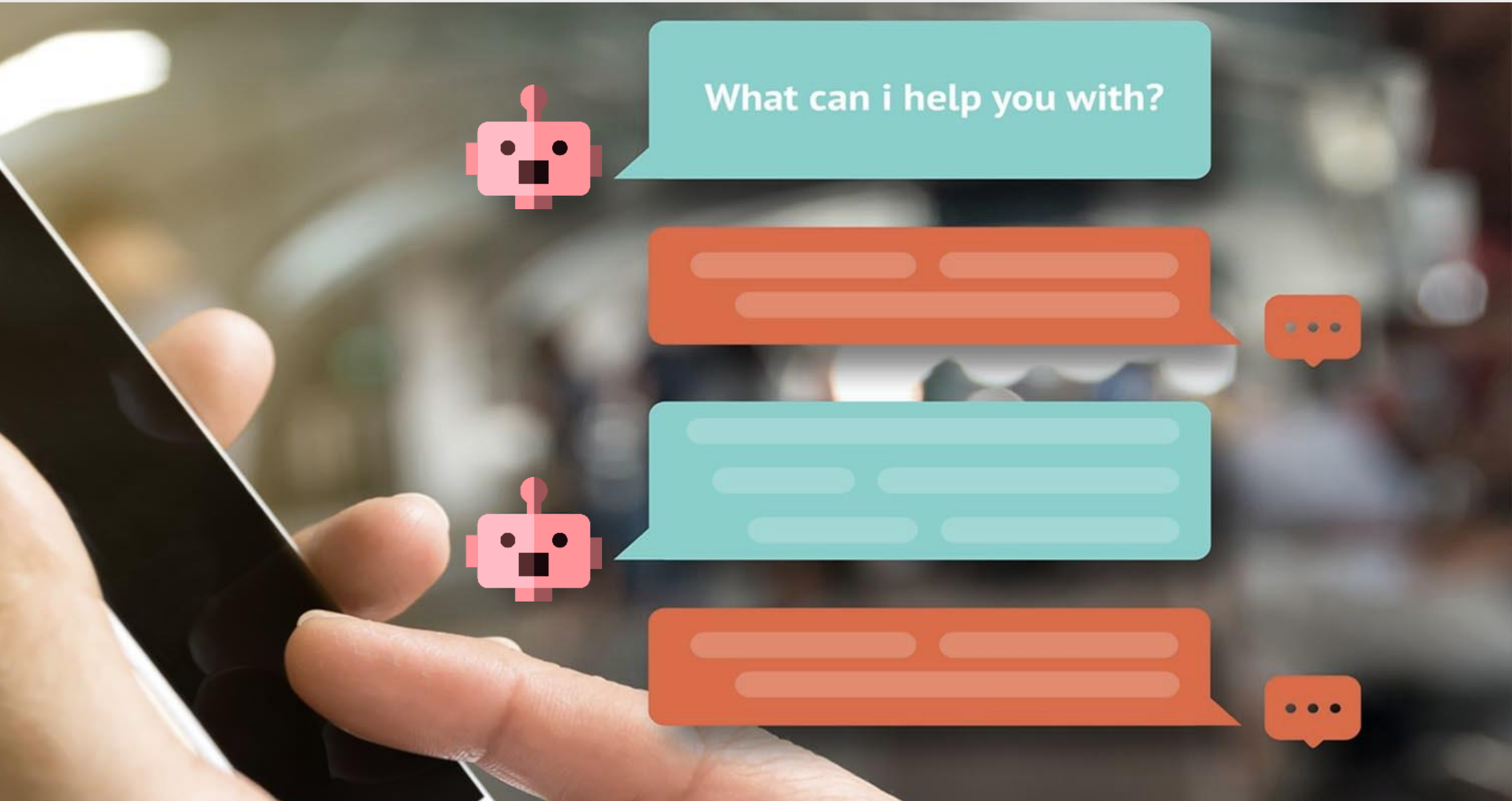
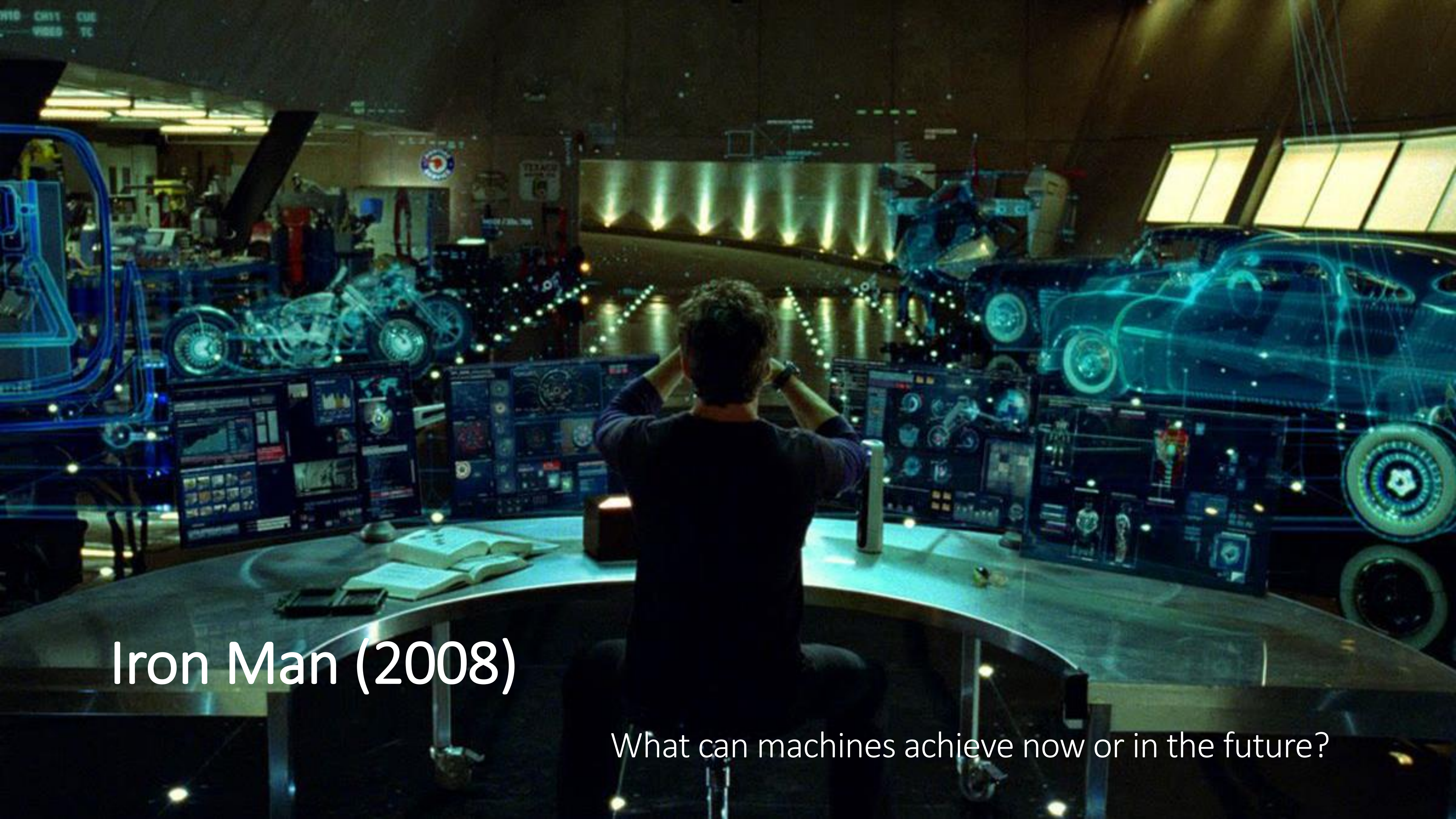


S9276: Towards Open-Domain Conversational AI

國立臺灣大學
National Taiwan University

YUN-NUNG (VIVIAN) CHEN 陳縕儂
HTTP://VIVIANCHEN.IDV.TW

GPU TECHNOLOGY CONFERENCE

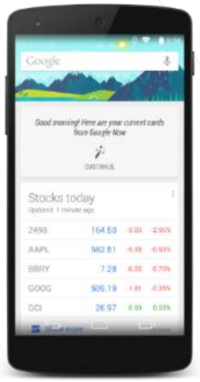


Iron Man (2008)

What can machines achieve now or in the future?

Language Empowering Intelligent Assistant

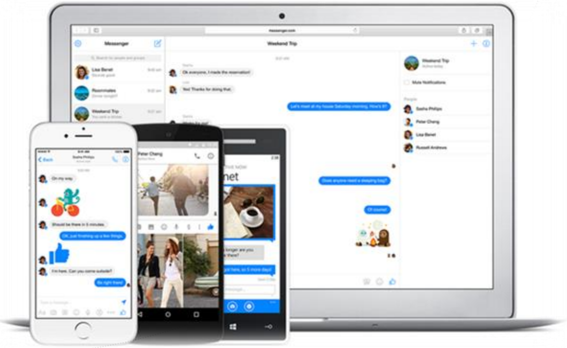
Apple Siri (2011)



Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

Amazon Alexa/Echo (2014)



Facebook M & Bot (2015)

Google Home (2016)

Apple HomePod (2017)

Why Natural Language?

- Global Digital Statistics (2018 January)

Total Population
7.59B

Internet Users
4.02B
7%

Active Social Media Users
3.20B
13%

Unique Mobile Users
5.14B
4%

Active Mobile Social Users
2.96B
14%

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

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	Task-Oriented Dialogues
"What should I do?"	Decision support	Task-Oriented Dialogues

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

- *Book me the flight ticket from Taipei to San Francisco*
- *Reserve a table at Din Tai Fung for 5 people, 7PM tonight*

- *Is GTC good to attend?*

Intelligent Assistants

show blockers assigned to Peter

get me Javier's contact info

who does Peter work for?

add a half hour meeting Friday at 11

current weather in London

what's my first meeting on Wed?

show me next week's schedule

when did Jennifer join our company?

how's my Jira sprint status?

show Rodney's supervisor

get today's agenda

when's the next meeting

translate "Good morning" to French

where does Charlotte sit?

how's my calendar tomorrow

when is Jim Harris anniversary?

what is the local time in Dublin Ireland?

cancel my 9 o'clock meeting tomorrow

Who is Shannon Anderson

find all Tracker's high cycle stories

when is Jonathan's birthday?

Show github leaderboard

show all my pull requests

which team does Olivia part of?

get all Github commits in last 7d

→ Task-Oriented

Conversational Agents

Chit-Chat

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

Task-Oriented

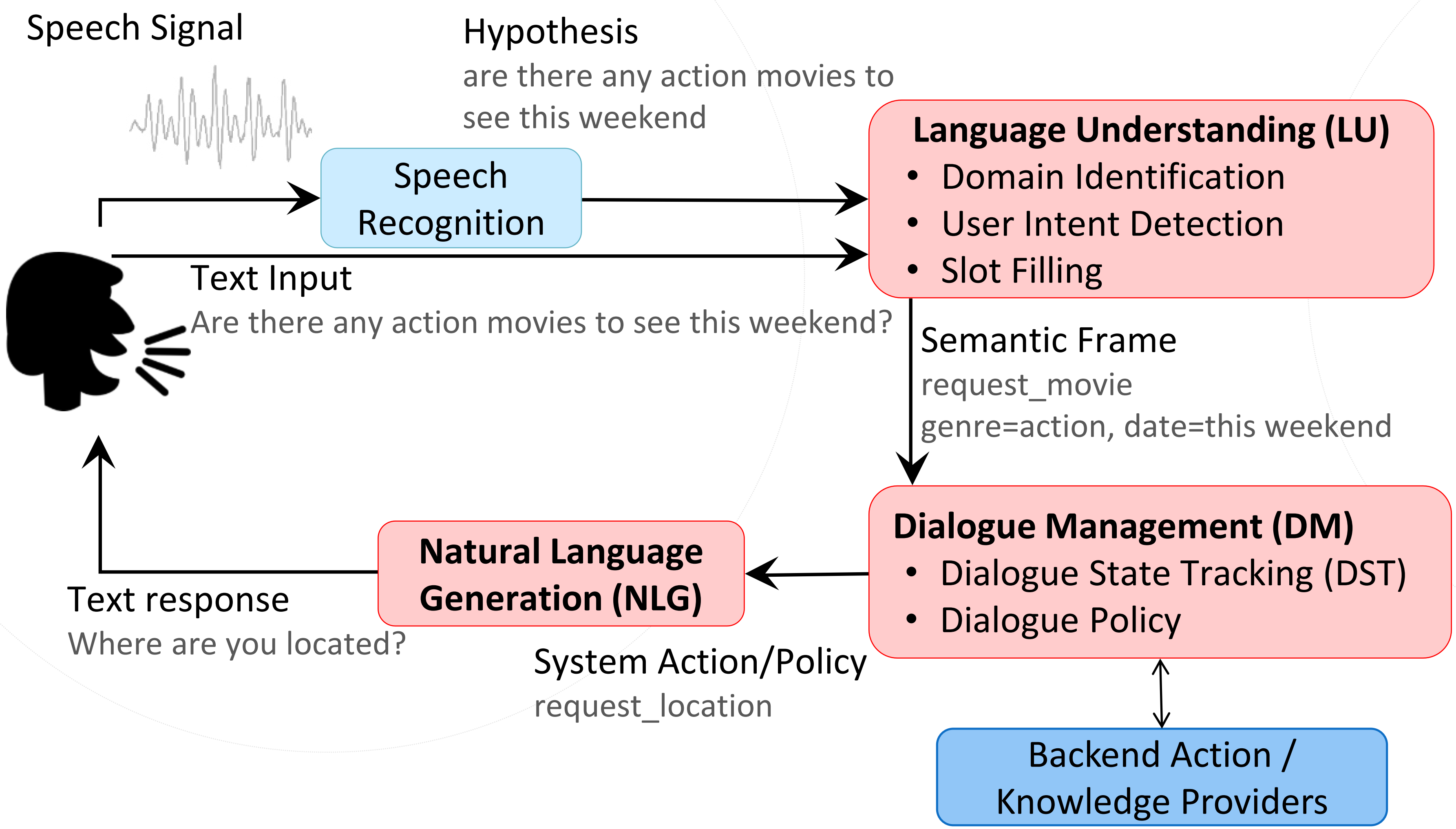
Single-domain, system-initiative → Multi-domain, contextual, mixed-initiative → End-to-end learning, massively multi-domain

