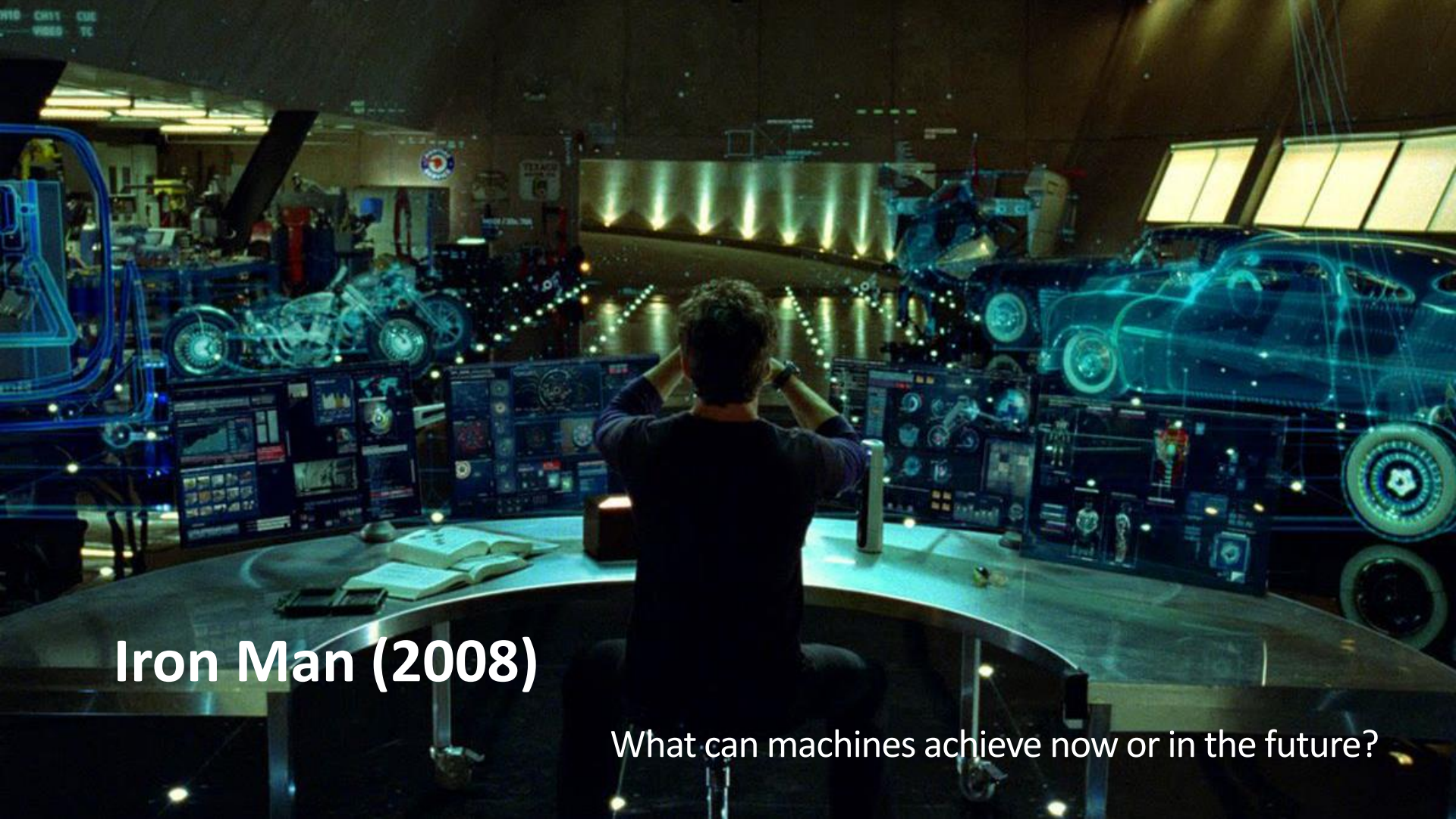
Towards Open-Domain Conversational AI

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

國立臺灣大學
National Taiwan University

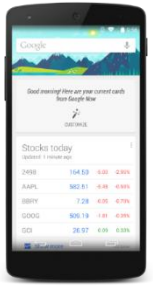


Iron Man (2008)

What can machines achieve now or in the future?

Language Empowering Intelligent Assistants

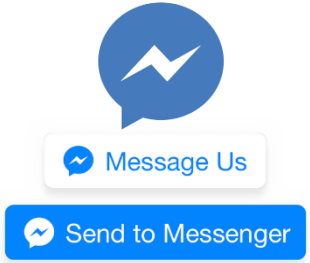
Apple Siri (2011)



Google Now (2012)
Google Assistant (2016)

Microsoft Cortana (2014)

Amazon Alexa/Echo (2014)



Facebook M & Bot (2015)

Google Home (2016)

Apple HomePod (2017)

Why and When We Need?

“I want to chat”	Turing Test (talk like a human)
“I have a question”	Information consumption
“I need to get this done”	Task completion
“What should I do?”	Decision support

Why and When We Need?

“I want to chat”	Turing Test (talk like a human)
“I have a question”	Information consumption
“I need to get this done”	Task completion
“What should I do?”	Decision support

Why and When We Need?

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

- *What is today's agenda?*
- *Which room is SCAI workshop in?*
- *What does SCAI stand for?*

Why and When We Need?

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

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

Why and When We Need?

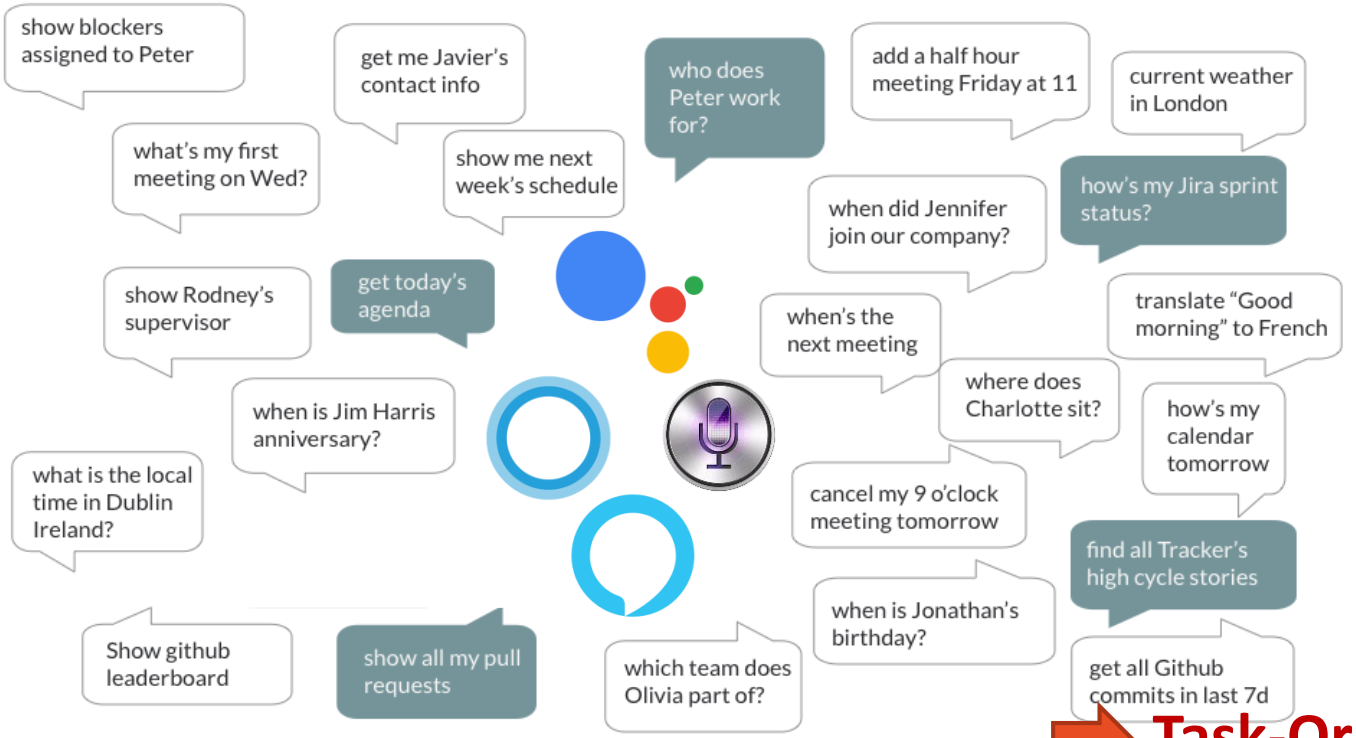
"I want to chat"	Turing Test (talk like a human)
"I have a question"	Information consumption
"I need to get this done"	Task completion
"What should I do?"	Decision support

- *Is this brussels waffle worth to try?*
- *Is the SCAI workshop good to attend?*