Task-Oriented Dialogue Systems

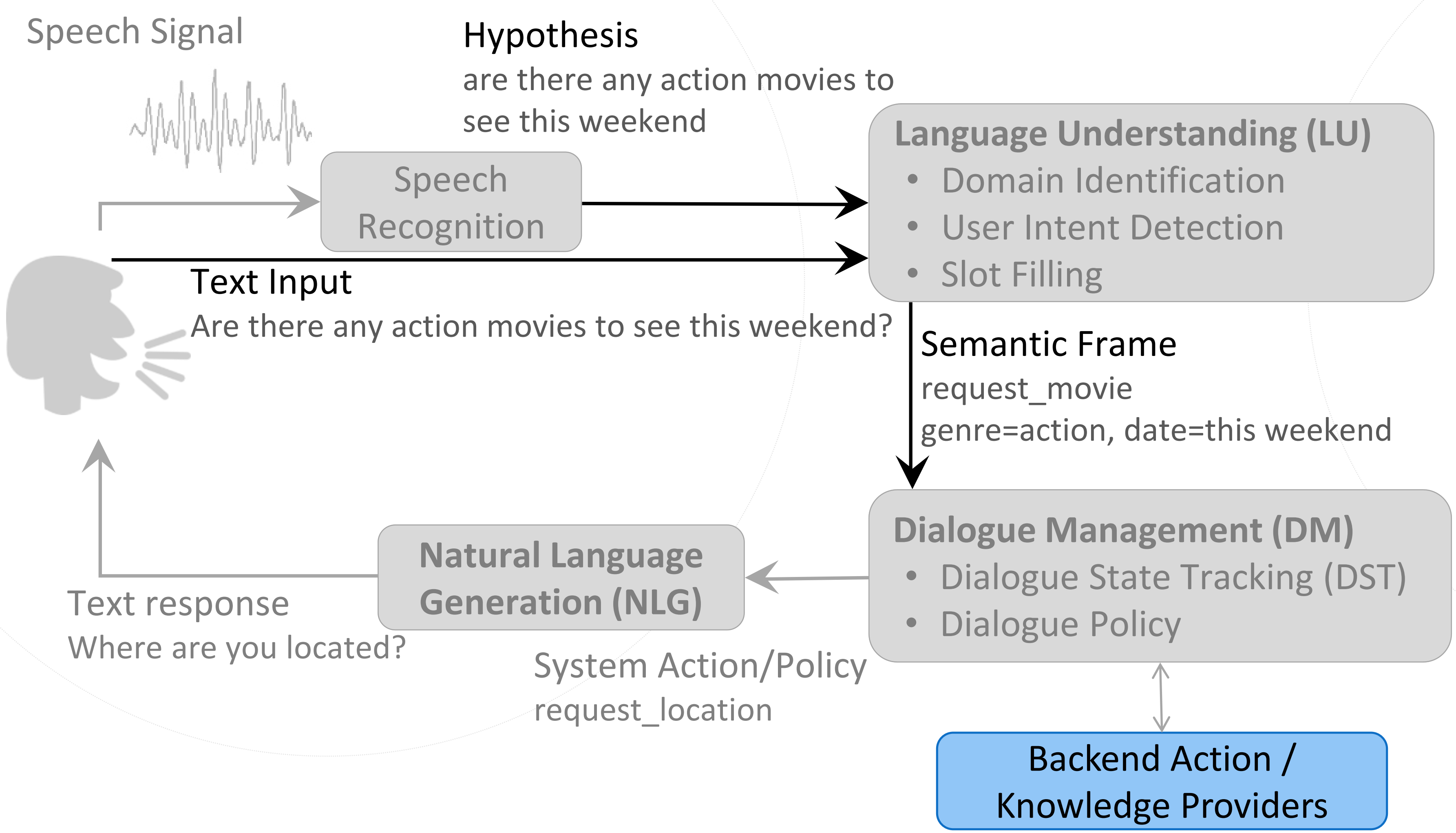
JARVIS – Iron Man's Personal Assistant

Baymax – Personal Healthcare Companion

Task-Oriented Dialogue System (Young, 2000)

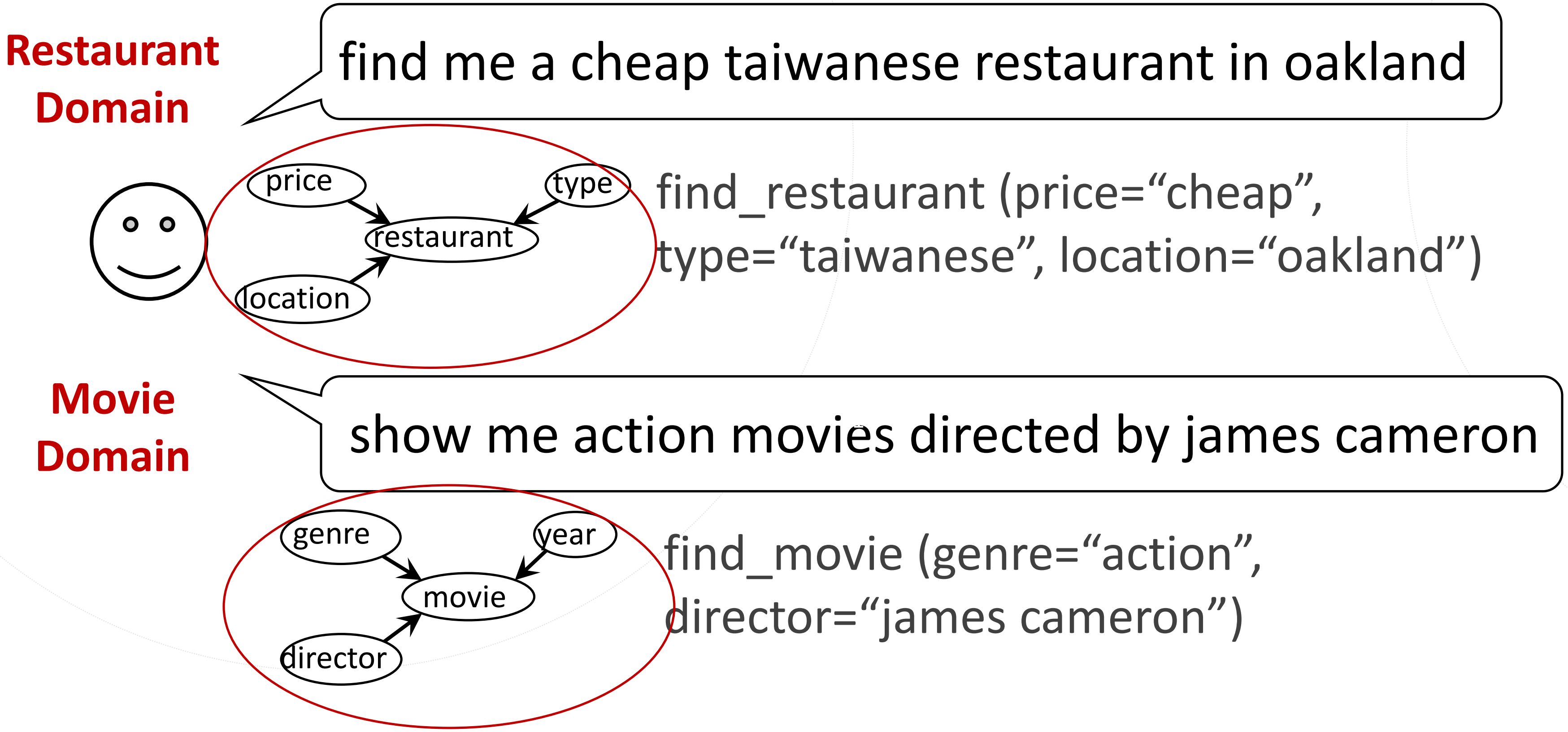


Task-Oriented Dialogue System (Young, 2000)



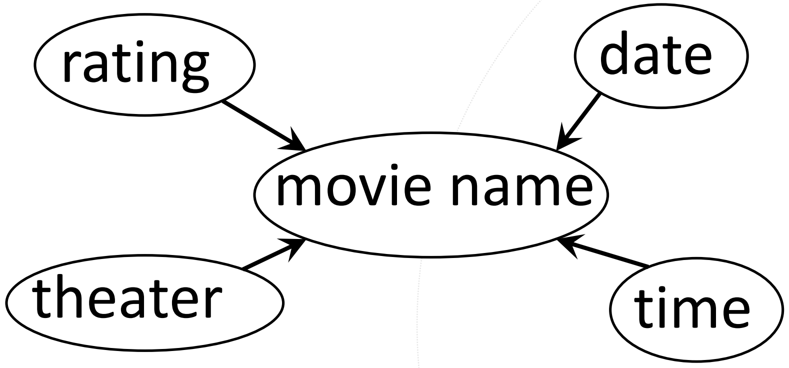
Semantic Frame Representation

- Requires a domain ontology: early connection to backend
- Contains **core content (intent, a set of slots with fillers)**



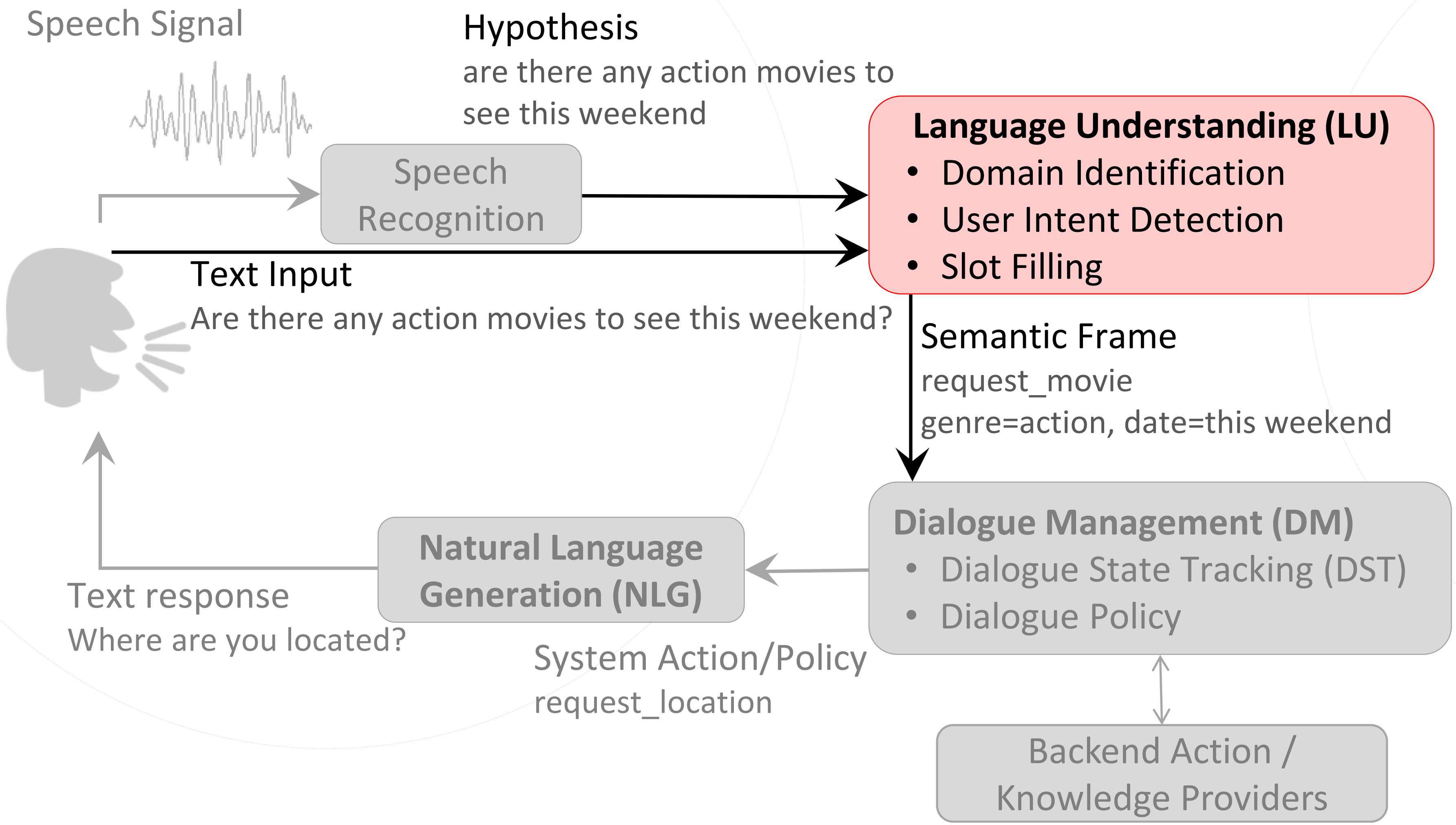
Backend Database / Ontology

- Domain-specific table
  - Target and attributes
- Functionality
  - Information access: find specific entries
  - Task completion: find the row that satisfies the constraints



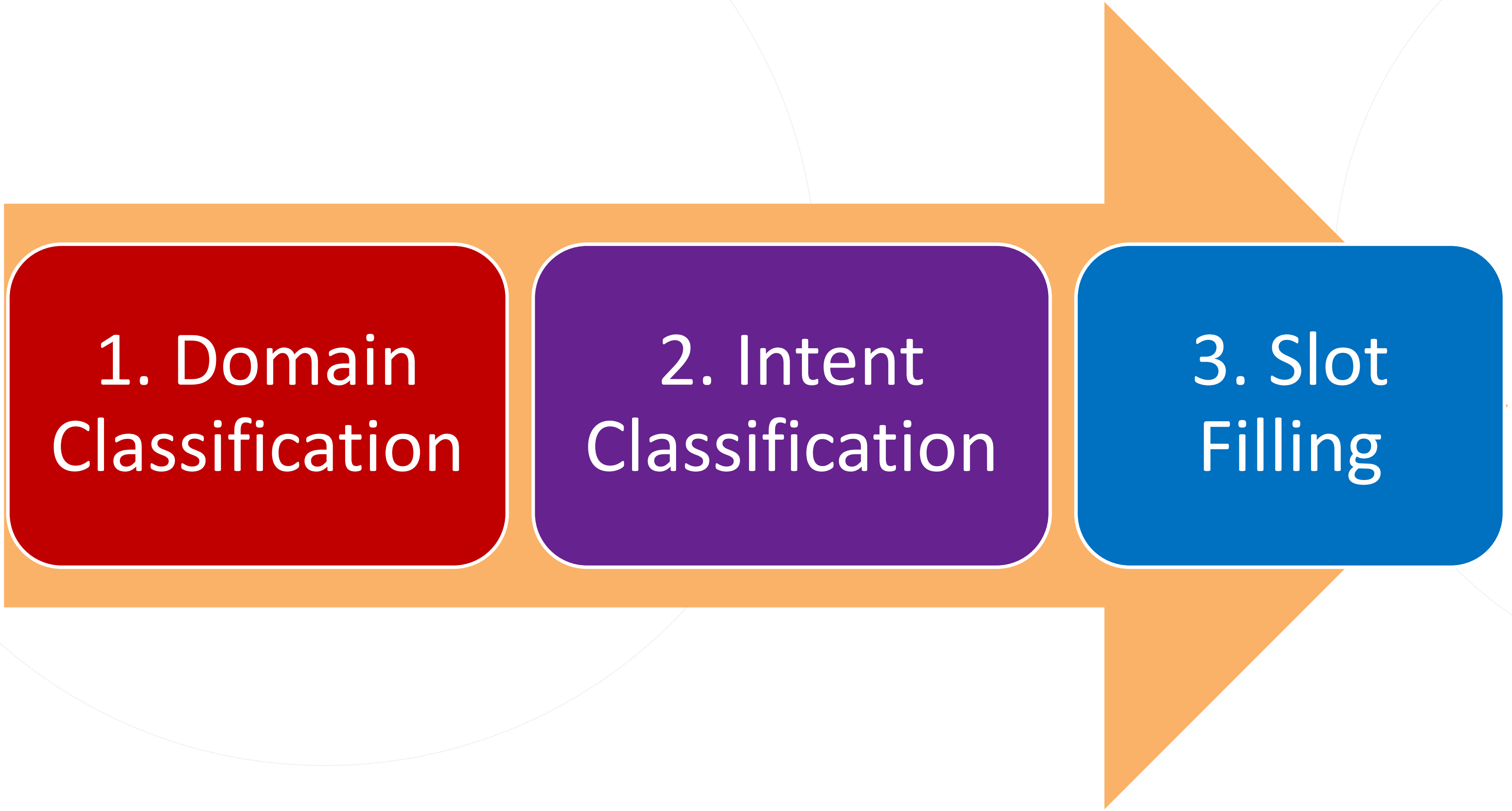
Movie Name	Theater	Rating	Date	Time
Iron Man Last	Taipei A1	8.5	2018/10/31	09:00
Iron Man Last	Taipei A1	8.5	2018/10/31	09:25
Iron Man Last	Taipei A1	8.5	2018/10/31	10:15
Iron Man Last	Taipei A1	8.5	2018/10/31	10:40