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	

Intelligent Assistants



Task-Oriented

Conversational Agents

Chit-Chat

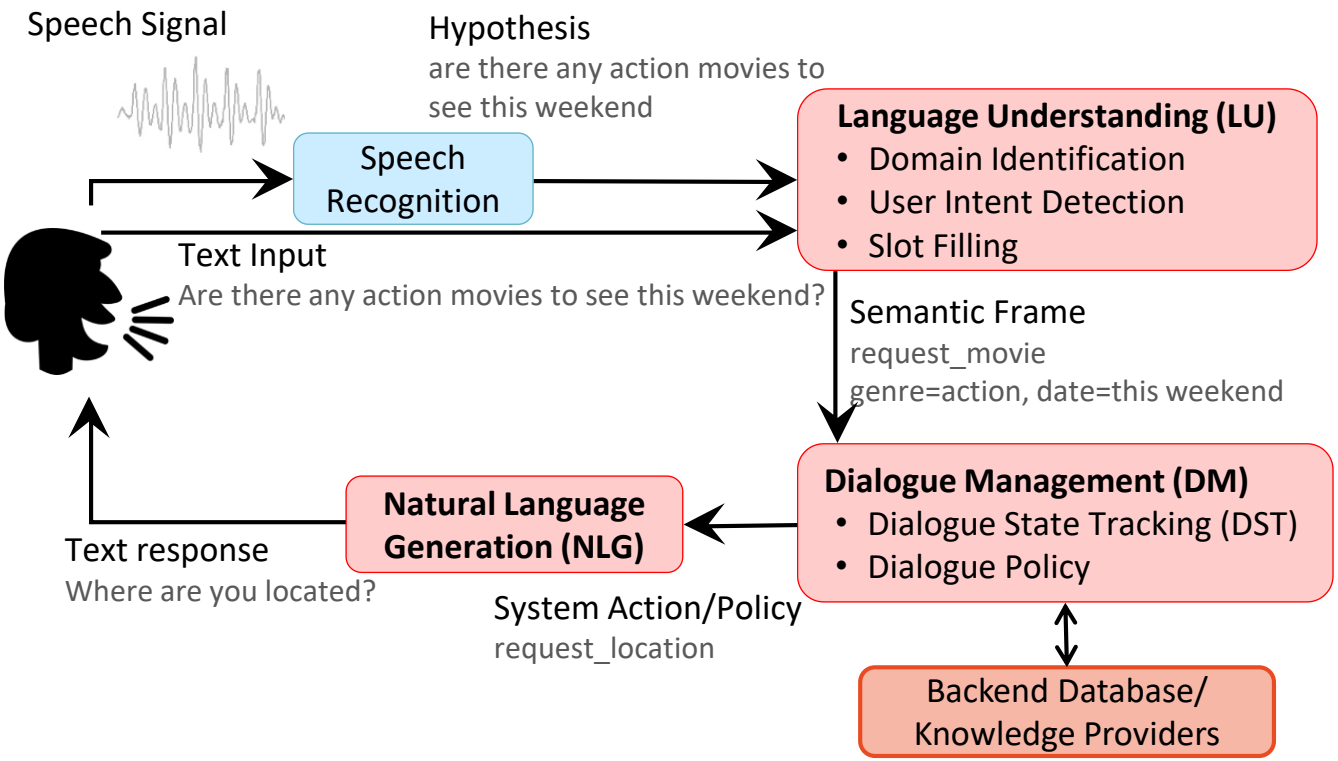
Task-Oriented

Task-Oriented Dialogue Systems

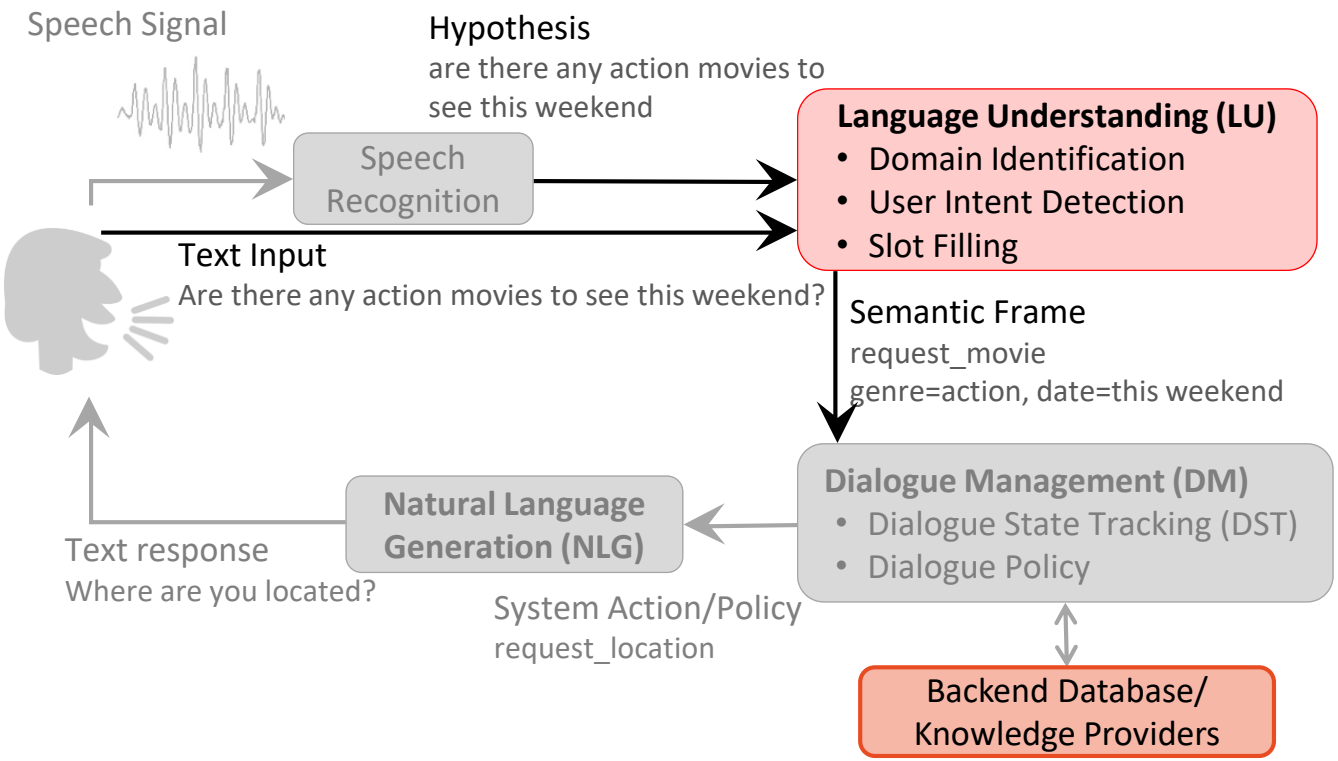
JARVIS – Iron Man's Personal Assistant

Baymax – Personal Healthcare Companion

Task-Oriented Dialogue Systems (Young, 2000)



Task-Oriented Dialogue Systems (Young, 2000)

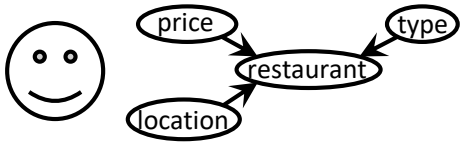


Semantic Frame Representation

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

Restaurant Domain

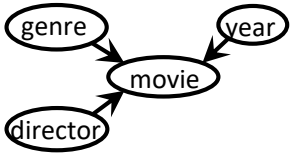
find me a cheap taiwanese restaurant in oakland



find_restaurant (price="cheap", type="taiwanese", location="oakland")

Movie Domain

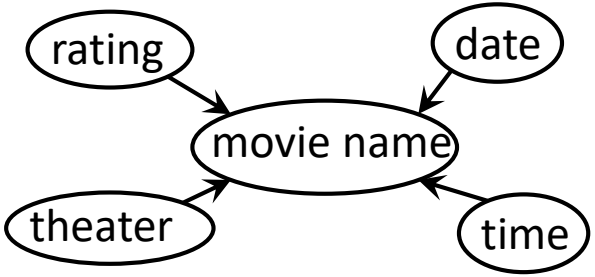
show me action movies directed by james cameron



find_movie (genre="action", director="james cameron")

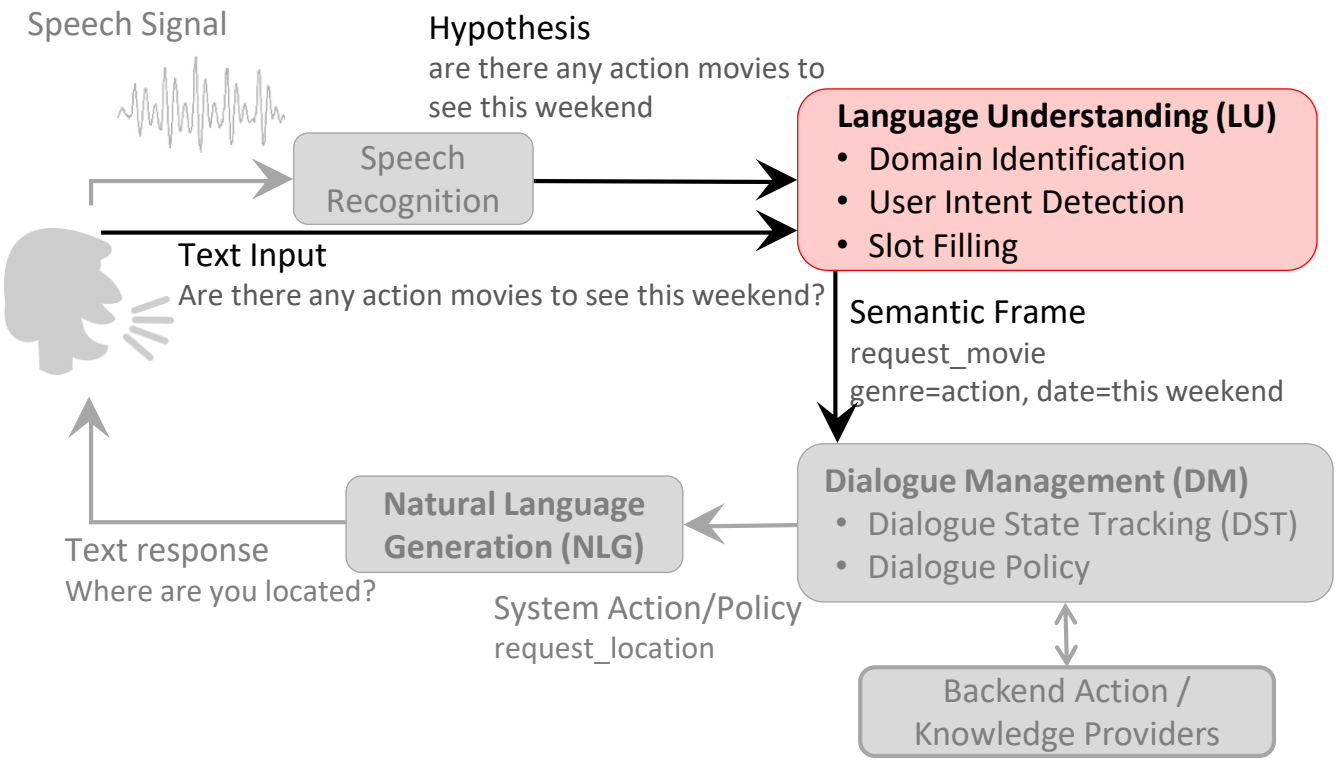
Backend Database / Ontology

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



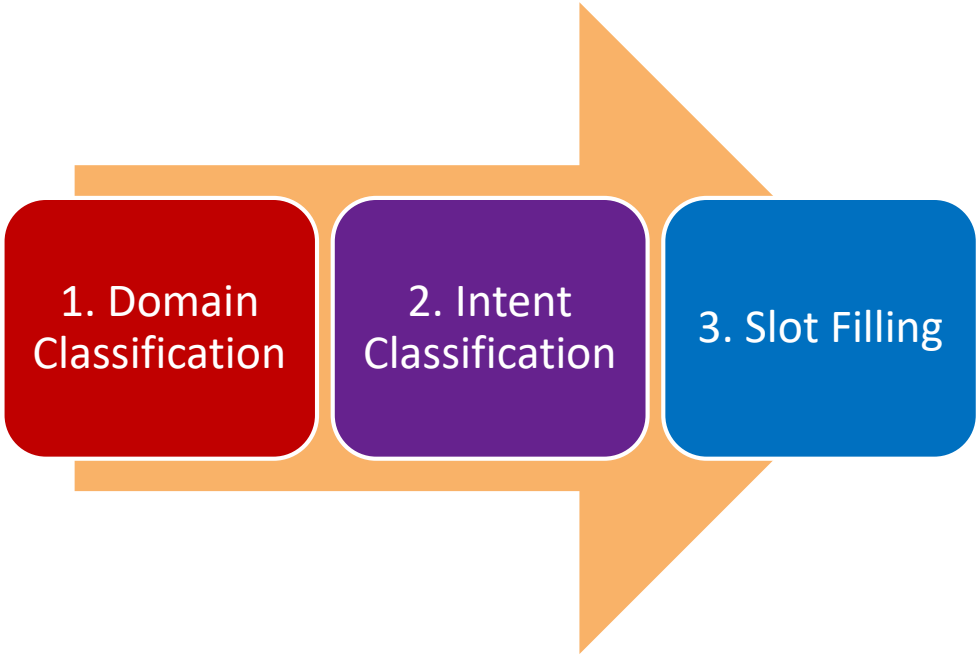
Movie Name	Theater	Rating	Date	Time
Iron Man Last	Taipei A1	8.5	2018/10/31	09:00
Iron Man Last	Taipei A1	8.5	2018/10/31	09:25
Iron Man Last	Taipei A1	8.5	2018/10/31	10:15
Iron Man Last	Taipei A1	8.5	2018/10/31	10:40

Task-Oriented Dialogue Systems (Young, 2000)



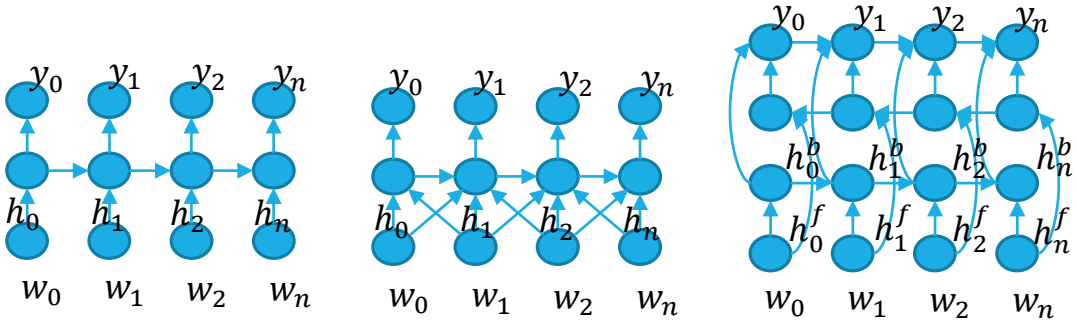
Language Understanding (LU)

- Pipelined



RNN for Slot Tagging – I (Yao et al, 2013; Mesnil et al, 2015)

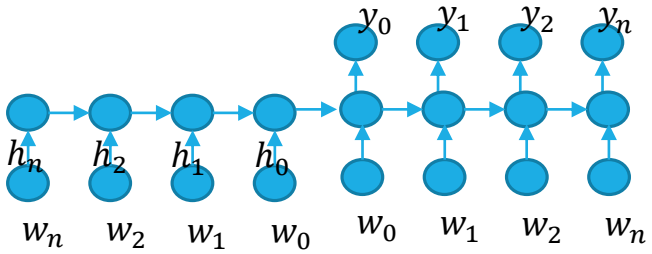
- Variations:
  a. RNNs with LSTM cells
  b. Input, sliding window of n-grams
  c. Bi-directional LSTMs



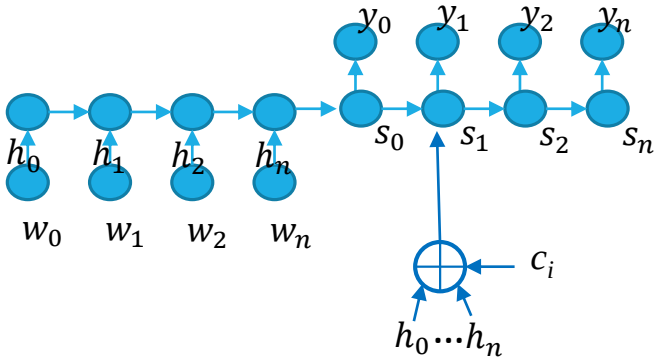
(a) LSTM (b) LSTM-LA (c) bLSTM

RNN for Slot Tagging – II (Kurata et al., 2016; Simonnet et al., 2015)

- Encoder-decoder networks
  - Leverages sentence level information



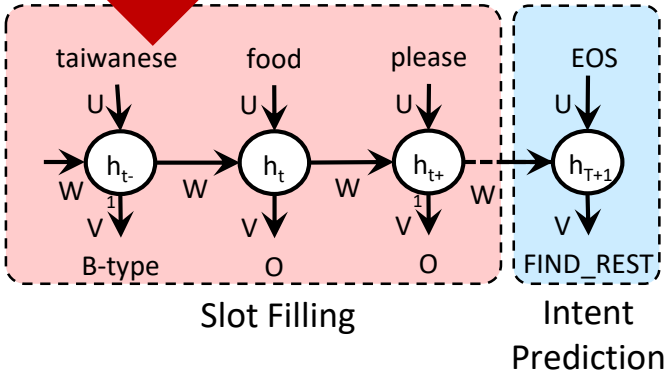
- Attention-based encoder-decoder
  - Use attention (as in MT) in the encoder-decoder network



Joint Semantic Frame Parsing

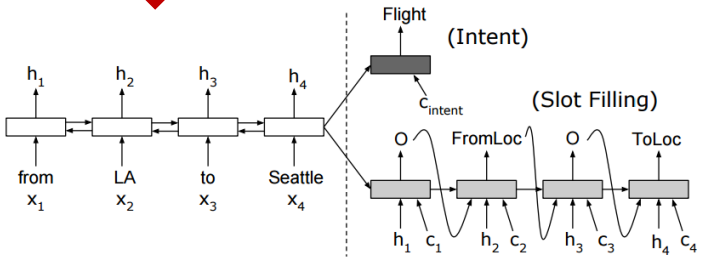
Sequence-based (Hakkani-Tur et al., 2016)

- Slot filling and intent prediction in the same output sequence



Parallel (Liu and Lane, 2016)

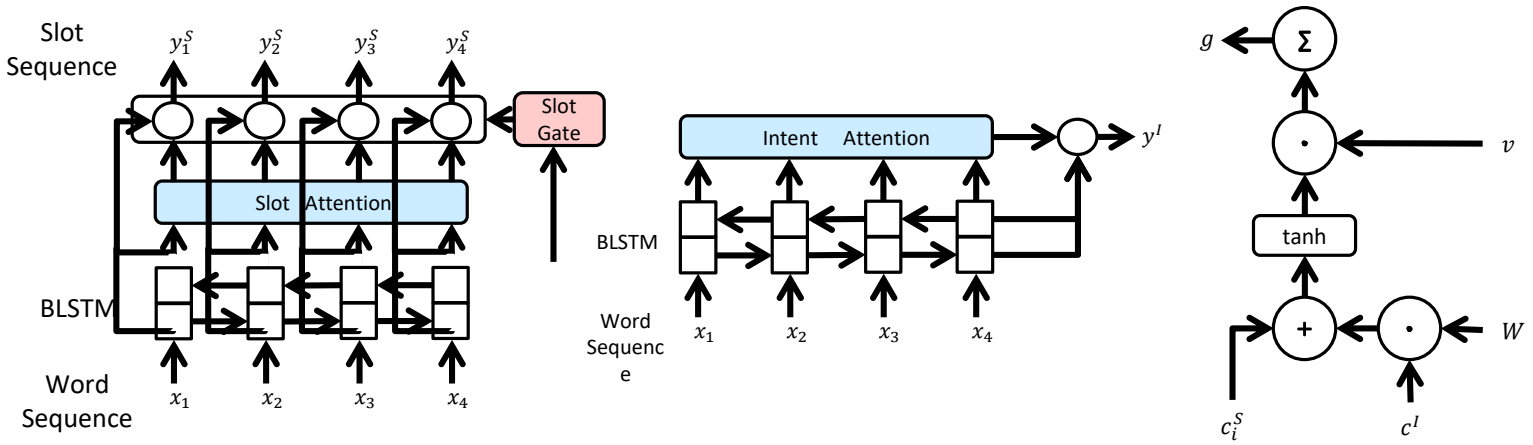
- Intent prediction and slot filling are performed in two branches



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 et al., 2018)



- Slot Gate $g = \sum v \cdot \tanh(c_i^S + W \cdot c^I)$
- Slot Prediction $y_i^S = softmax(W^S(h_i + c_i^S) + b^S)$ ➡ $y_i^S = softmax(W^S(h_i + \boldsymbol{g} \cdot c_i^S) + b^S)$

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

W^S: matrix for output layer
b^S : bias for output layer

c_i^S: slot context vector
c^I : intent context vector

W: trainable matrix
v : trainable vector
g : scalar gate value

Contextual Language Understanding

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")

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"

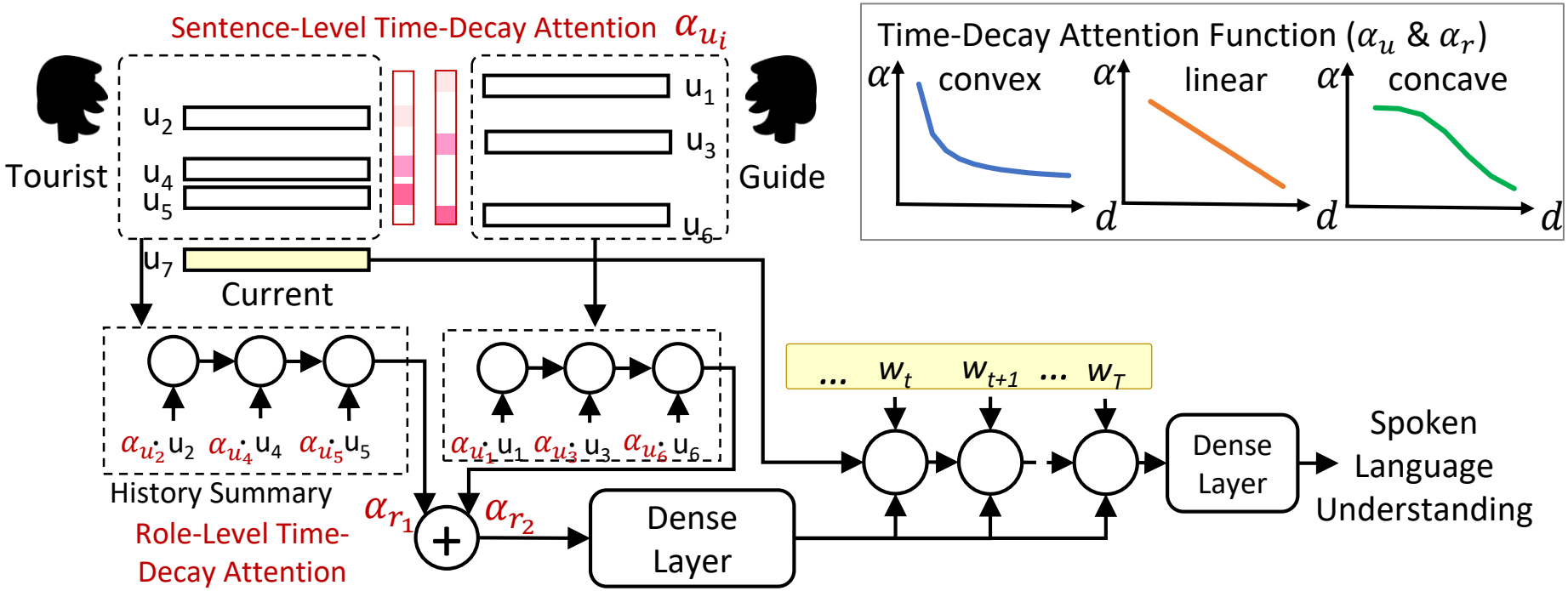
S: "What time would you like ?"