Task-Oriented Dialogue System (Young, 2000)



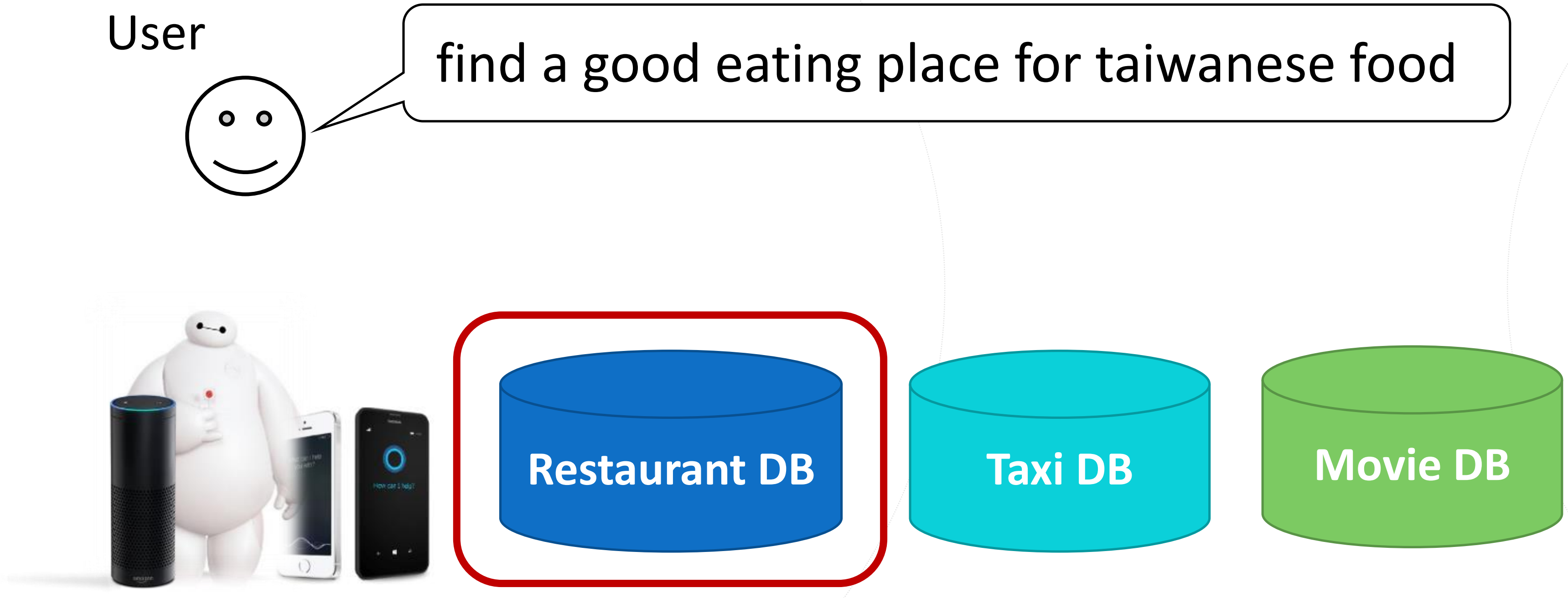
Language Understanding (LU)

- Pipelined



1. Domain Identification

Requires Predefined Domain Ontology



Intelligent Agent

Organized Domain Knowledge (Database)

Classification!

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
:	:	:

Intelligent Agent

Restaurant DB

FIND_RESTAURANT
rating=“good”
type=“taiwanese”

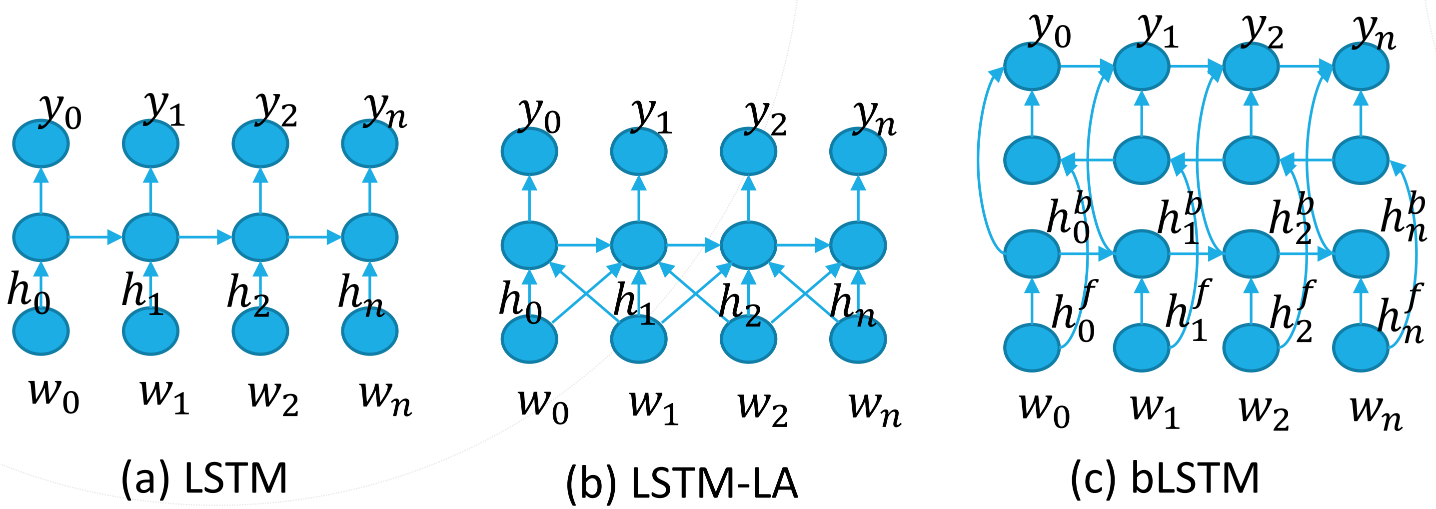
Semantic Frame

SELECT restaurant {
rest.rating=“good”
rest.type=“taiwanese”
}

Sequence Labeling

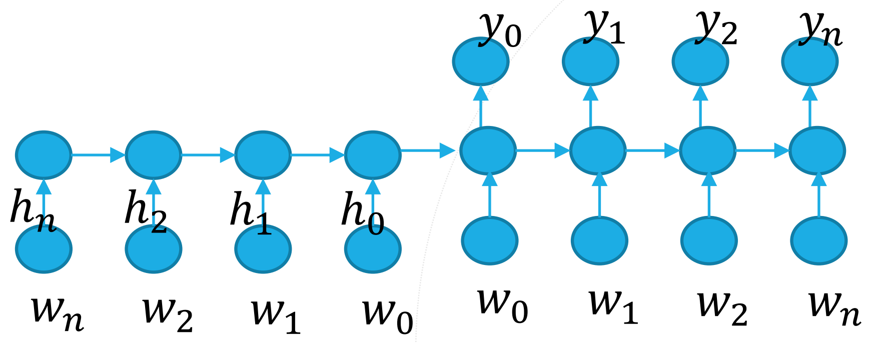
Slot Tagging (Yao+, 2013; Mesnil+, 2015)

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

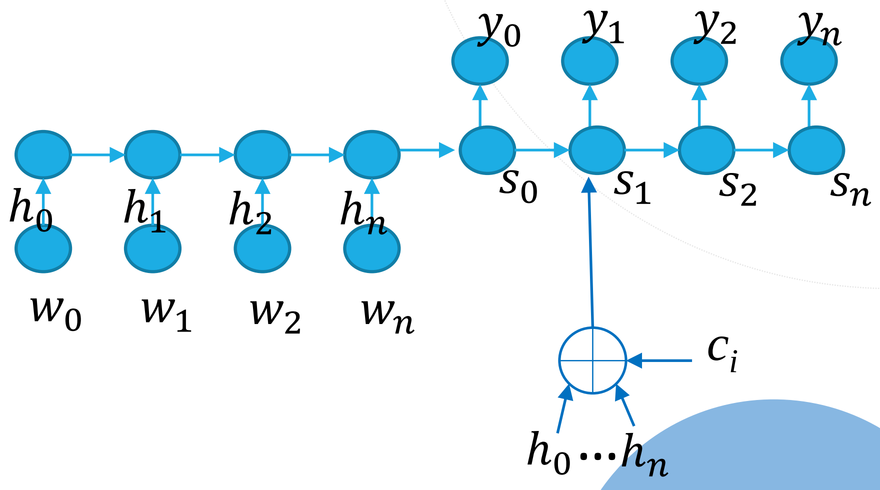


Slot Tagging (Kurata+, 2016; Simonnet+, 2015)

- Encoder-decoder networks
  - Leverages sentence level information



- Attention-based encoder-decoder
  - Use of attention (as in MT) in the encoder-decoder network
  - Attention is estimated using a feed-forward network with input: h_t and s_t at time t



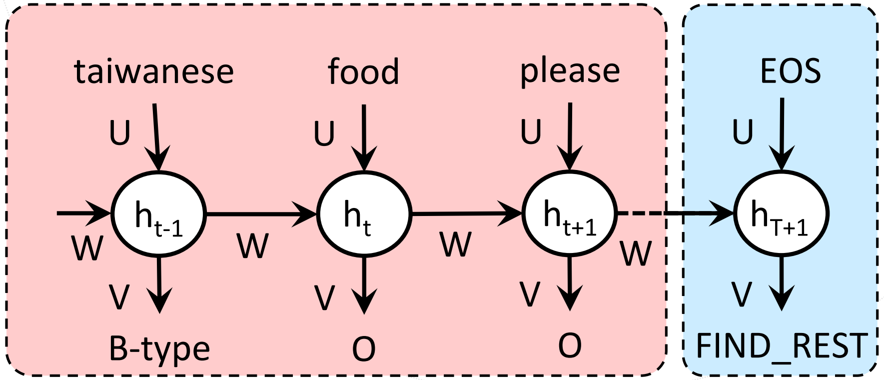
Joint Semantic Frame Parsing

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

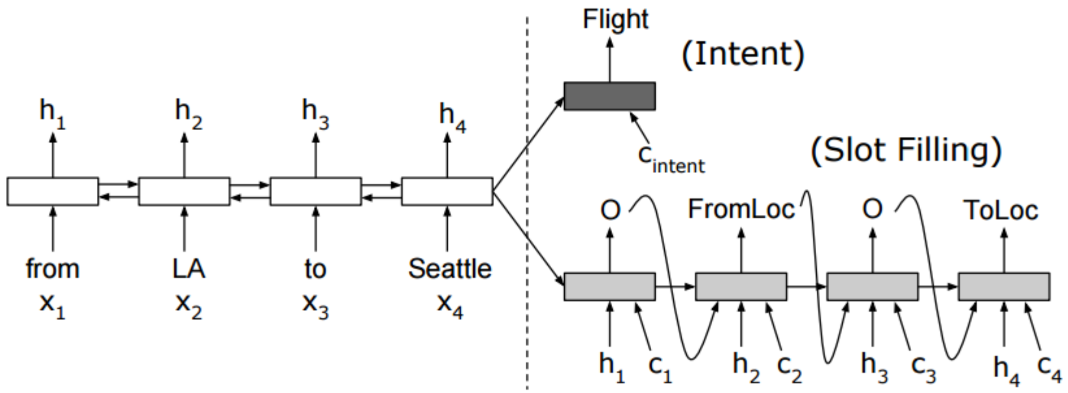
- Slot filling and intent prediction in the same output sequence

Parallel (Liu and Lane, 2016)

- Intent prediction and slot filling are performed in two branches



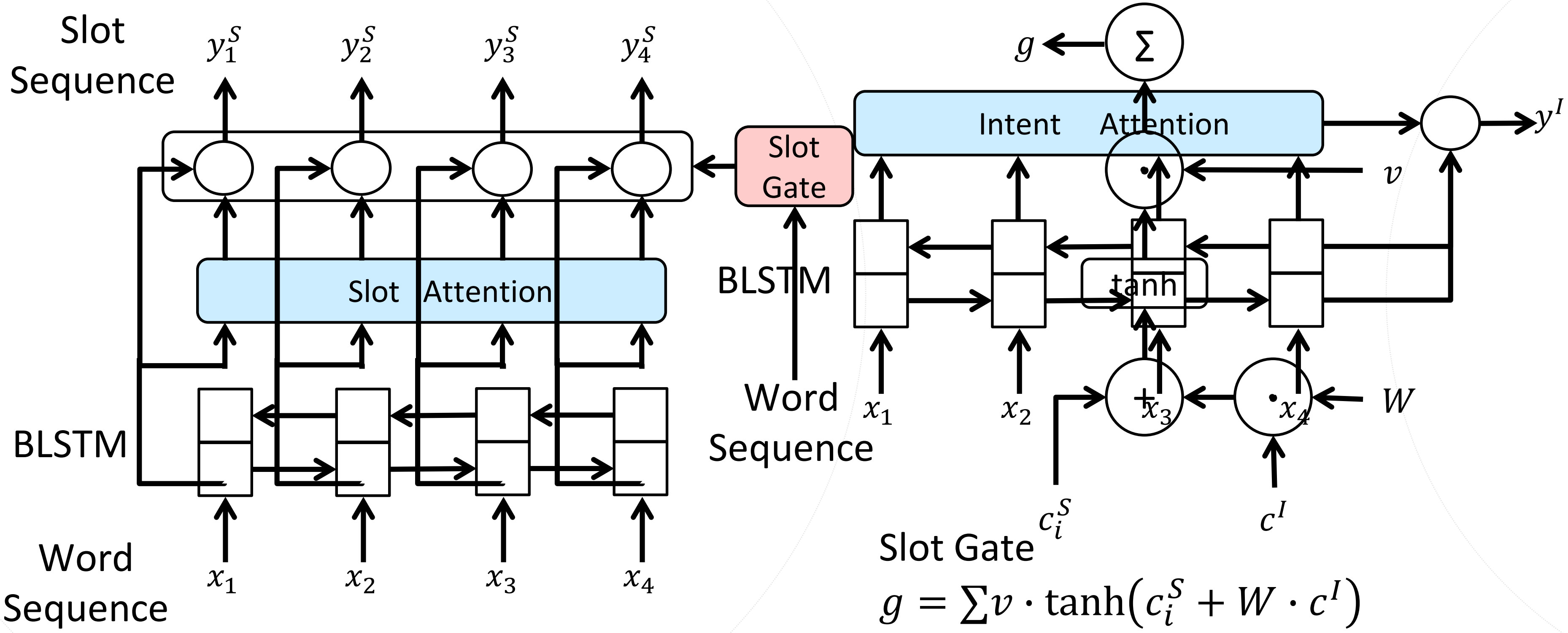
Slot Filling Intent Prediction



Joint Model Comparison

	Attention Mechanism	Intent-Slot Relationship
Joint bi-LSTM	X	Δ (Implicit)
Attentional Encoder-Decoder	√	Δ (Implicit)
Slot Gate Joint Model	√	√ (Explicit)

Slot-Gated Joint SLU (Goo+, 2018)



Slot Gate

$$g = \sum v \cdot \tanh\left(c_i^S + W \cdot c^I\right)$$

Slot Prediction

$$y_i^S = softmax\left(W^S\left(h_i + \boldsymbol{g} \cdot c_i^S\right) + b^S\right)$$

$\boldsymbol{g}$ will be larger if slot and intent are better related

Contextual LU

Domain Identification → Intent Prediction → Slot Filling

D communication *I* send_email

Single Turn

U	just	sent	email	to	bob	about	fishing	this	weekend
S	O	O	O	O	B-contact_name	O	B-subject	I-subject	I-subject

→ send_email(contact_name="bob", subject="fishing this weekend")

Multi-Turn

U_1	send	email	to	bob
S_1				B-contact_name

→ send_email(contact_name="bob")

U_2	are	we	going	to	fish	this	weekend
S_2	B-message	I-message	I-message	I-message	I-message	I-message	I-message

→ send_email(message="are we going to fish this weekend")

Contextual LU

- 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

E2E MemNN for Contextual LU (Chen+, 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

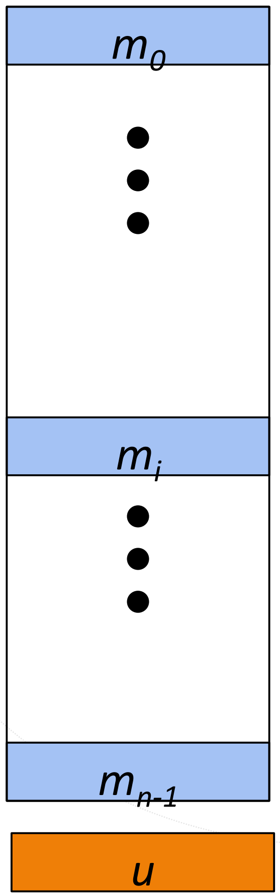
U: "3 tickets for saturday"

S: "What time would you like ?"

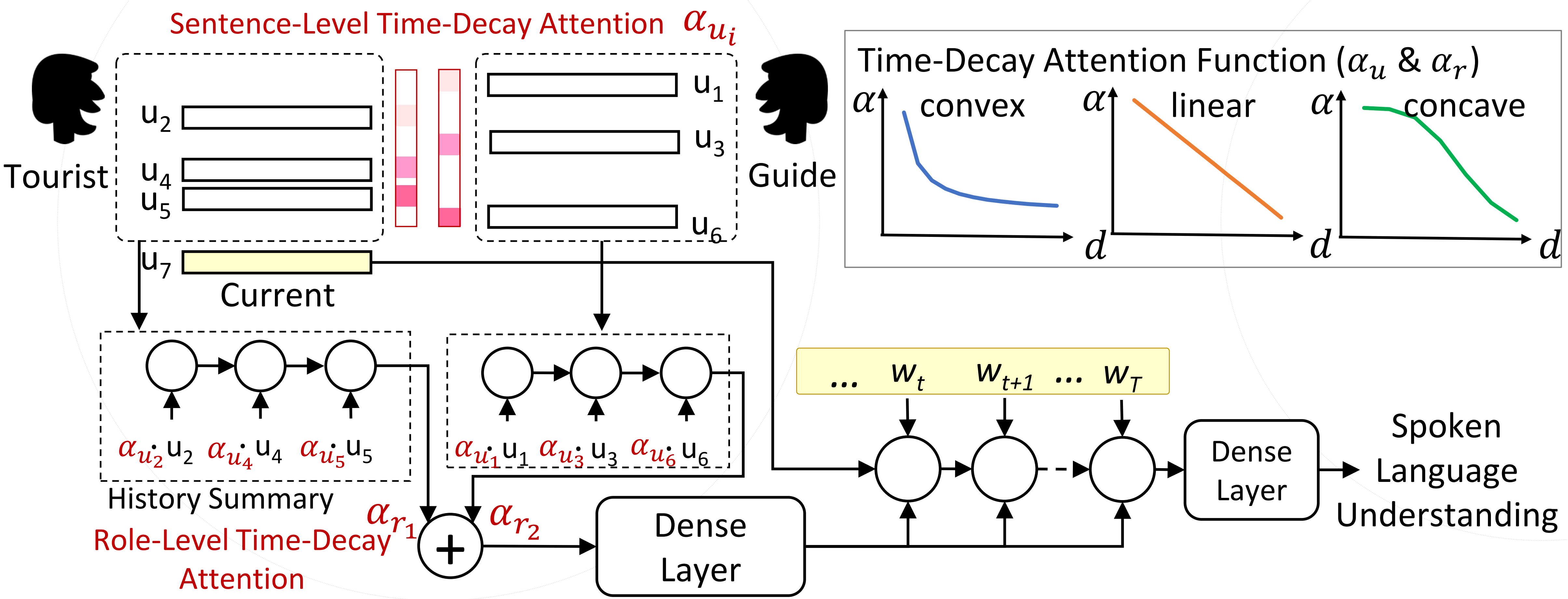
U: "Any time on saturday is fine"

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

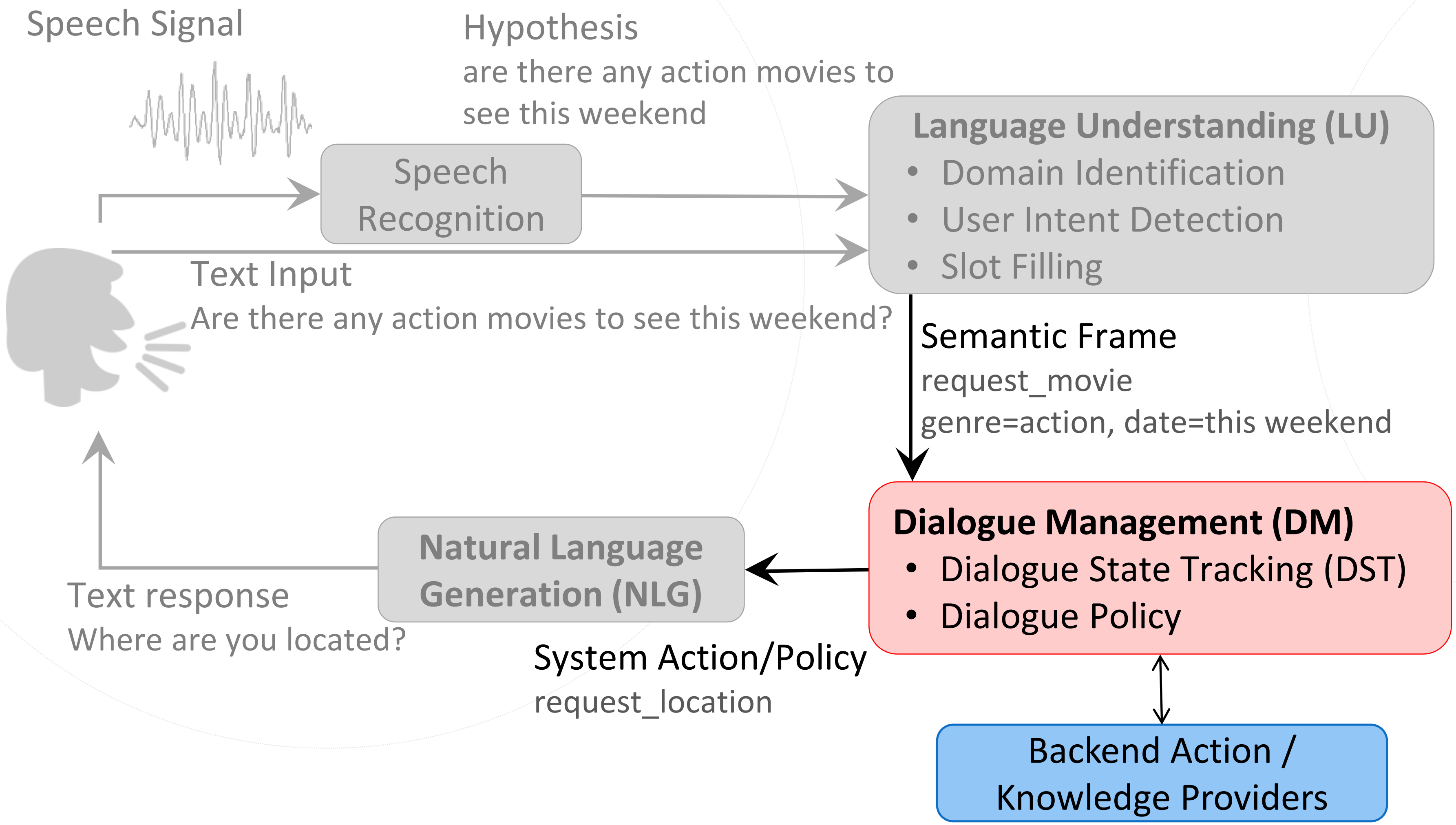
U: "Let's do 5:40"



Time-Aware & Role-Based Attention (Su+, 2018)



Task-Oriented Dialogue System (Young, 2000)



Dialogue State Tracking

Utterance	Dialogue act
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 ()

DNN for DST

Hello, how may I help you?

I'm looking for a Thai restaurant. — inform(type=restaurant, food=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.

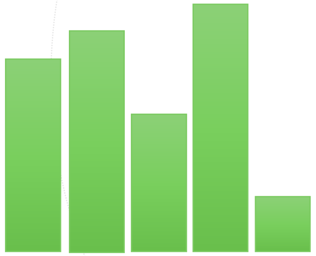
Thank you, bye. — bye()

multi-turn conversation

feature extraction

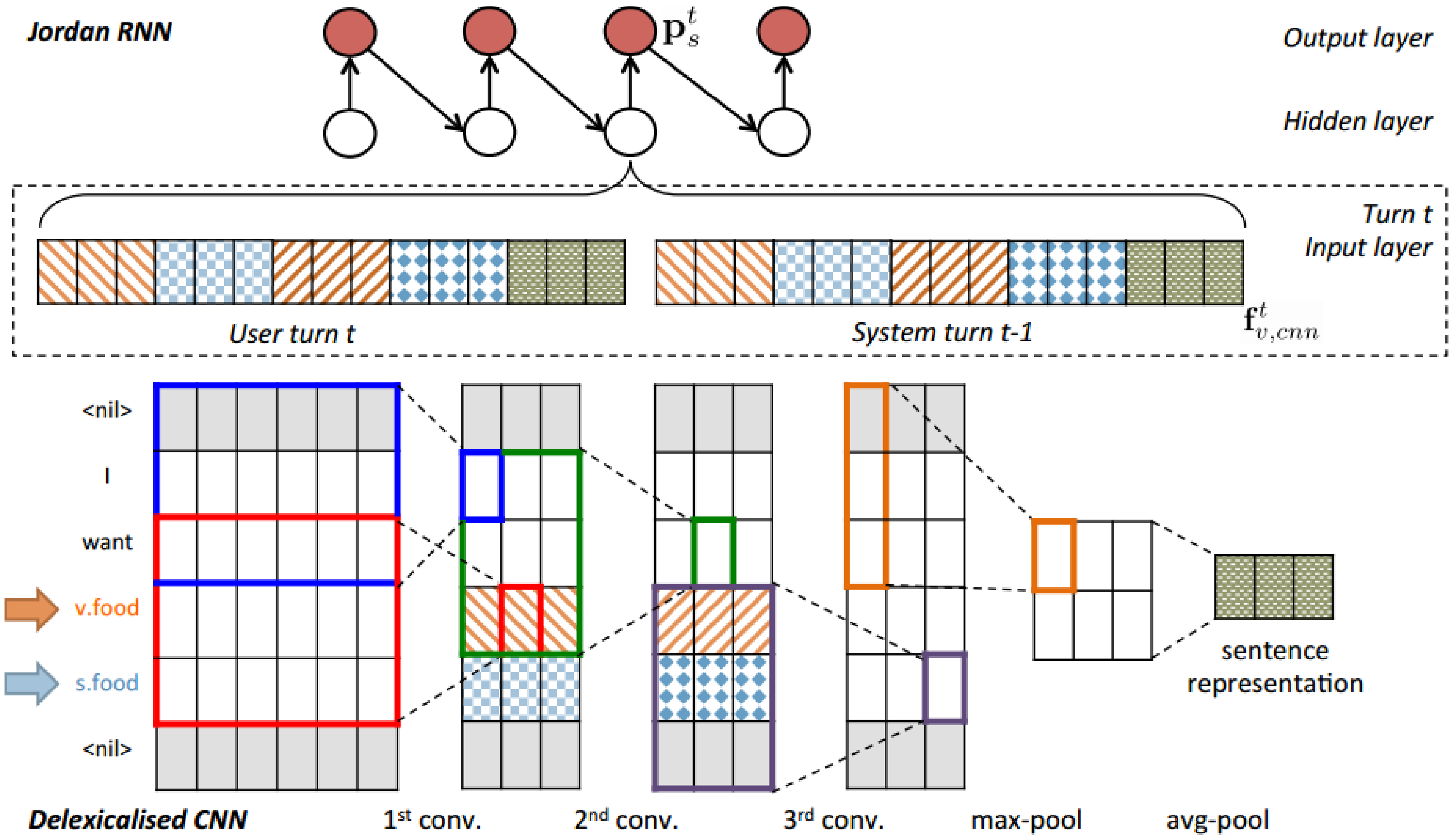
DNN

state of this turn



A slot value distribution for each slot

RNN-CNN DST (Mrkšić+, 2015)



(Figure from Wen et al, 2016)

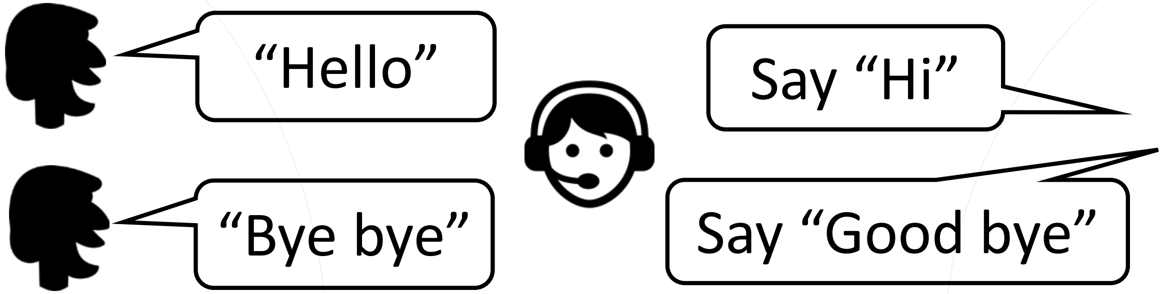
Dialogue Policy Optimization

Utterance	Dialogue act
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 ()