U: "Any time on saturday is fine"

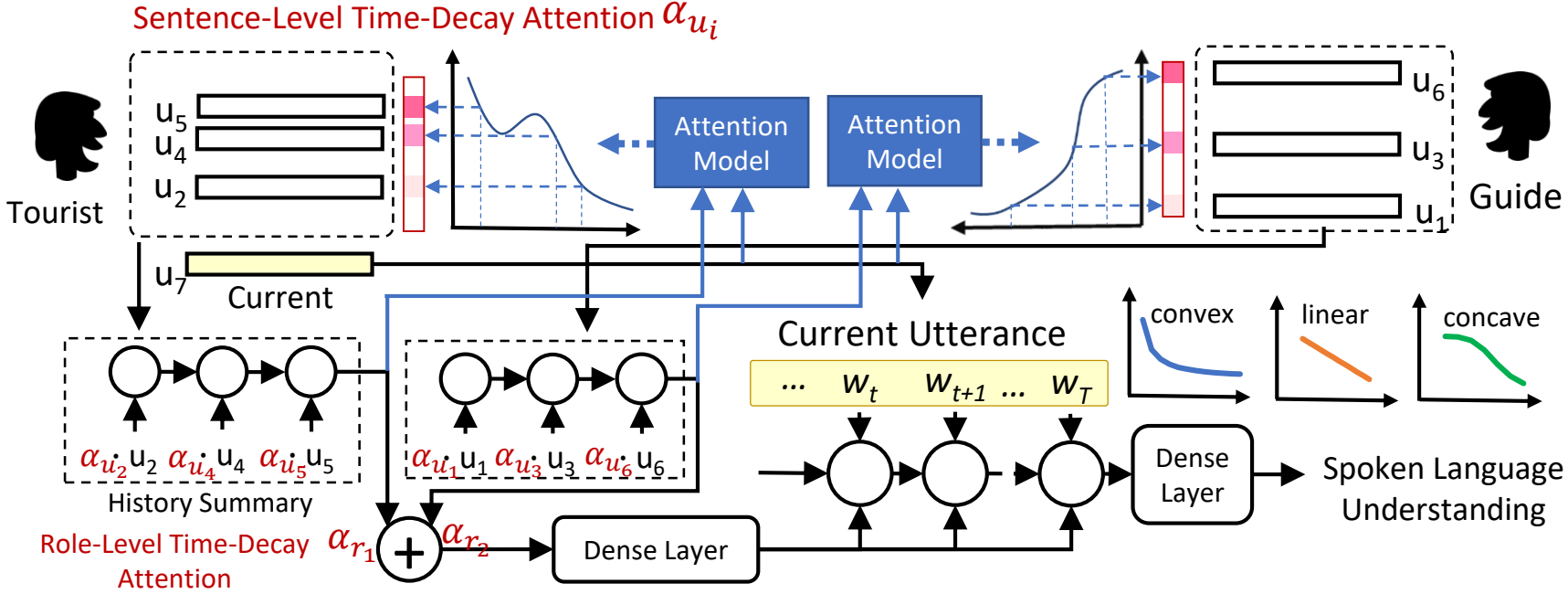
S: "okay , there is 4:10 pm , 5:40 pm and 9:20 pm" → 0.16

U: "Let's do 5:40"

Role-Based & Time-Aware Attention (Su et al., 2018)

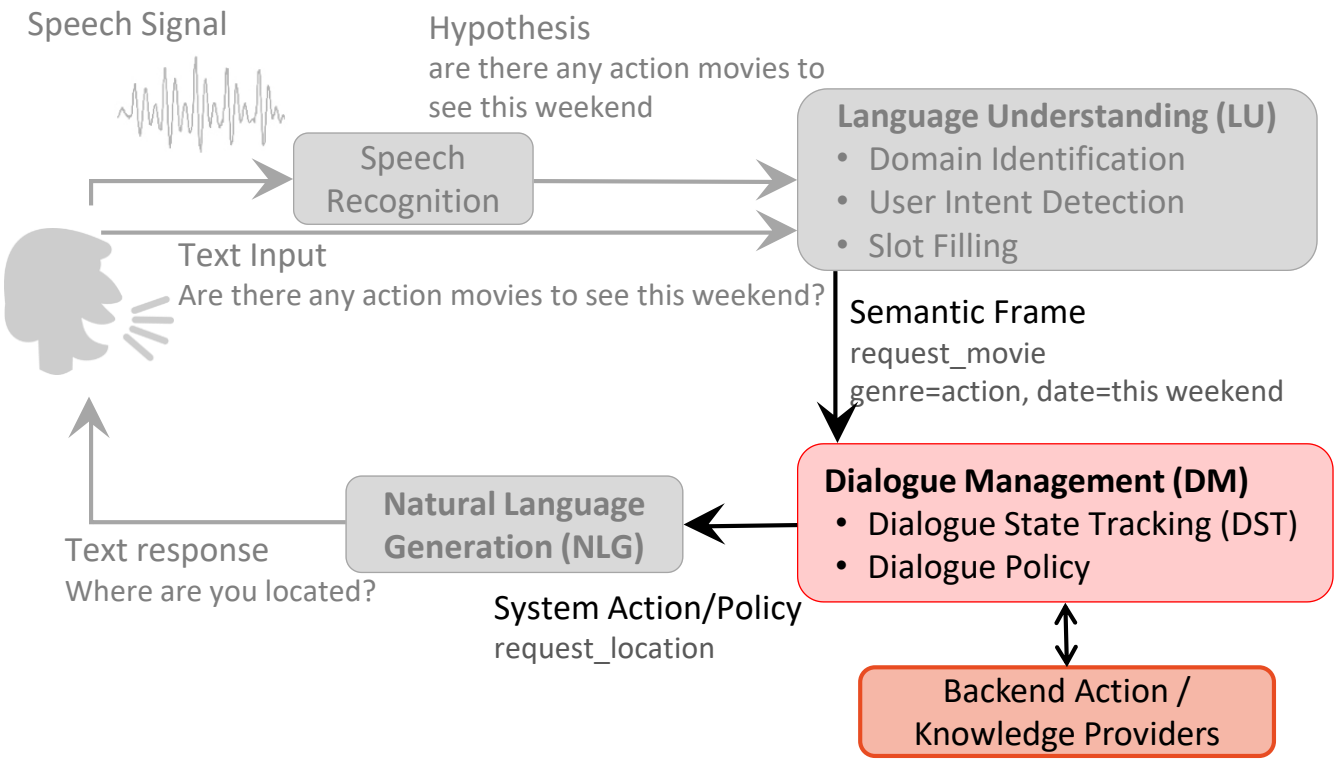


Context-Sensitive Time-Decay Attention (Su et al., 2018)



Time-decay attention significantly improves the understanding results

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)

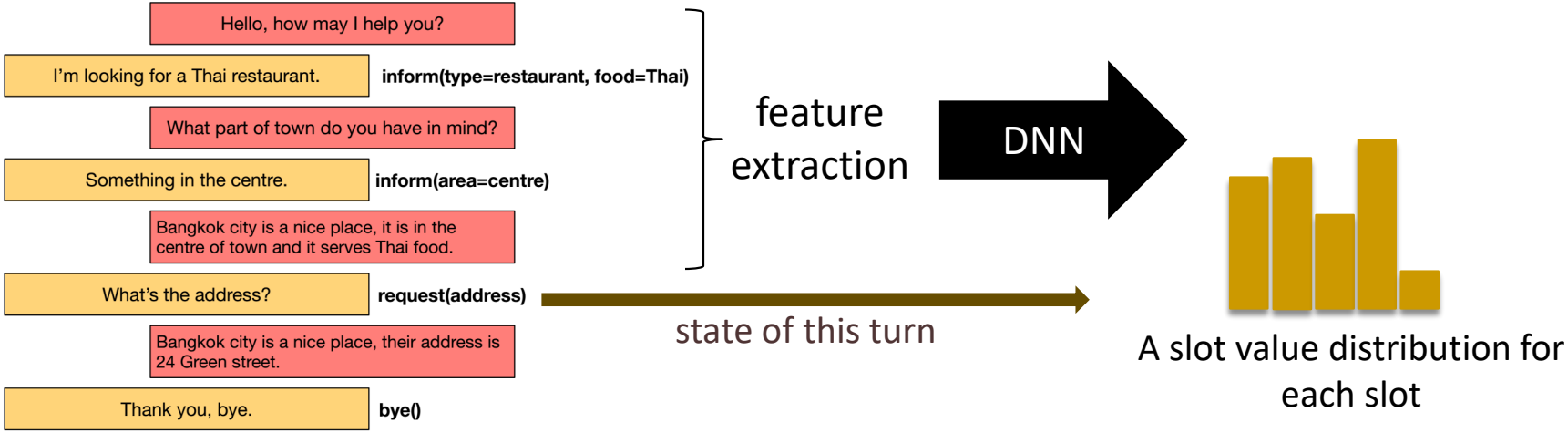
Bangkok city is a nice place, it is in the centre of town and it serves Thai food.

What's the address? — request (address)

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

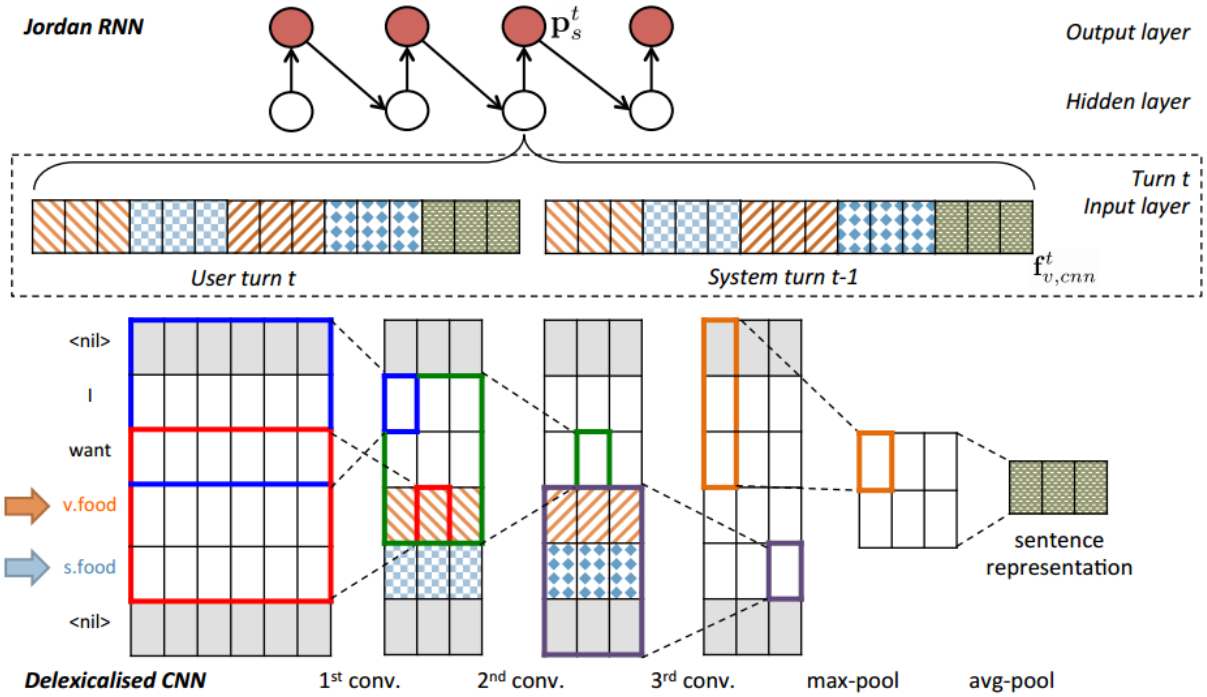
Thank you, bye. — bye ()

DNN for DST



multi-turn conversation

RNN-CNN DST (Wen et al., 2016)



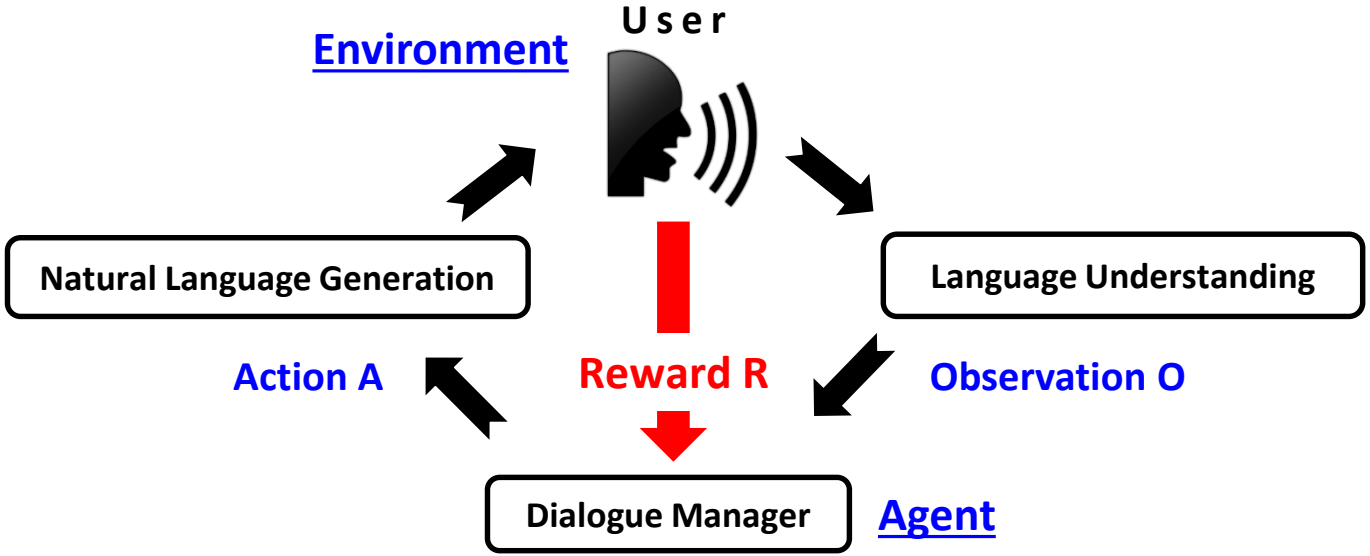
(Figure from Wen et al, 2016)