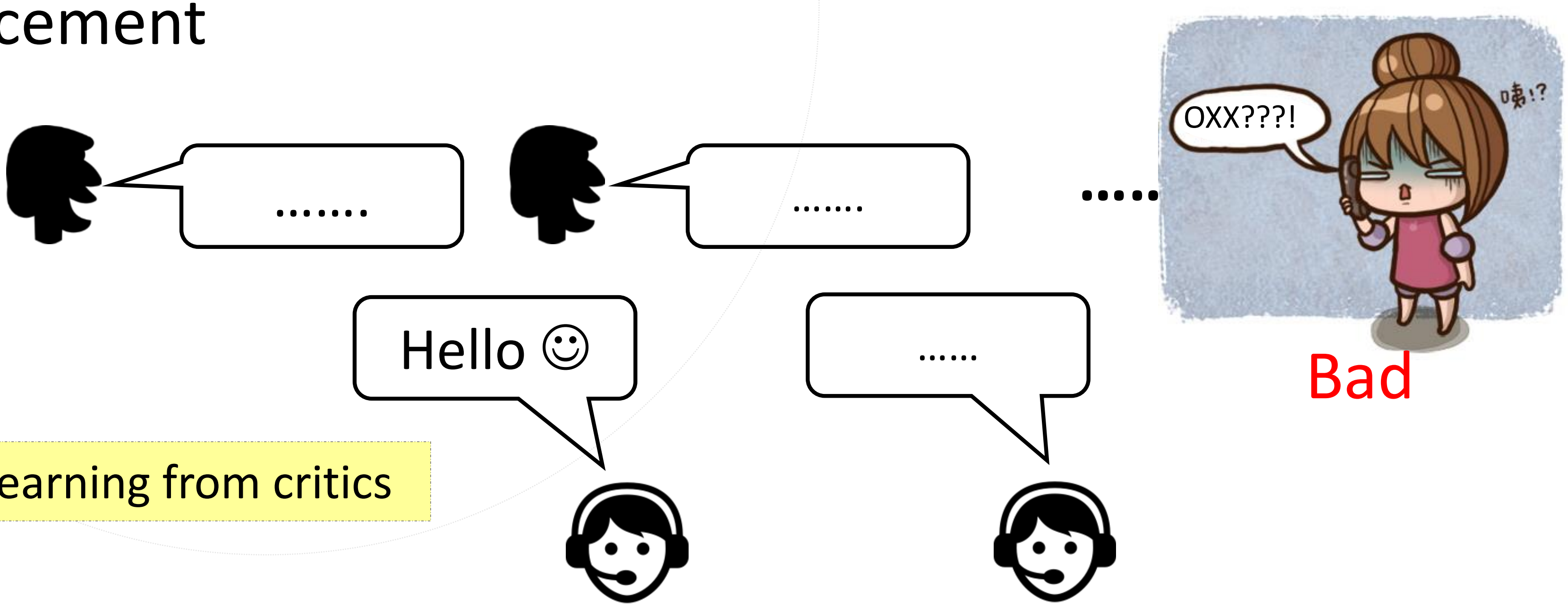
Supervised v.s. Reinforcement

- Supervised

Learning from teacher



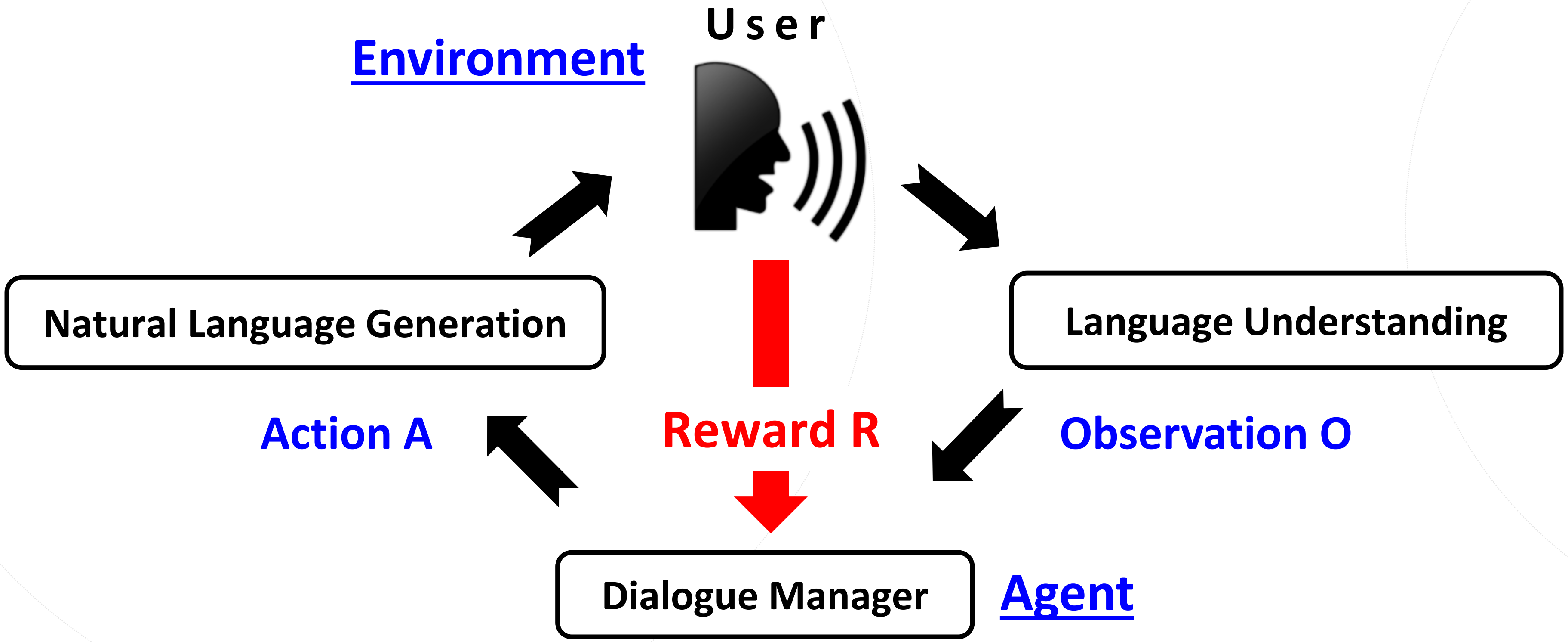
- Reinforcement



Learning from critics

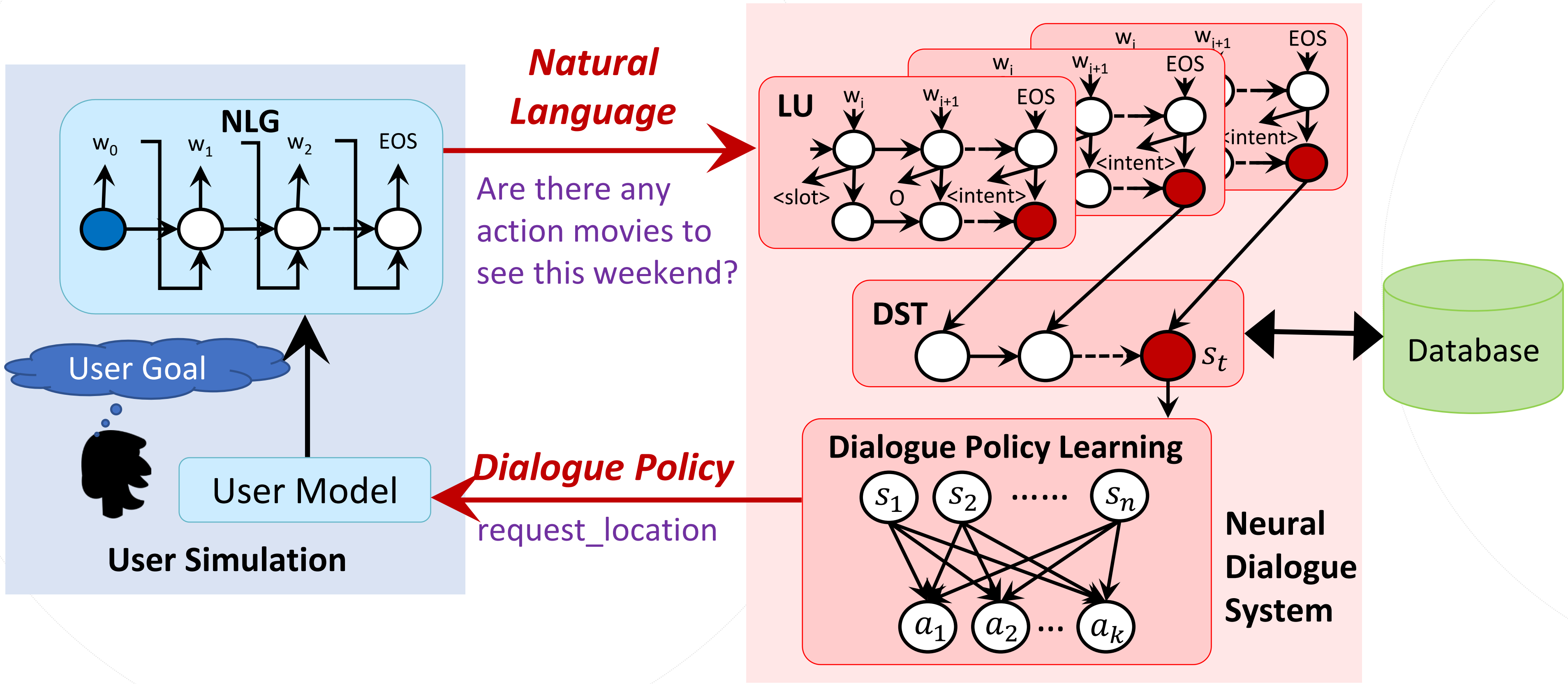
Dialogue Policy Optimization

- Dialogue management in a RL framework



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

E2E Task-Completion Bot (TC-Bot) (Li+, 2017)



Idea: SL for each component and RL for end-to-end training

Reinforcement Learning Agent (Li+, 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, thanks.
Agent: Thanks!

Issue 1: no notion about what requests can be skipped

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!

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

Reinforcement Learning Agent (Li+, 2017)

User goal – ***Two tickets for "Hail Caesar" tomorrow 7:15PM at AMC Lowes Oak tree 6, Seattle.***

RULE BASED SYSTEM

User: Do you any ticket for movie 'Hail Caesar' at theater AMC Lowes Oak Tree 6?
Agent: What start time do you prefer?
User: **7:15 pm** Tomorrow
Agent: Which city you are looking for?
User: Seattle
Agent: How many tickets do you want?
User: Two
Agent: Okay, I have issued two tickets for movie 'Hail Caeser' tomorrow **7:50 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Oh, no.
Agent: Thanks!

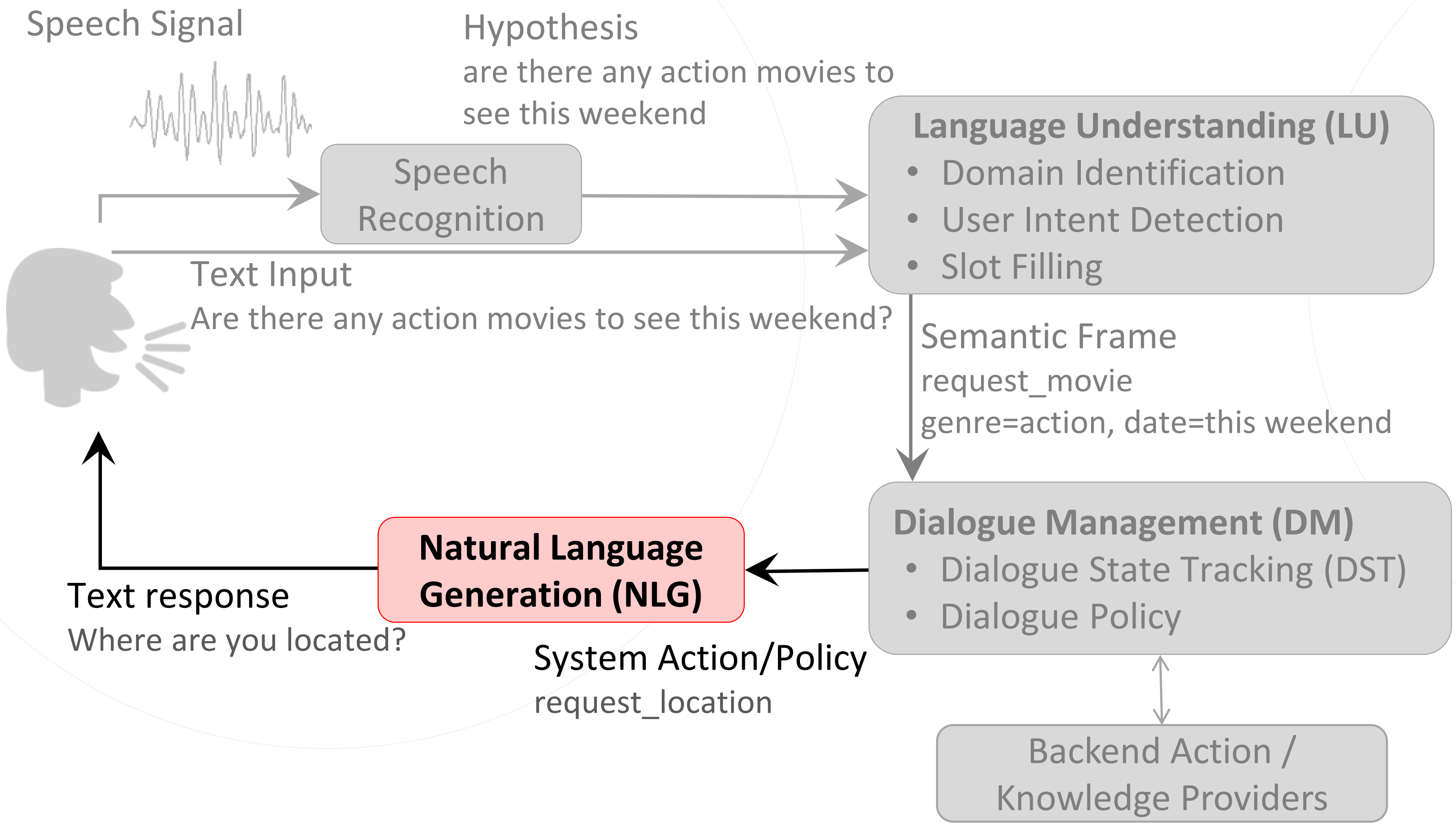
Issue 2: no principle way to deal with LU errors/uncertainty

REINFORCEMENT LEARNING SYSTEM

User: Do you any ticket for movie 'Hail Caesar' at theater AMC Lowes Oak Tree 6?
Agent: What start time do you prefer?
User: **7:15 pm** Tomorrow
Agent: Which city you are looking for?
User: Seattle
Agent: Okay, I will issue you tickets for movie 'Hail Caeser' tomorrow **7:50 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Oh, no.
Agent: **It is 7:50 pm, right?**
User: Two tickets for 7:15 pm, please.
Agent: Okay, I will issue you two tickets for movie 'Hail Caeser' tomorrow **7:15 pm** at theater AMC Lowes Oak Tree 6, Seattle.
User: Thanks!
Agent: Thanks!

Automatically learn when and what to confirm for better robustness

Task-Oriented Dialogue System (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

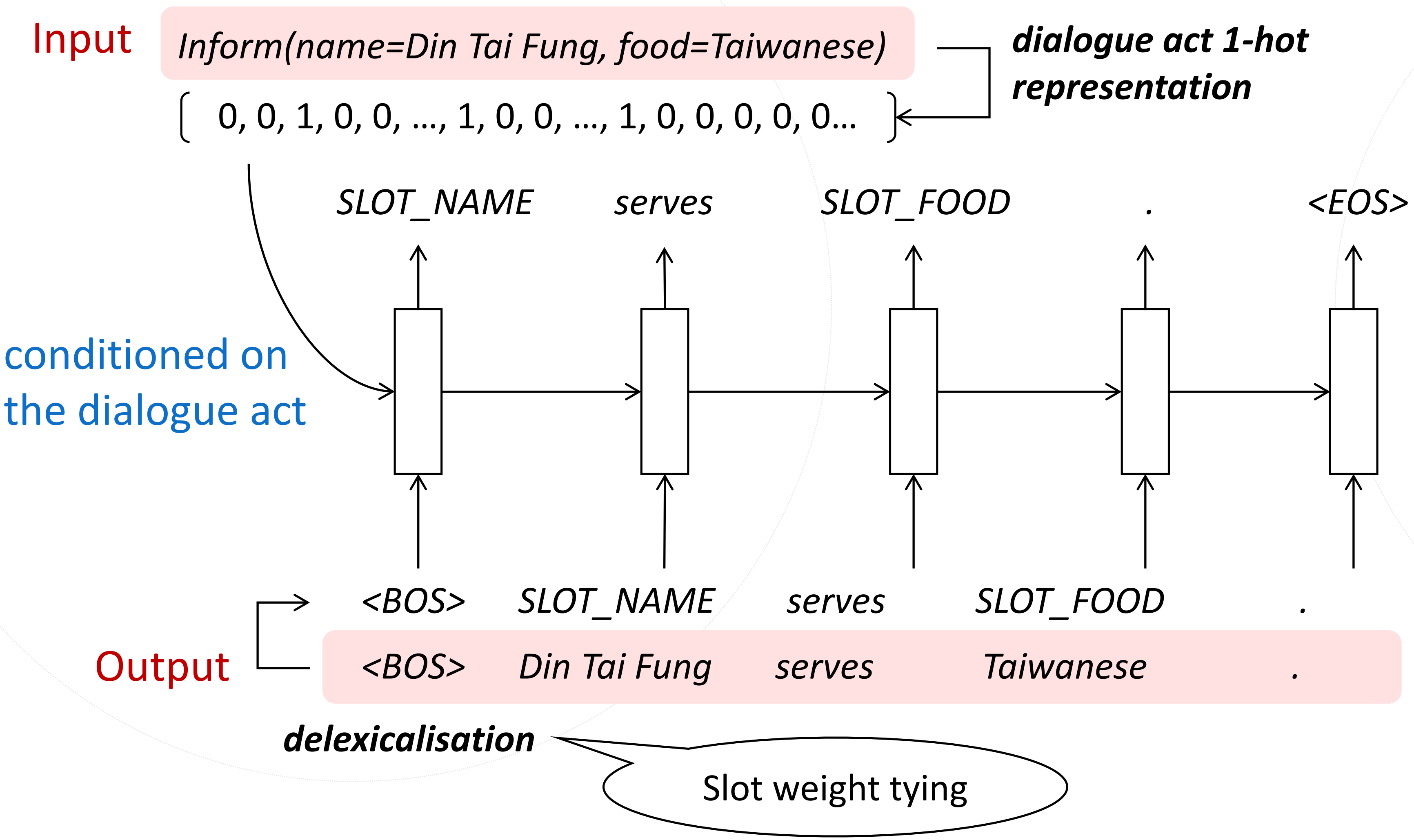
Template-Based NLG

- Define a set of rules to map frames to NL

Semantic Frame	Natural Language
confirm()	"Please tell me more about the product your 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, poor scalability

RNN-Based LM NLG (Wen+, 2015)



Handling Semantic Repetition

- 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.
- Deficiency in either model or decoding (or both)
- Mitigation
  - Post-processing rules (Oh & Rudnicky, 2000)
  - Gating mechanism (Wen et al., 2015)
  - Attention (Mei et al., 2016; Wen et al., 2015)

Semantic Conditioned LSTM (Wen+, 2015)

- Original LSTM cell

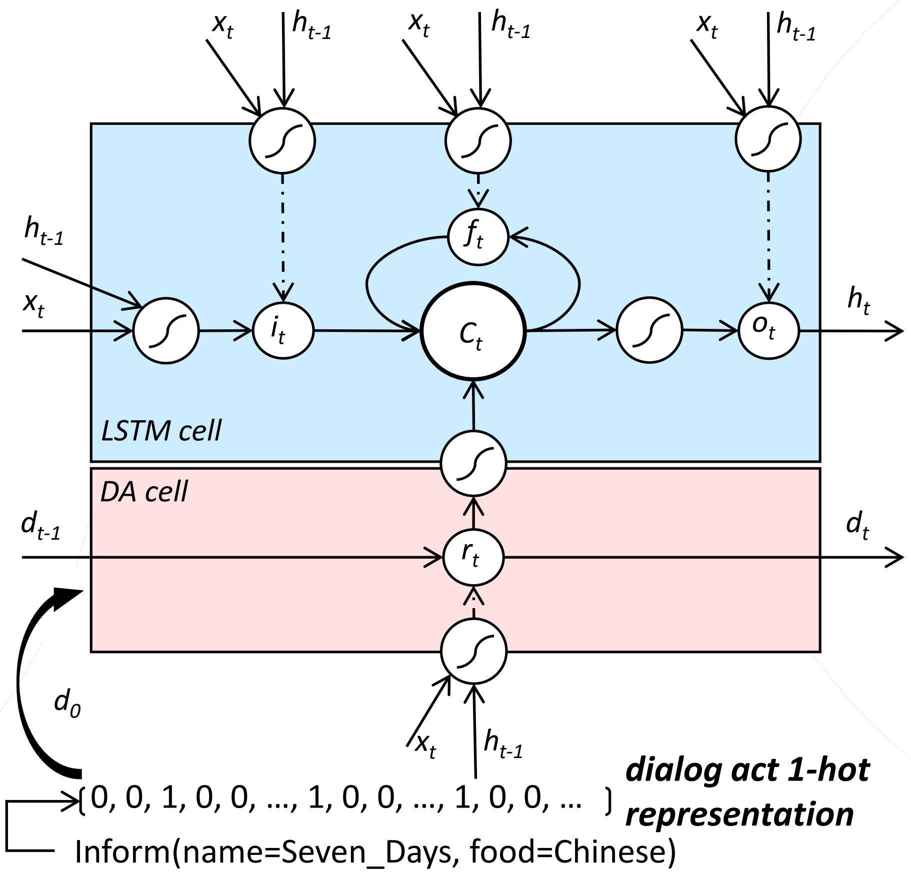
$$\mathbf{i}_t = \sigma(\mathbf{W}_{wi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1})$$
$$\mathbf{f}_t = \sigma(\mathbf{W}_{wf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1})$$
$$\mathbf{o}_t = \sigma(\mathbf{W}_{wo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1})$$
$$\hat{\mathbf{c}}_t = \tanh(\mathbf{W}_{wc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1})$$
$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t$$
$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t)$$

- Dialogue act (DA) cell

$$\mathbf{r}_t = \sigma(\mathbf{W}_{wr}\mathbf{x}_t + \mathbf{W}_{hr}\mathbf{h}_{t-1})$$
$$\mathbf{d}_t = \mathbf{r}_t \odot \mathbf{d}_{t-1}$$

- Modify Ct

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t + \tanh(\mathbf{W}_{dc}\mathbf{d}_t)$$

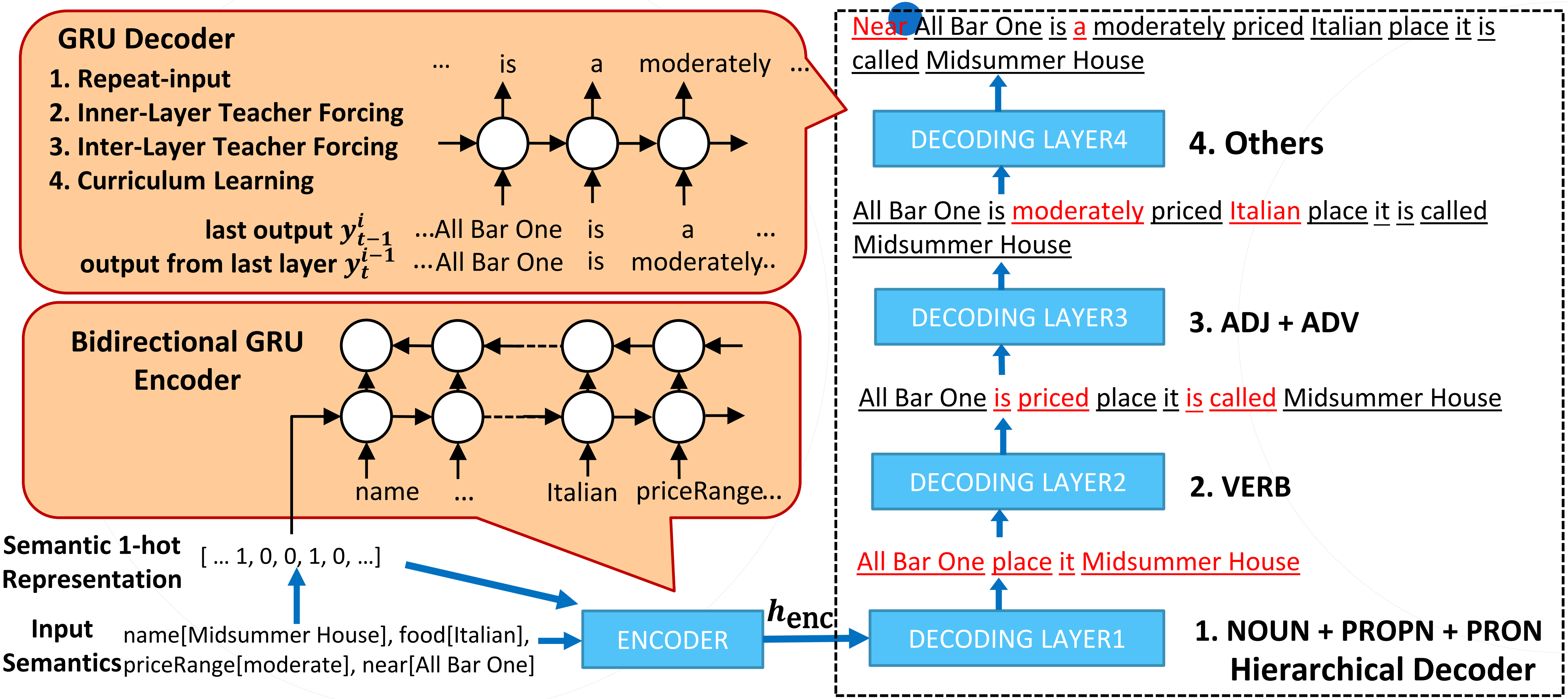


Idea: using gate mechanism to control the generated semantics (dialogue act/slots)

Issues in NLG

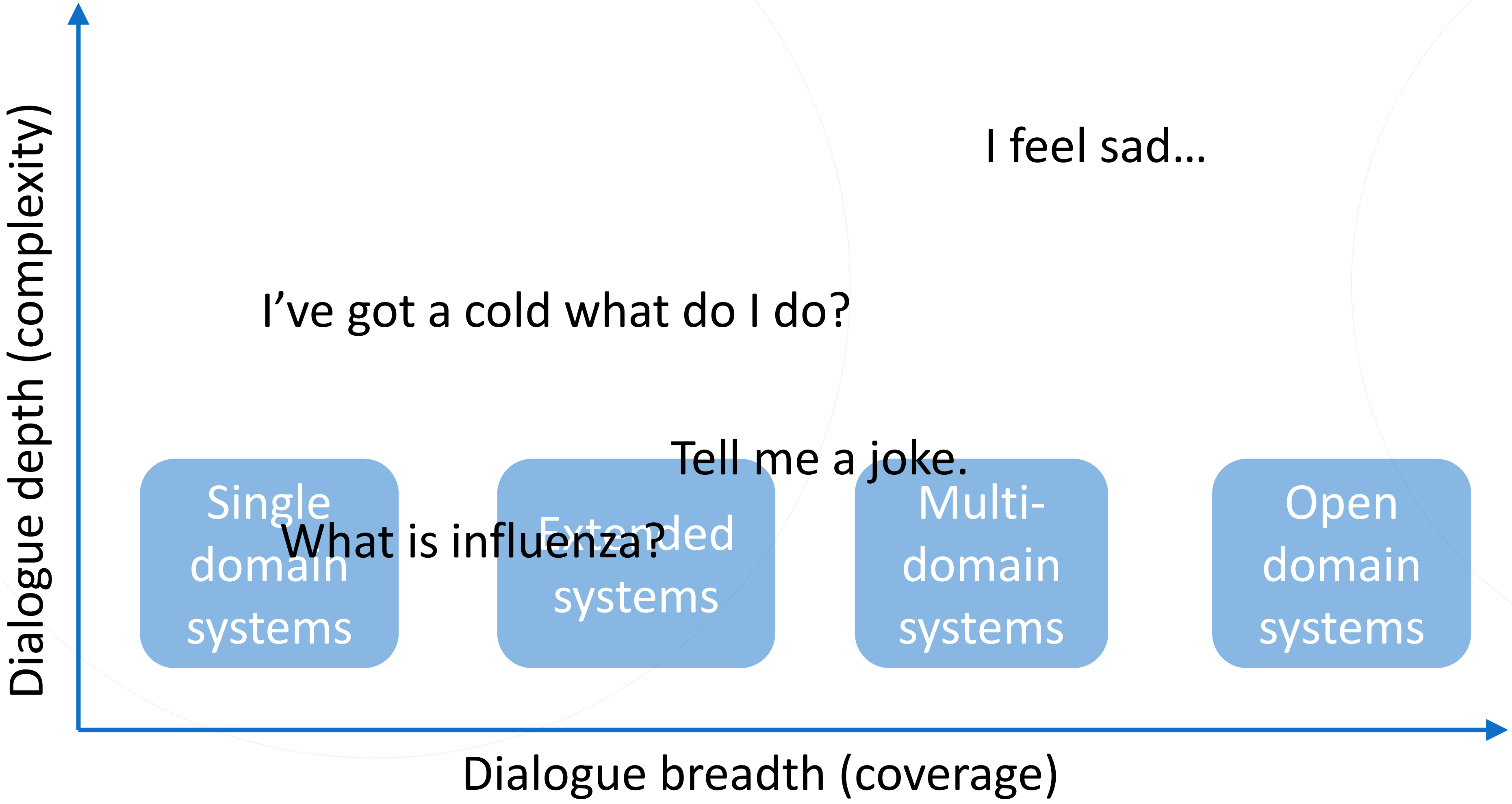
- Issue
  - NLG tends to generate shorter sentences
  - NLG may generate grammatically-incorrect sentences
- Solution
  - Generate word patterns in a order
  - Consider linguistic patterns

Hierarchical NLG w/ Linguistic Patterns (Su+, 2018)