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 ()

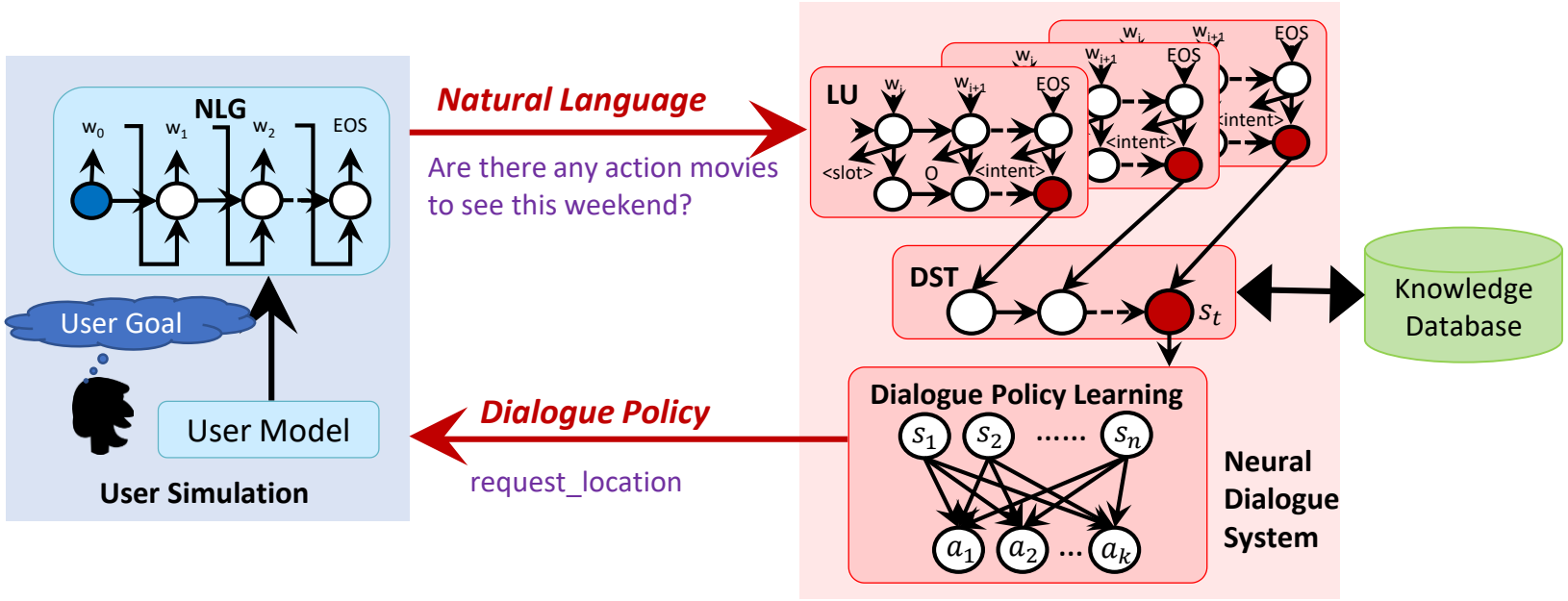
Dialogue Policy Optimization

- 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, 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 et al., 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!

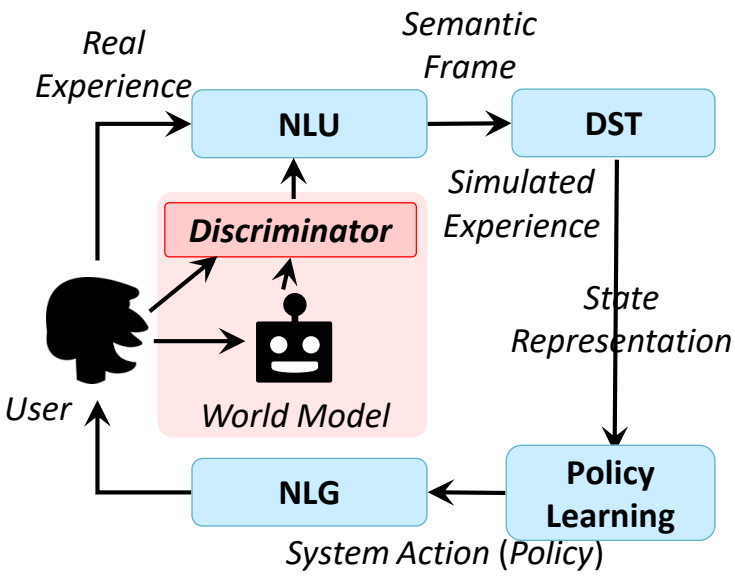
Automatically learn when and what to confirm for better robustness

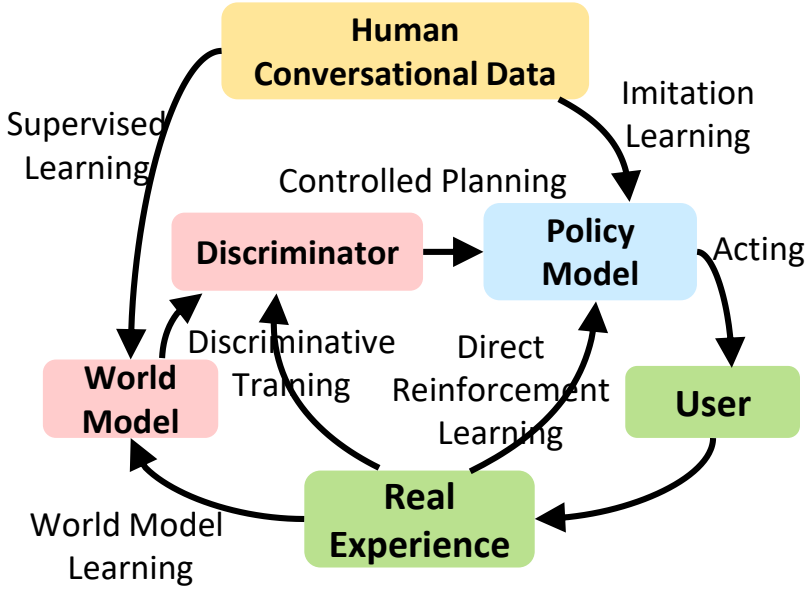
RL in Dialogue Systems

1. Sample inefficient, hard to design reward function, local optima…
2. Real users are expensive
3. Discrepancy between real users and simulators

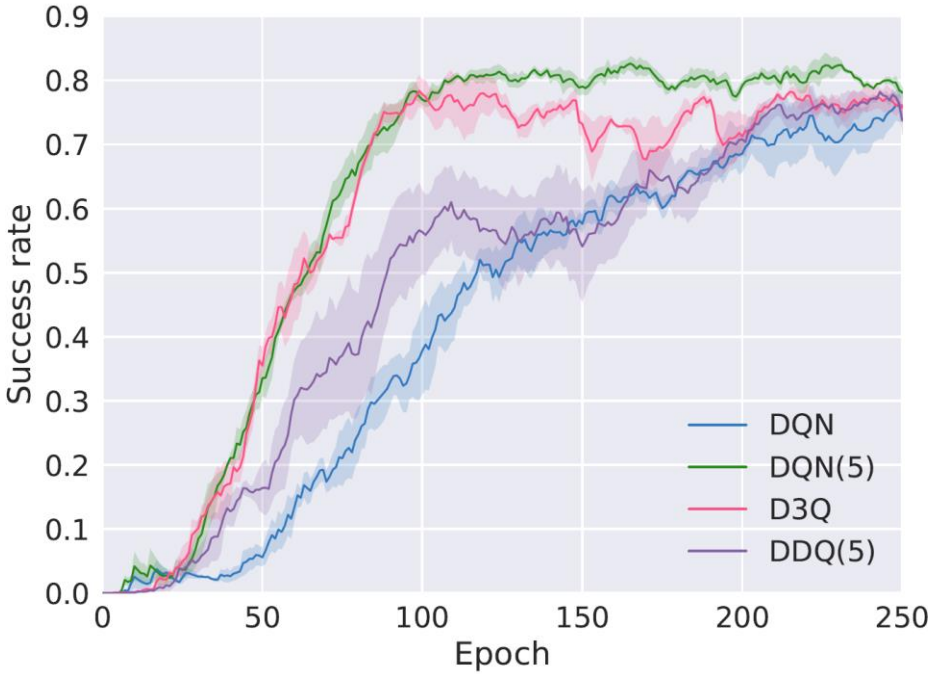
D3Q: Discriminative Deep Dyna-Q (Su et al., 2018)

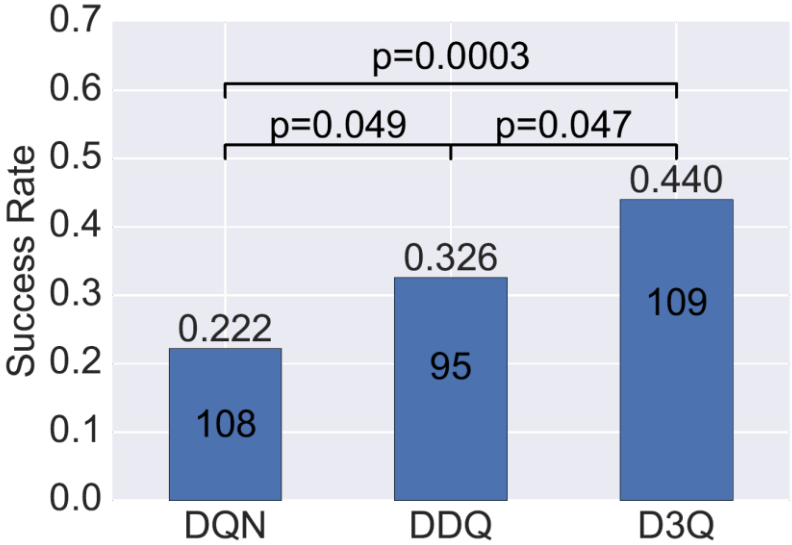
- Idea
  - learning with real users with planning
  - add a *discriminator* to filter out the bad experiences