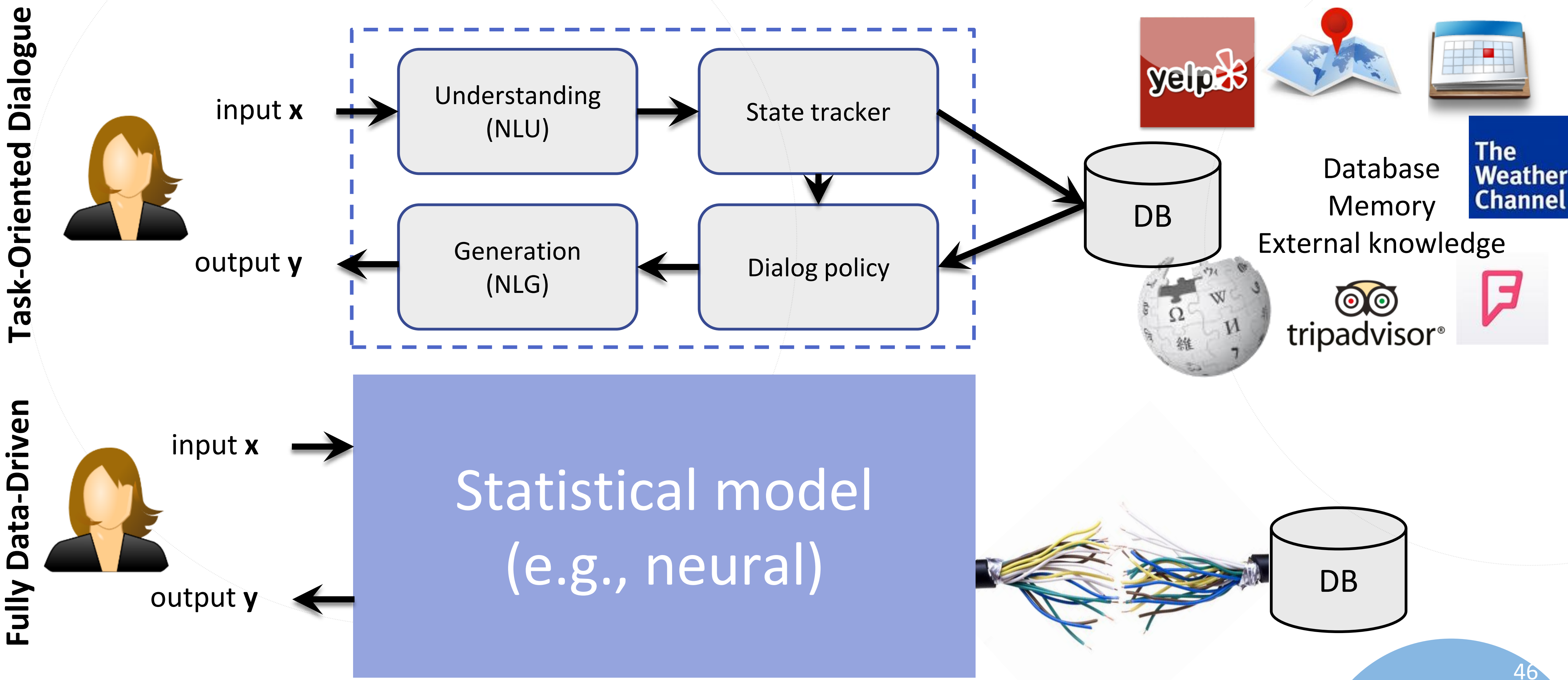


Idea: gradually generate words based on the linguistic knowledge

Evolution Roadmap



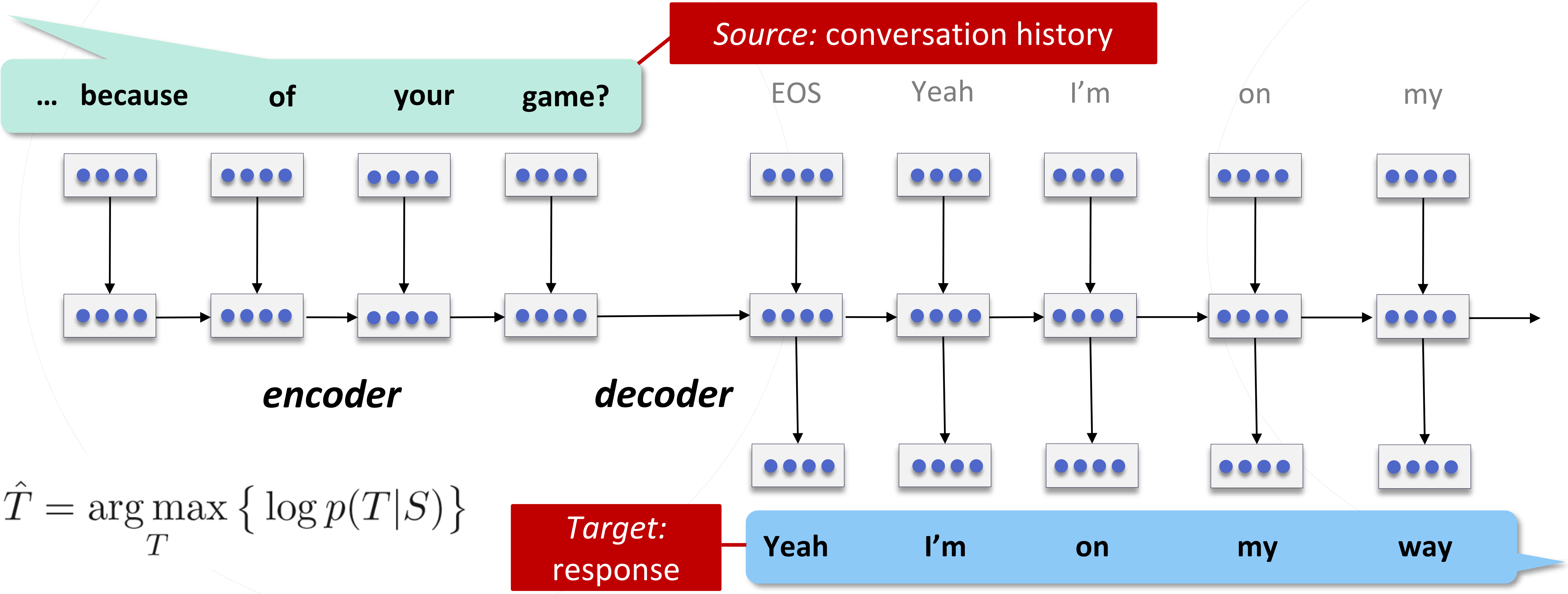
Dialogue Systems



Chit-Chat Social Bots

Non-Task-Oriented

Neural Response Generation (Sordoni+, 2015; Vinyals & Le, 2015)



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

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

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+, 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?

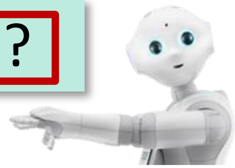
I think he'd be a good signing.

‘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+, 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.

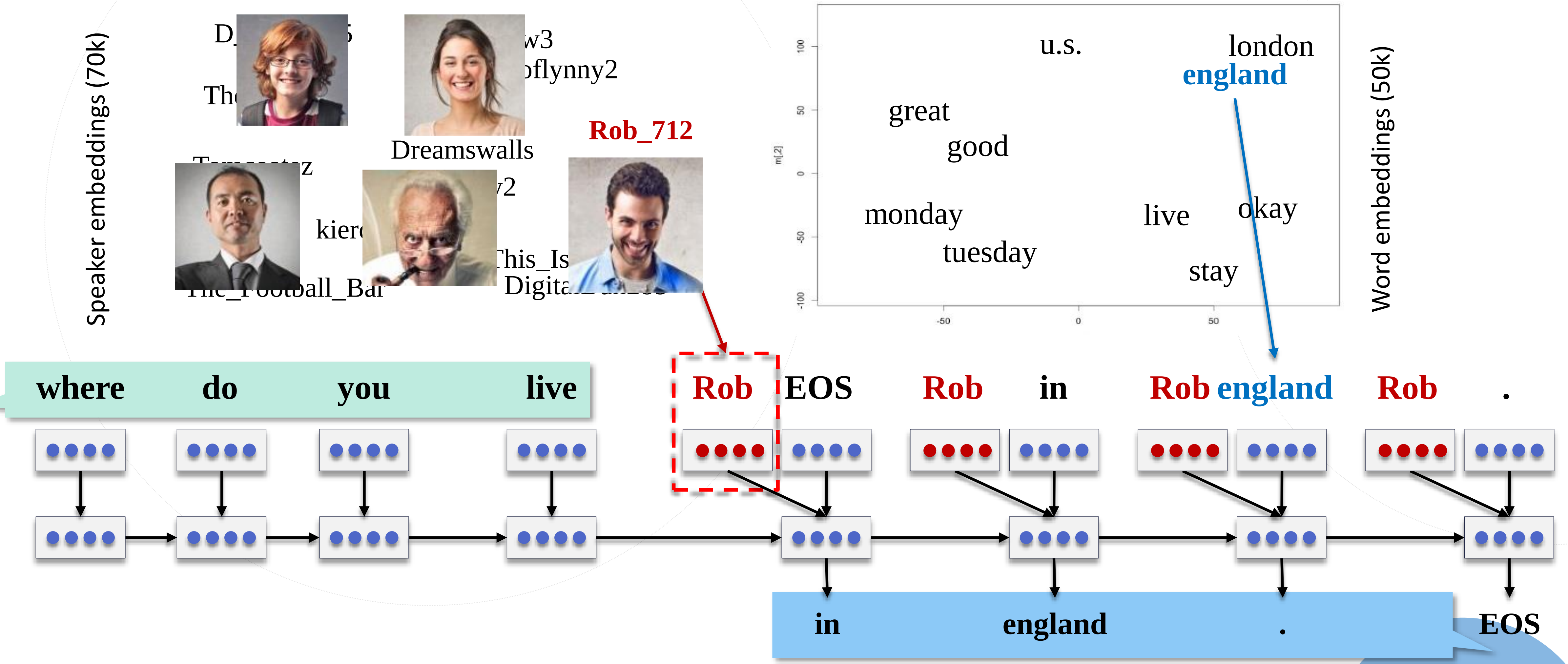
I've never seen him live.

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+, 2016)



Persona Model for Speaker Consistency (Li+, 2016)

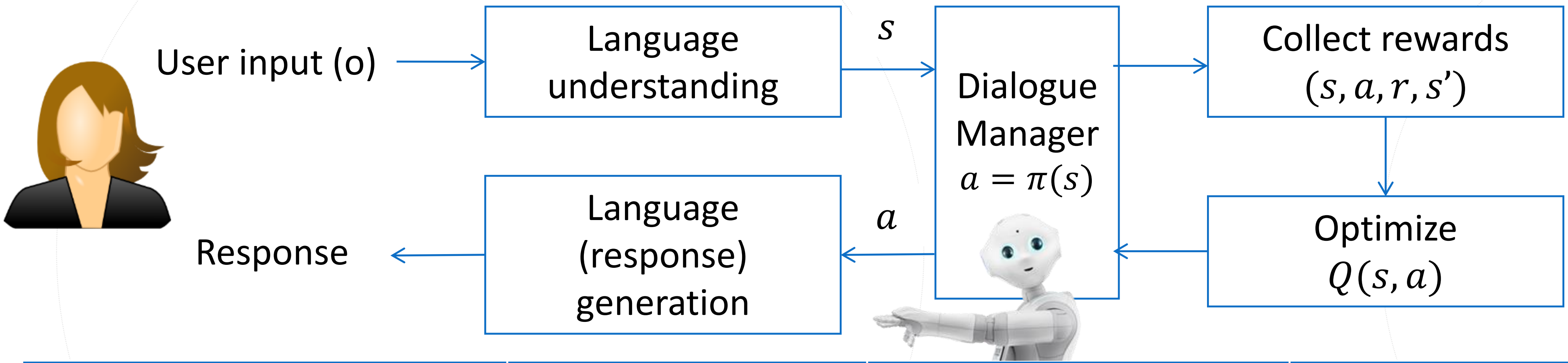
Baseline model → inconsistency

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+, 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+, 2015; Li+, 2016)

Neural model learns the general shape of conversations, and the system output is situationally appropriate and coherent.

No grounding into a real calendar, but the "shape" of the conversation is fluent and plausible.

H: hi
M: how are you?
H: not bad
M: what's wrong?
H: nothing really
M: wanna talk about it? i 'm here if you wanna talk
H: well, i'm just tired
M: me too, i'm here if you wanna talk

H: would thursday afternoon work sometime?
M: yeah , sure . just let me know when you're free.
H: after lunch is probably the best time
M: okay, sounds good . just let me know when you're free.
H: would 2 pm work for you?
M: works for me.
H: well let's say 2 pm then i 'll see you there
M: sounds good.

NTU MIULAB

Chit-Chat v.s. Task-Oriented

Social Chat
Engaging, Human-Like Interaction
(Ungrounded)

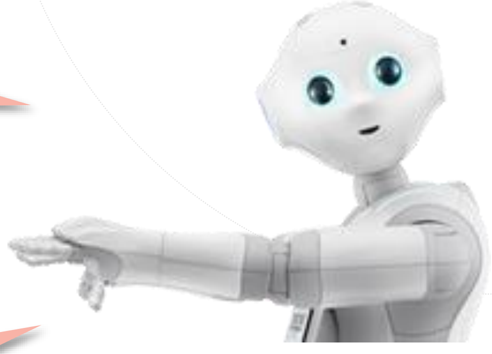
Task-Oriented
Task Completion, Decision Support
(Grounded)

The weather is so depressing these days.

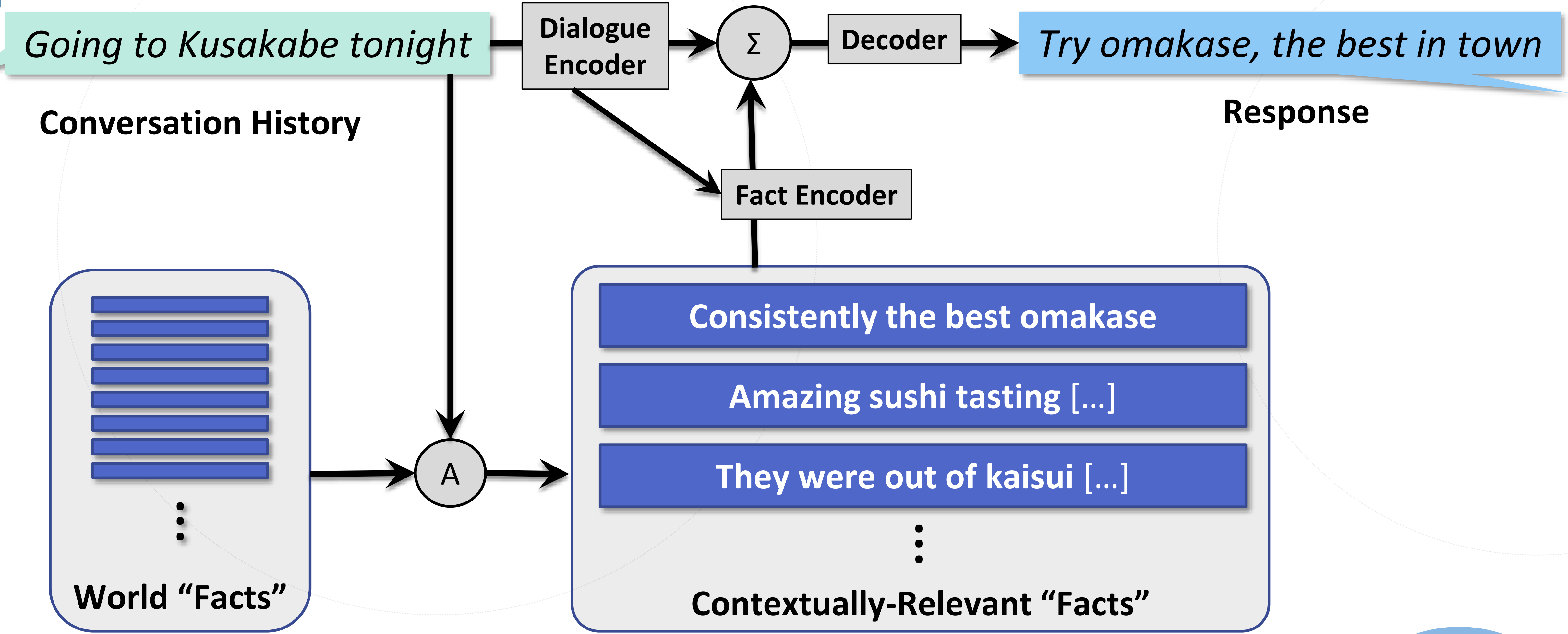
I know, I dislike rain too.
What about a day trip to eastern Washington?