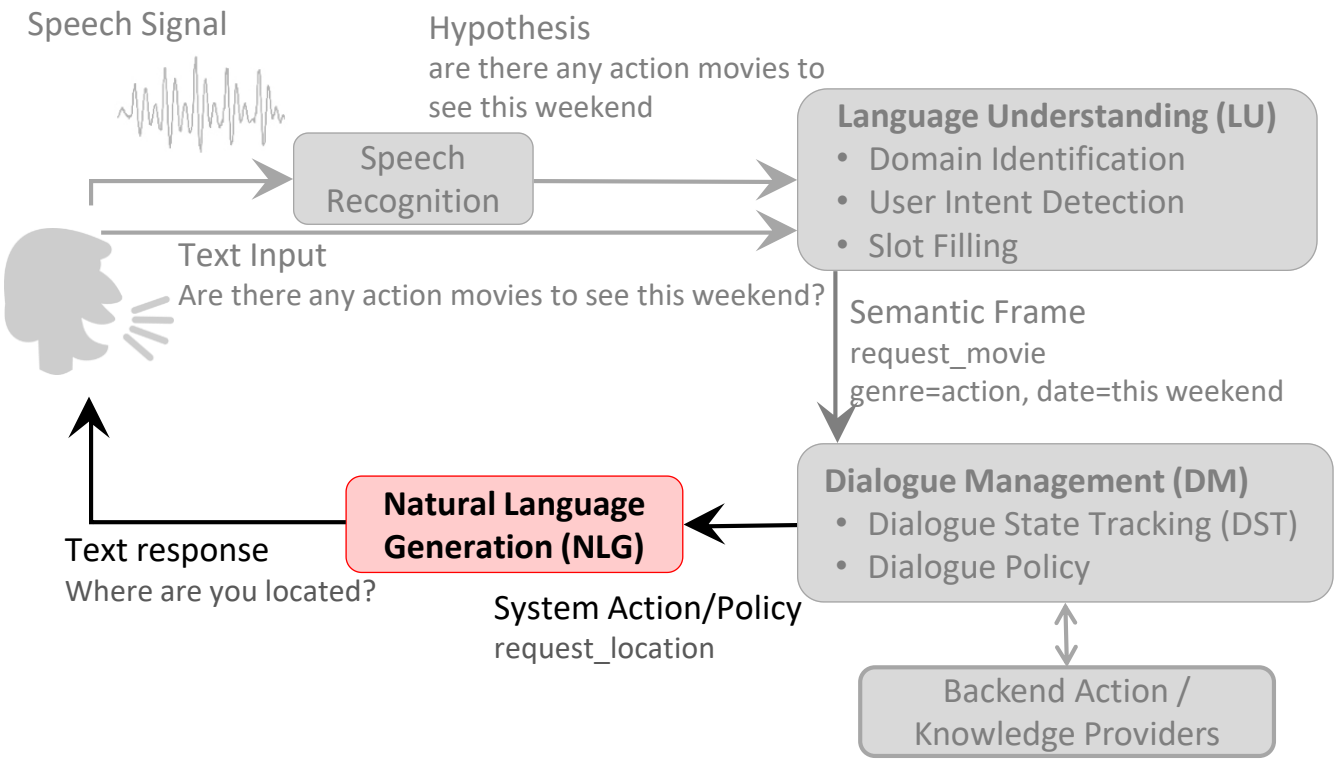
D3Q: Discriminative Deep Dyna-Q (Su et al., 2018)





The policy learning is more robust and shows the improvement in human evaluation

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

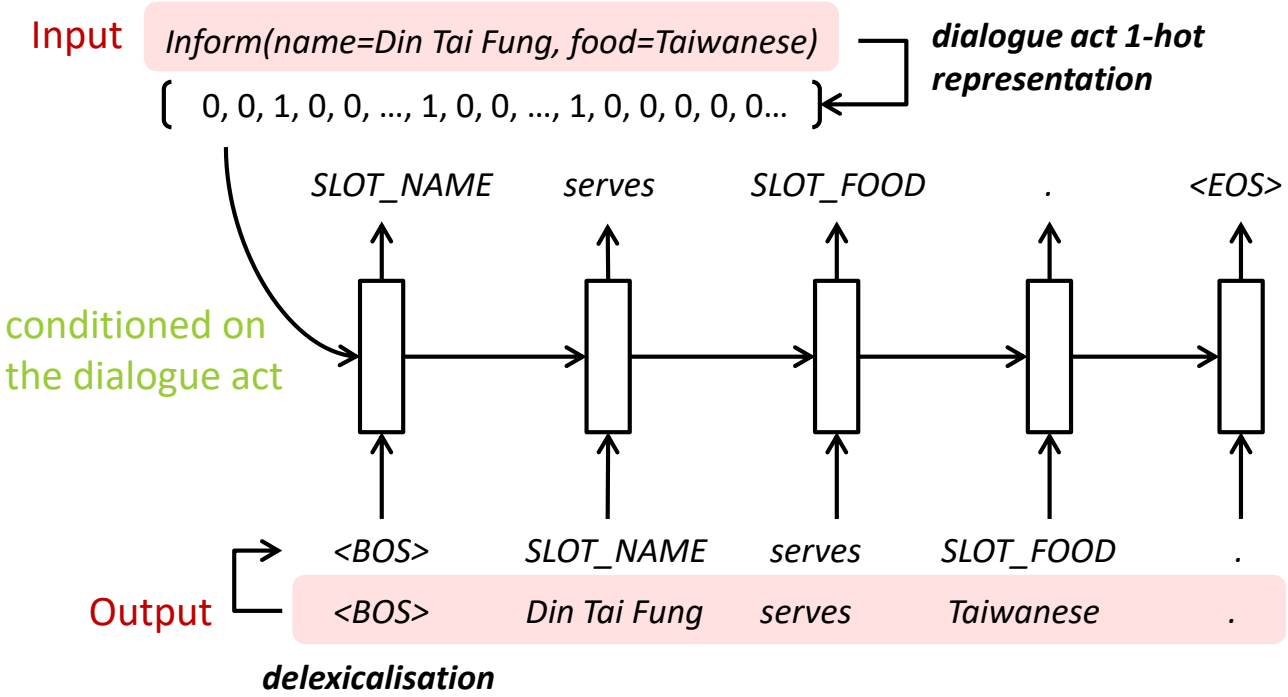
Template-Based NLG

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

Semantic Frame	Natural Language
confirm()	"Please tell me more about the product 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."

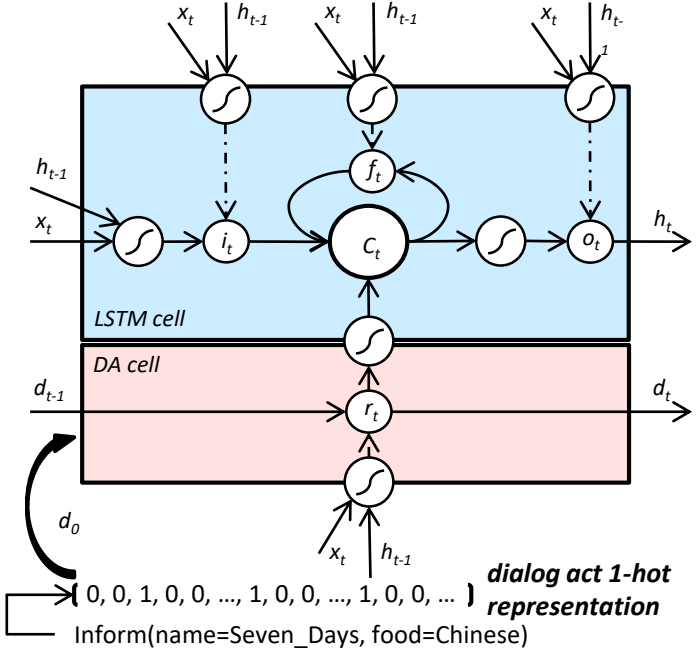
Pros: simple, error-free, easy to control
Cons: time-consuming, rigid, poor scalability

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.

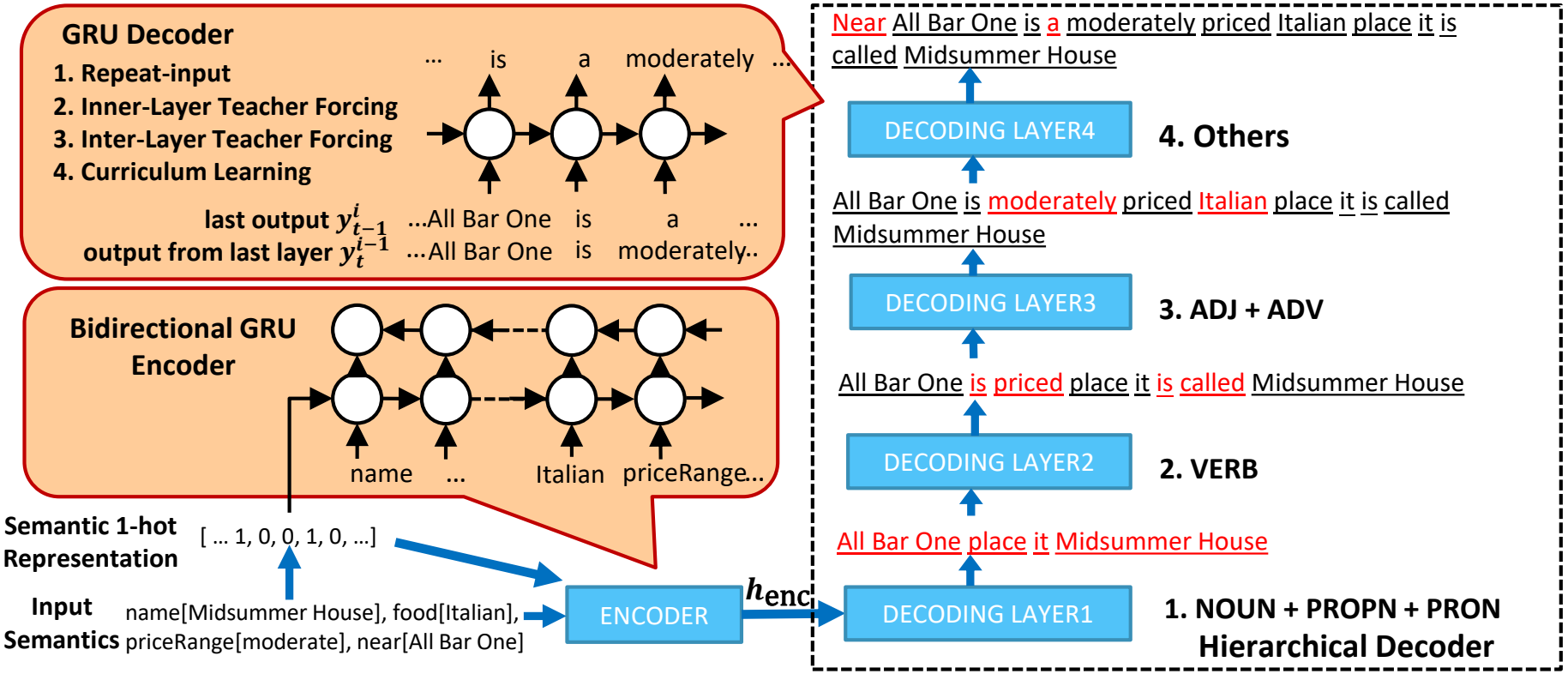


Idea: using gate mechanism to control the generated semantics (dialogue act/slots)