Any recommendation?

Try Dry Falls, it's spectacular!



Knowledge-Grounded Responses (Ghazvininejad+, 2017)



Conversation and Non-Conversation Data

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

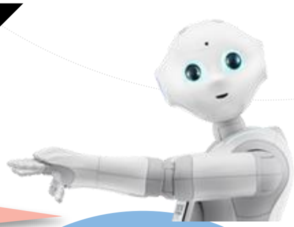
*Try **C**, one of the best **D** in the city.*

Conversation Data

Knowledge Resource

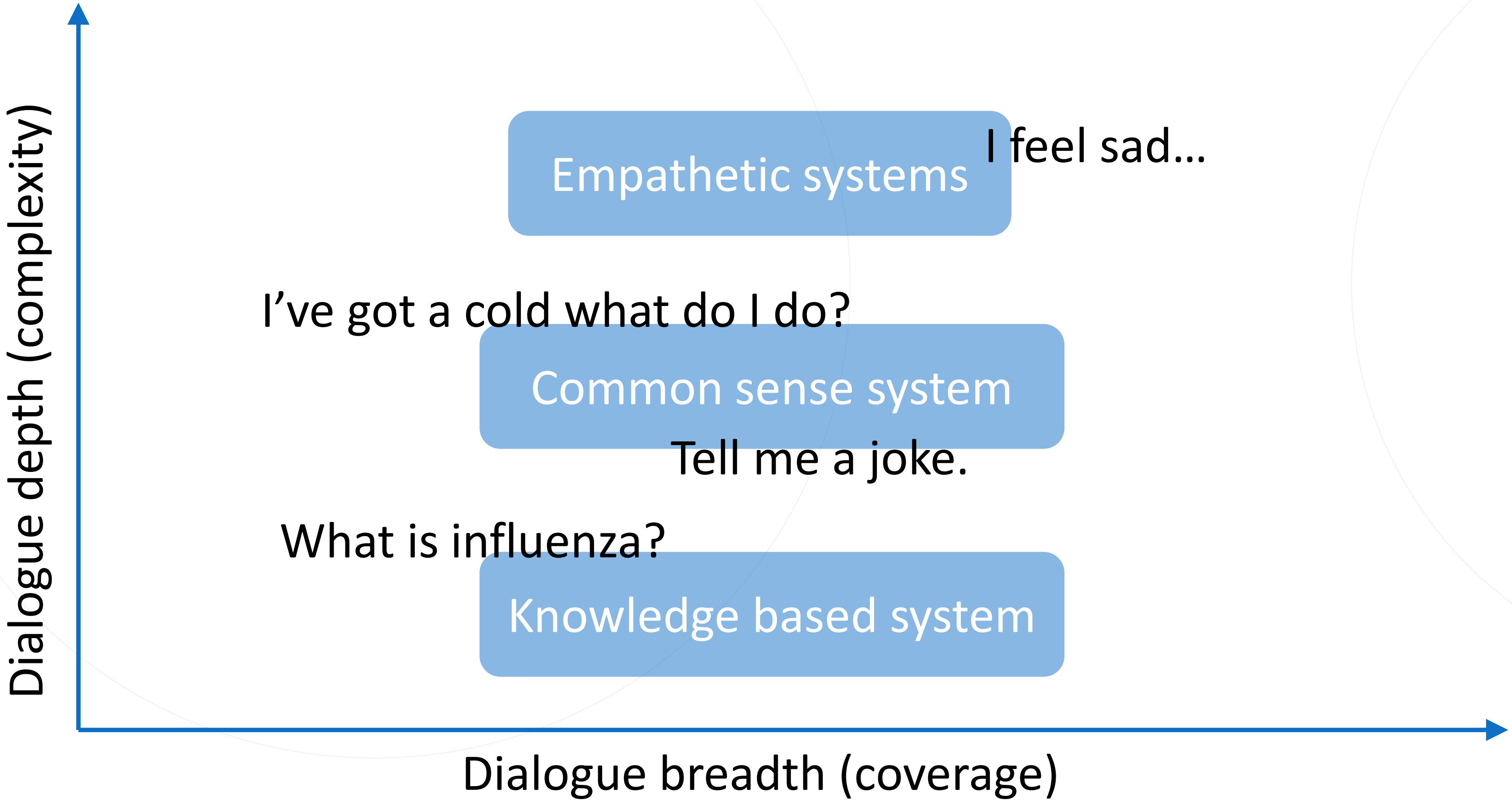
*You know any good **Japanese** restaurant in **Seattle**?*

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



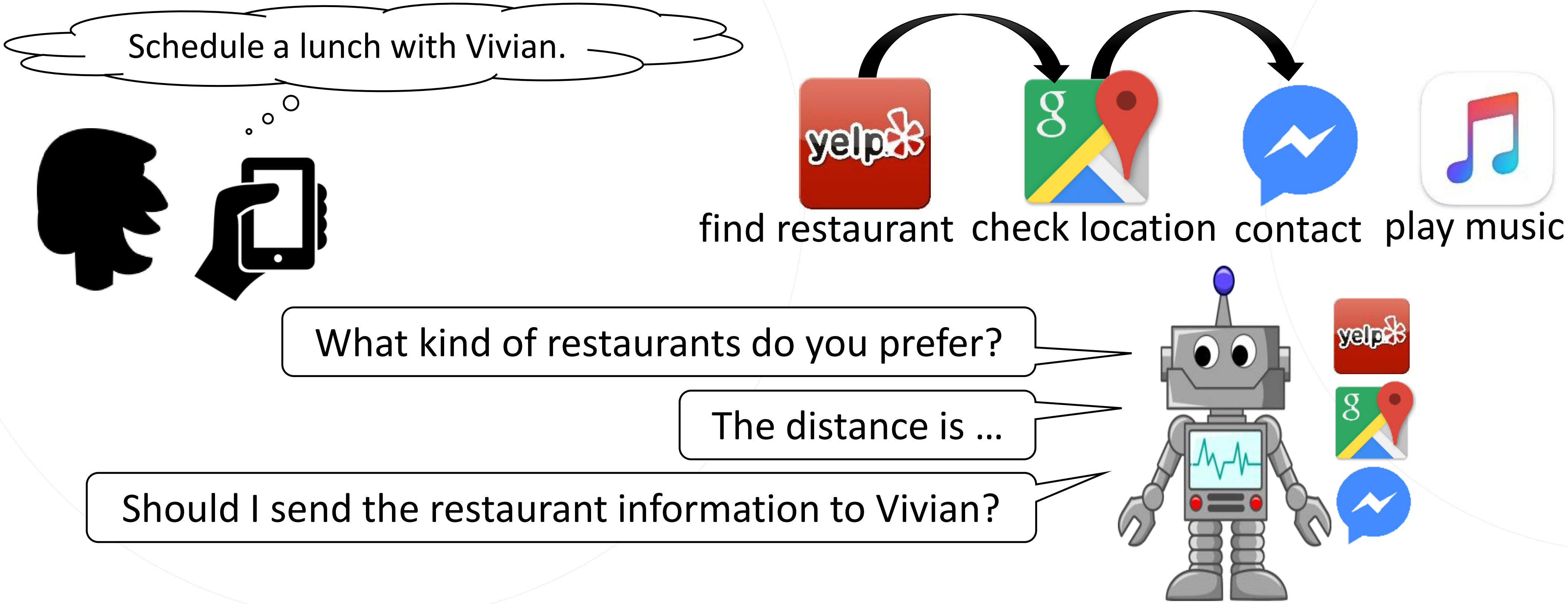
NTU MIULAB

Evolution Roadmap



Common Sense for Dialogue Planning (Sun+, 2016)

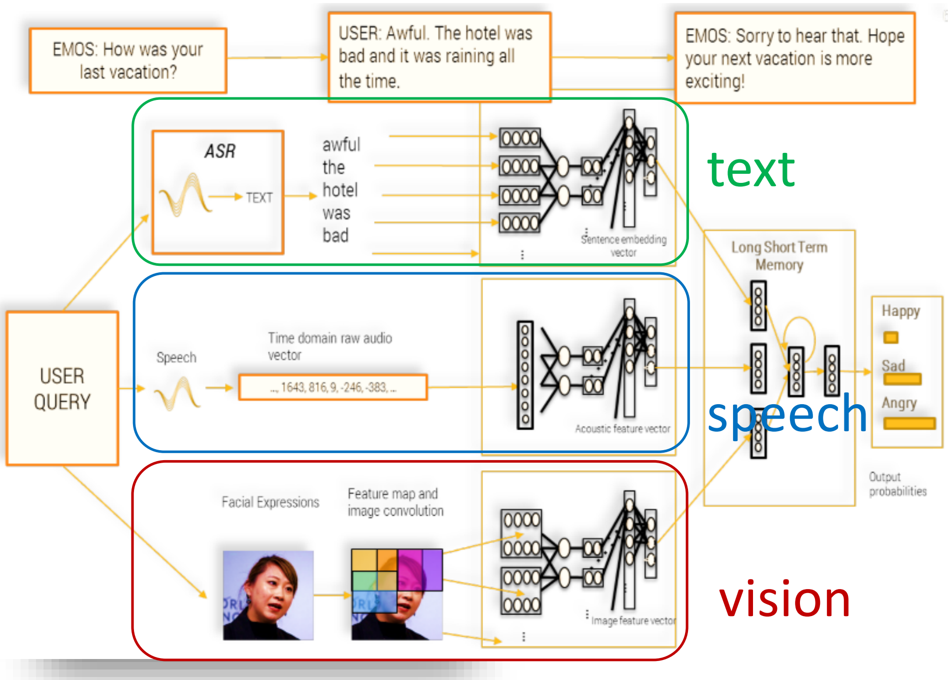
- High-level intention may span several domains



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

Empathy in Dialogue System (Fung+, 2016)

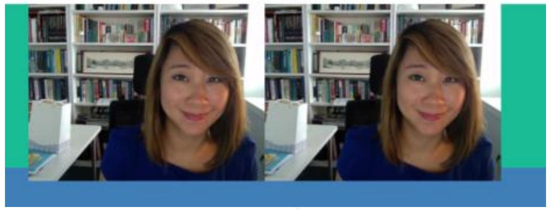
- Embed an empathy module
  - Recognize emotion using multimodality
  - Generate emotion-aware responses



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

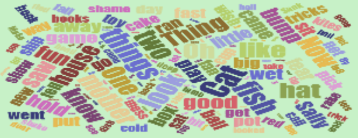
(index):1728
(index):1729

{
  "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"
}

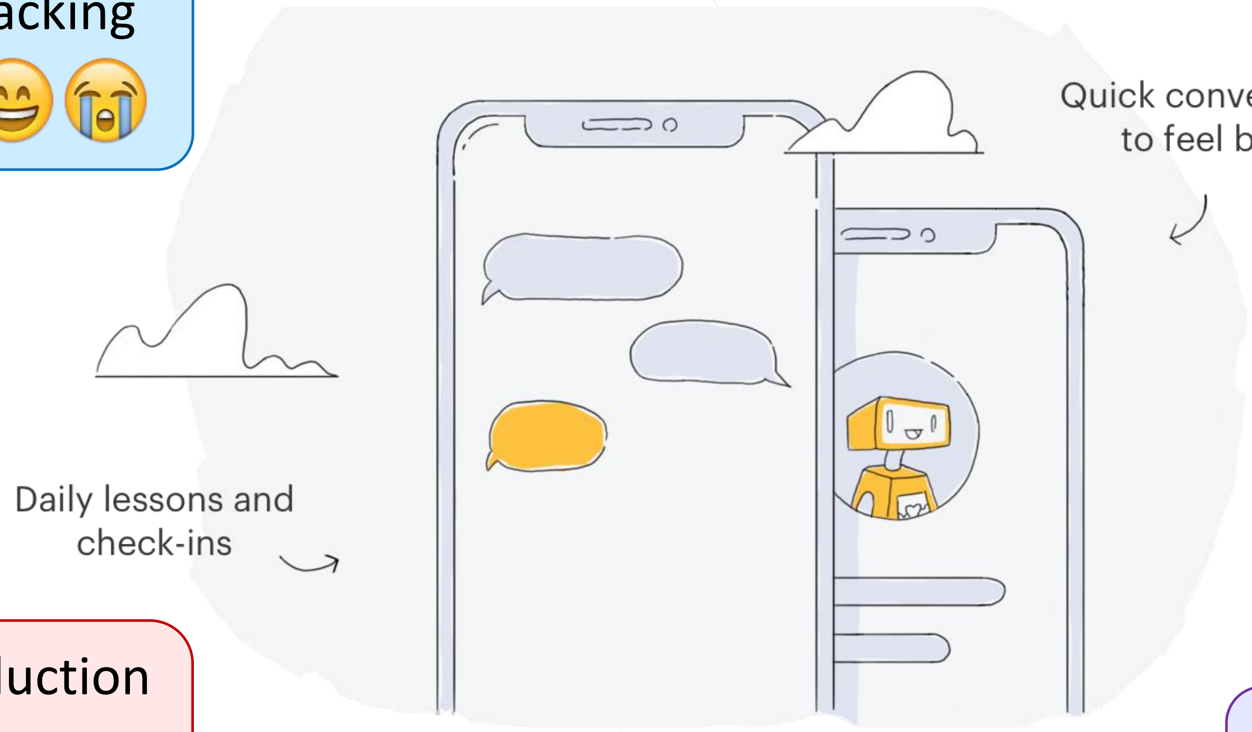
Cognitive Behavioral Therapy (CBT)

Mood Tracking

Pattern Mining



Depression Reduction



Content Providing

Always Be There

Know You Well

Summarized Challenges

- 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
- Data collection and analysis from un-structured data
- Complex-cascade systems requires high accuracy for working good as a whole

Framework & Resources

- MiuLab codes are available here: https://github.com/MiuLab/
- Frameworks
  - Tensorflow, PyTorch
- Resources
  - NVIDIA GTX 1070


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

NTU MIULAB

Q&A

Thanks for Your Attention!

Yun-Nung (Vivian) Chen
Assistant Professor
National Taiwan University
y.v.chen@ieee.org / http://vivianchen.idv.tw