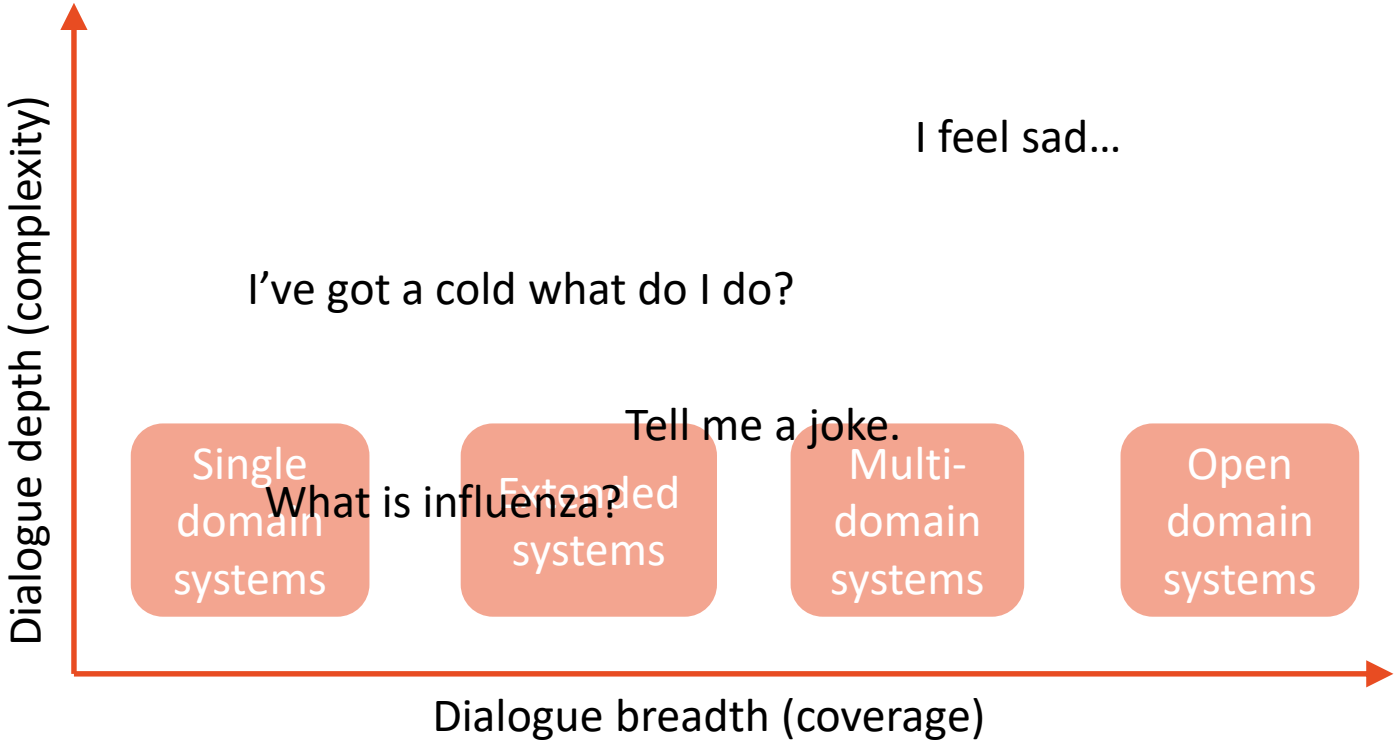
Issues in NLG

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

Hierarchical NLG w/ Linguistic Patterns (Su et al., 2018)



Evolution Roadmap



Dialogue Systems

Task-Oriented Dialogue

input **x**

output **y**

Understanding (NLU)

State tracker

Generation (NLG)

Dialog policy

DB

Database
Memory
External knowledge

Fully Data-Driven

input **x**

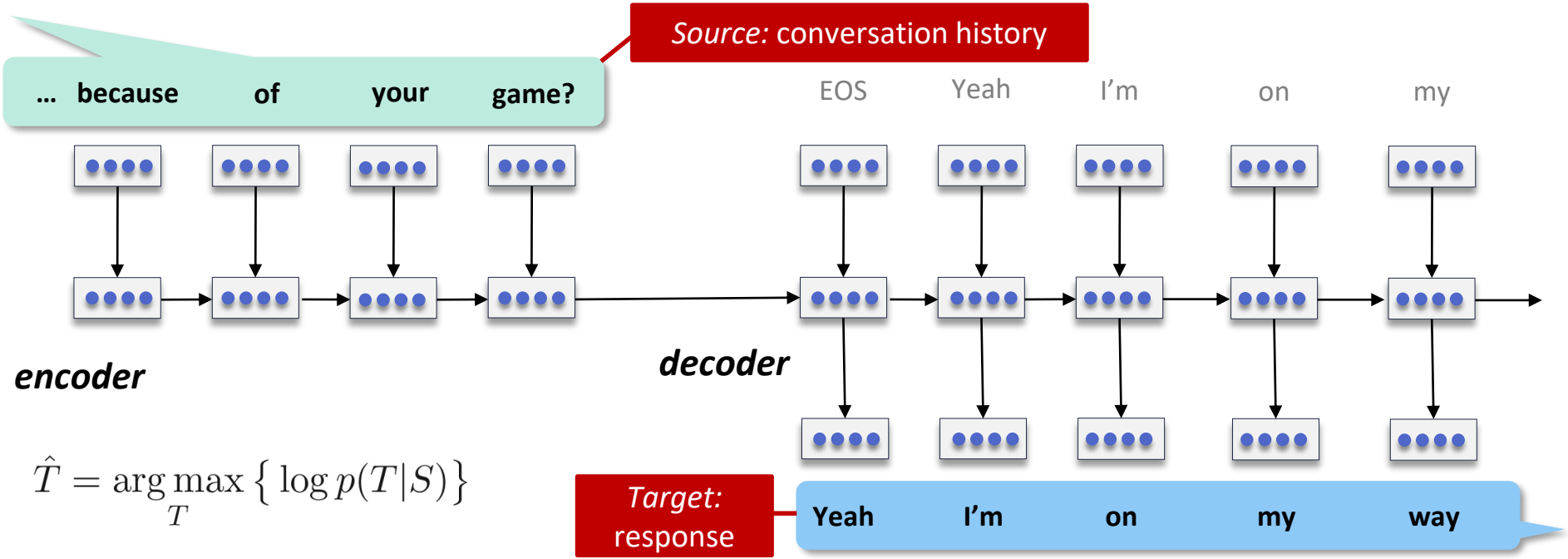
output **y**

Statistical model (e.g., neural)

DB

Chit-Chat Social Bots

Neural Response Generation (Sordoni et al., 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"

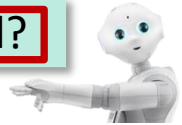
MMI for Response Diversity (Li et al., 2016)

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?

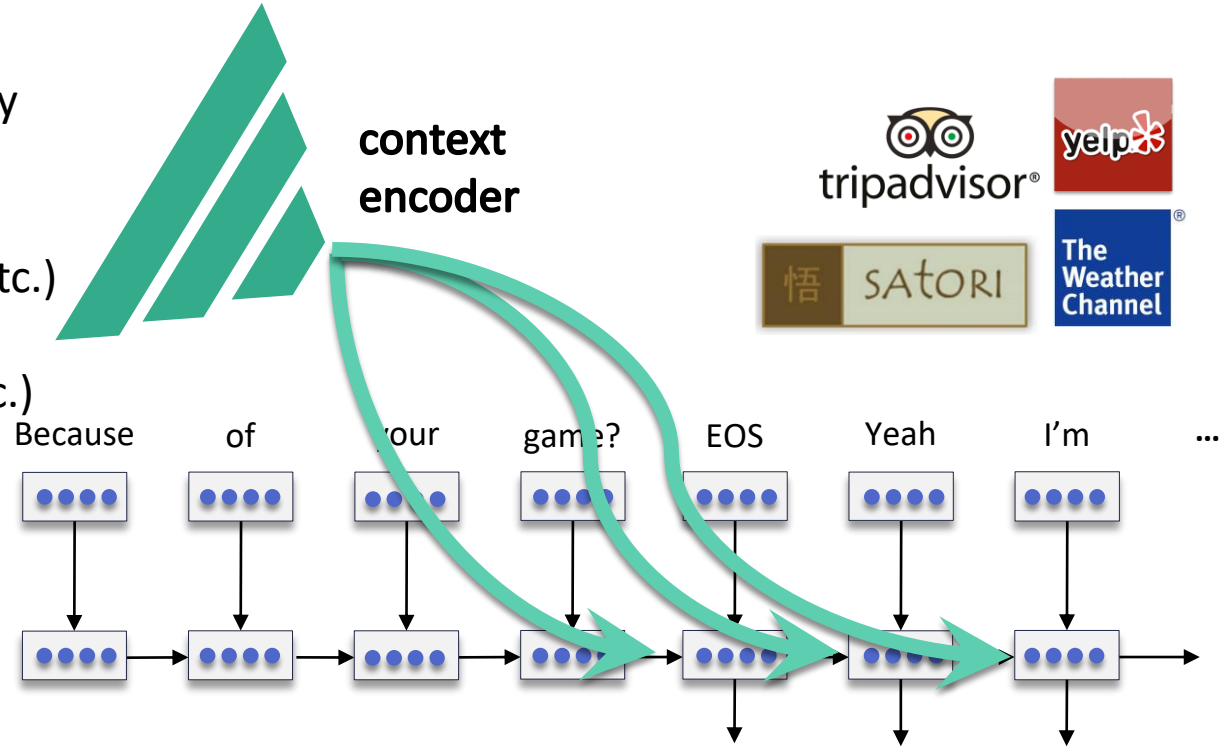
Depends on how much you drink!

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

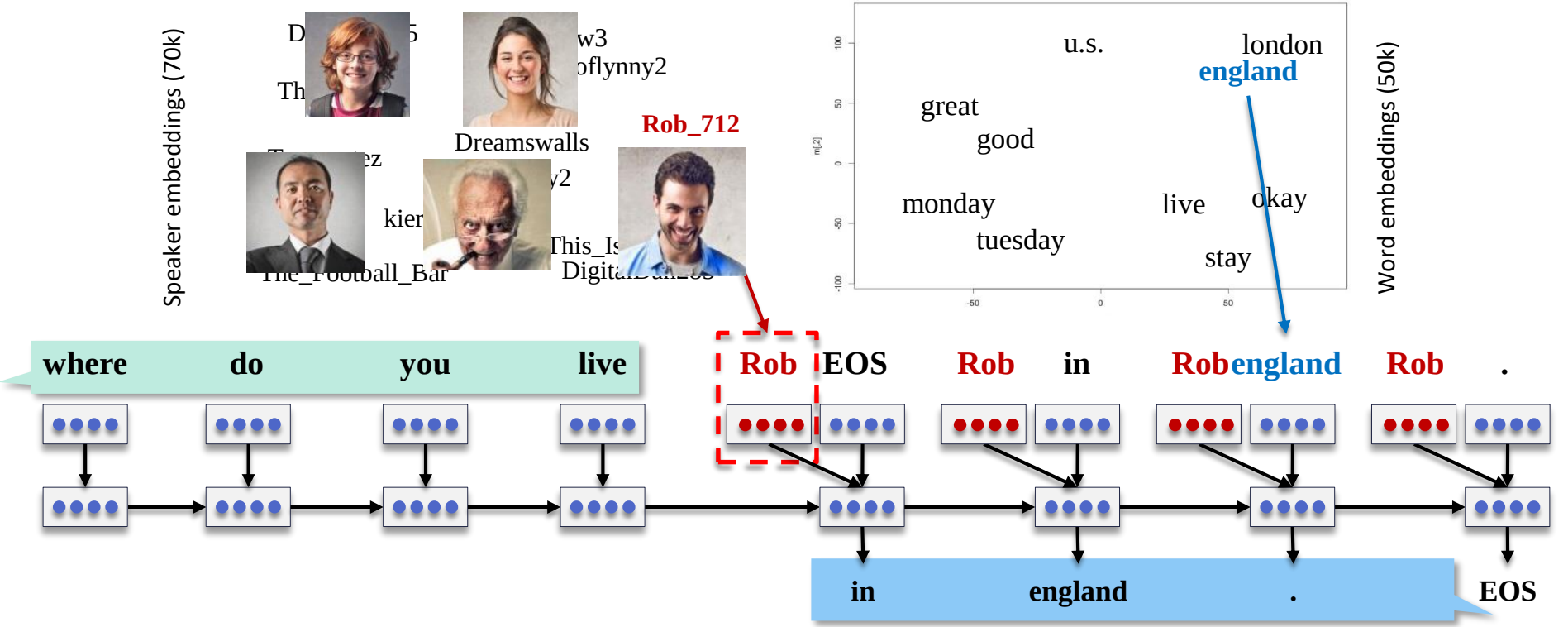


Issue 2: Response Inconsistency

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)

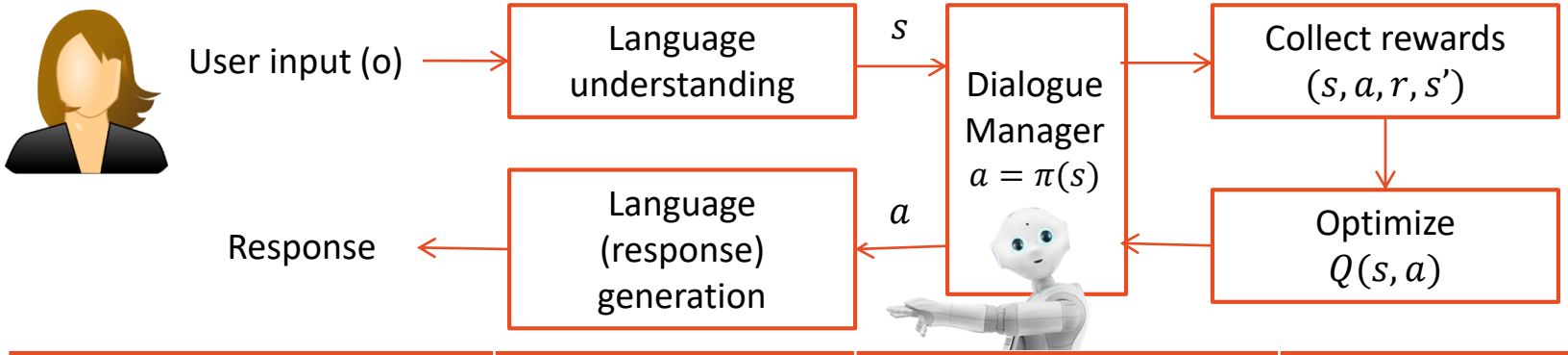
Baseline model → inconsistency

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)

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.

Chit-Chat v.s. Task-Oriented

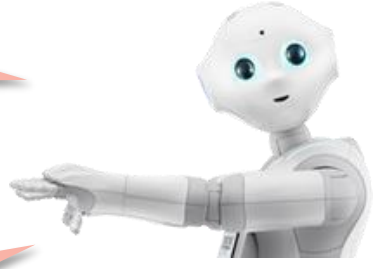
Social Chat
Engaging, Human-Like Interaction
(Ungrounded)

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!

Image-Grounded Dialogues (Mostafazadeh et al., 2017)

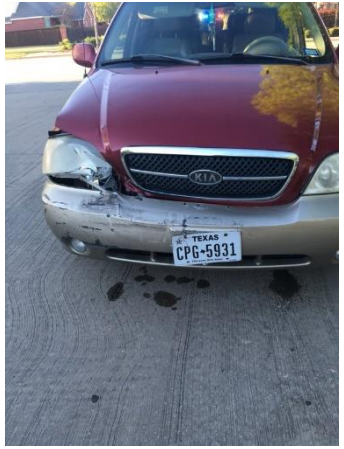
Neural model, trained on 250K Twitter convos:

-Turn 1: tweet + accompanying photo

-Turn 2: tweet is a question

Given data, other signals can be exploited in the same way: program state, database info, geolocation, etc.

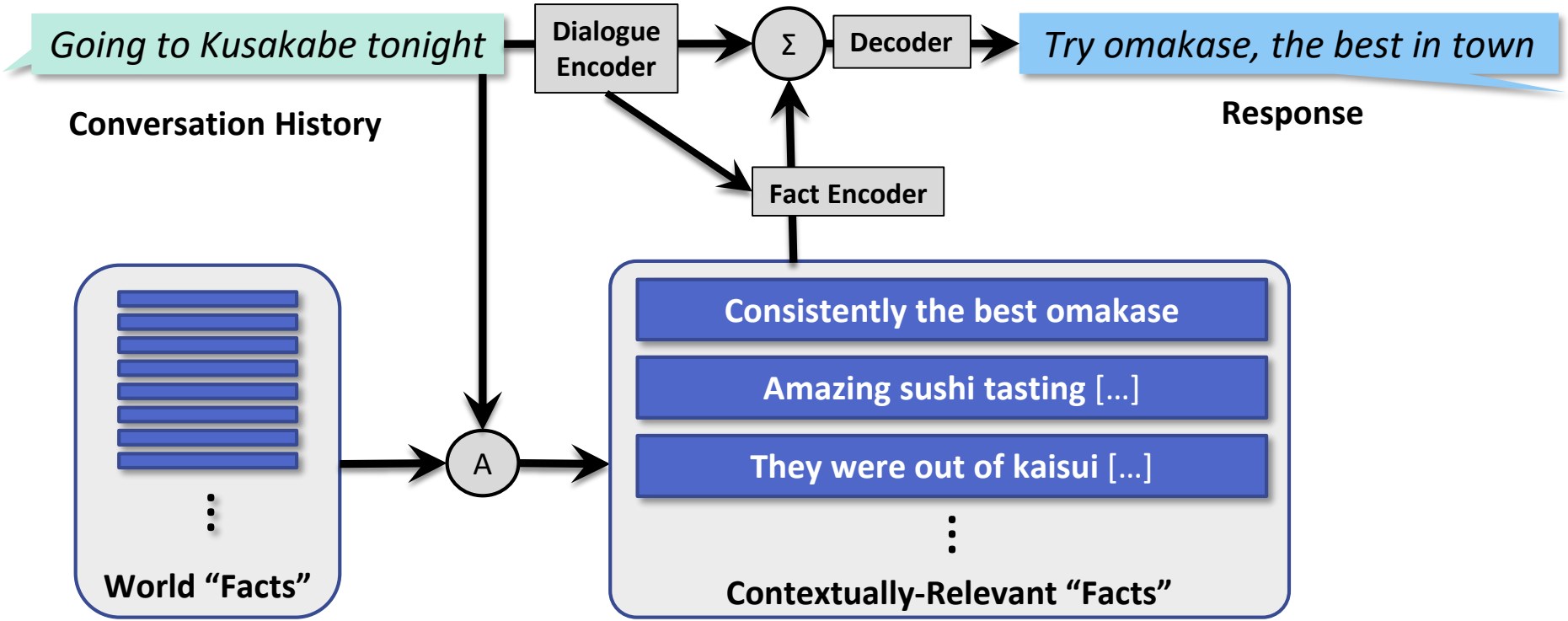
Tweet: some of the damage done
NeuroCon: **Is that at your house?**



Tweet: my poor baby moms saying it 's probably totaled so rip Bessie
NeuroCon: **How did you manage to do that?**

Tweet: as penitence I offer this
NeuroCon: **can I have one of those?**

Knowledge-Grounded Responses (Ghazvininejad et al., 2017)



Conversation and Non-Conversation Data

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

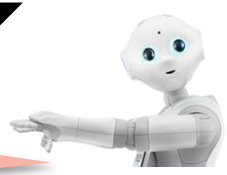
*Try **C**, one of the best **D** in the city.*

Conversation Data

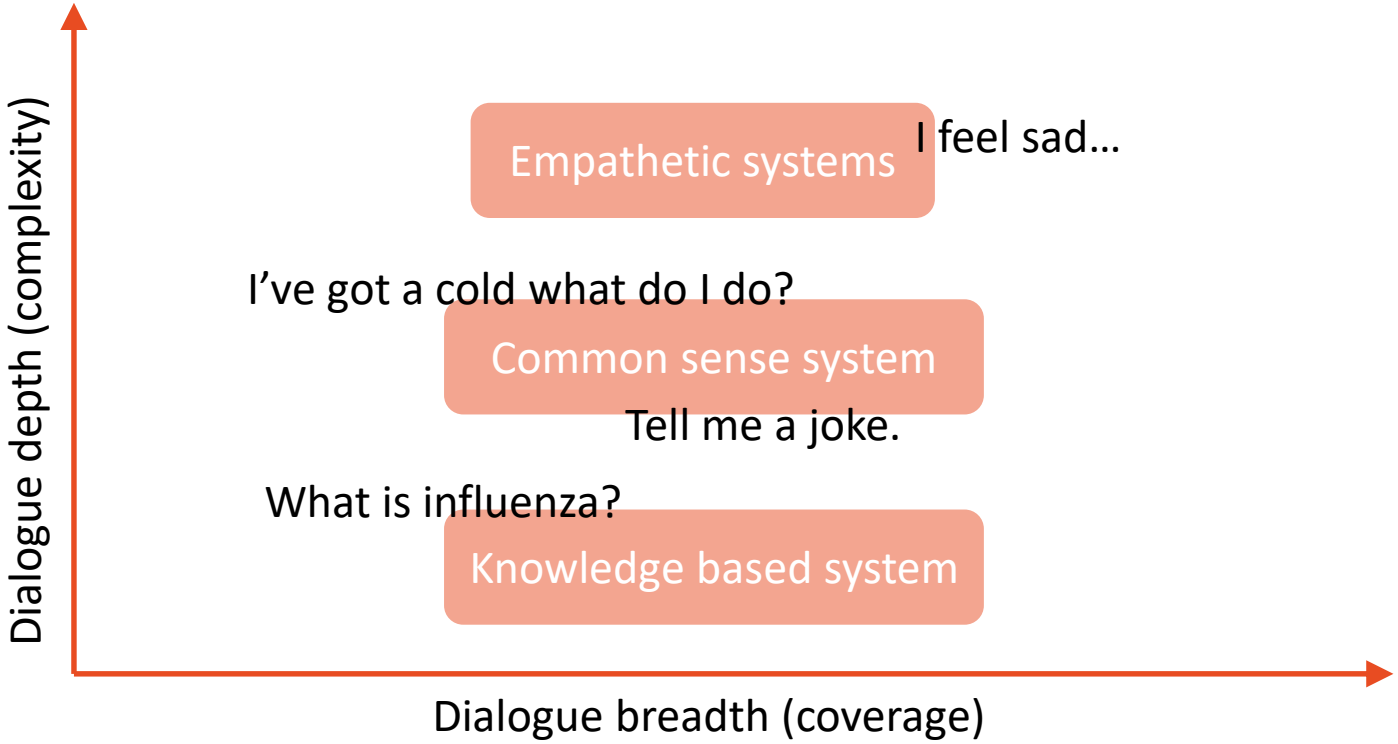
Knowledge Resource

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

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

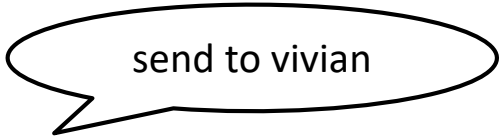


Evolution Roadmap

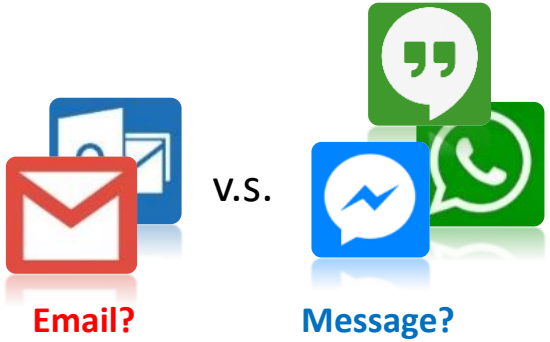


Multimodality & Personalization (Chen et al., 2018)

- Task: user intent prediction
- Challenge: language ambiguity



Communication



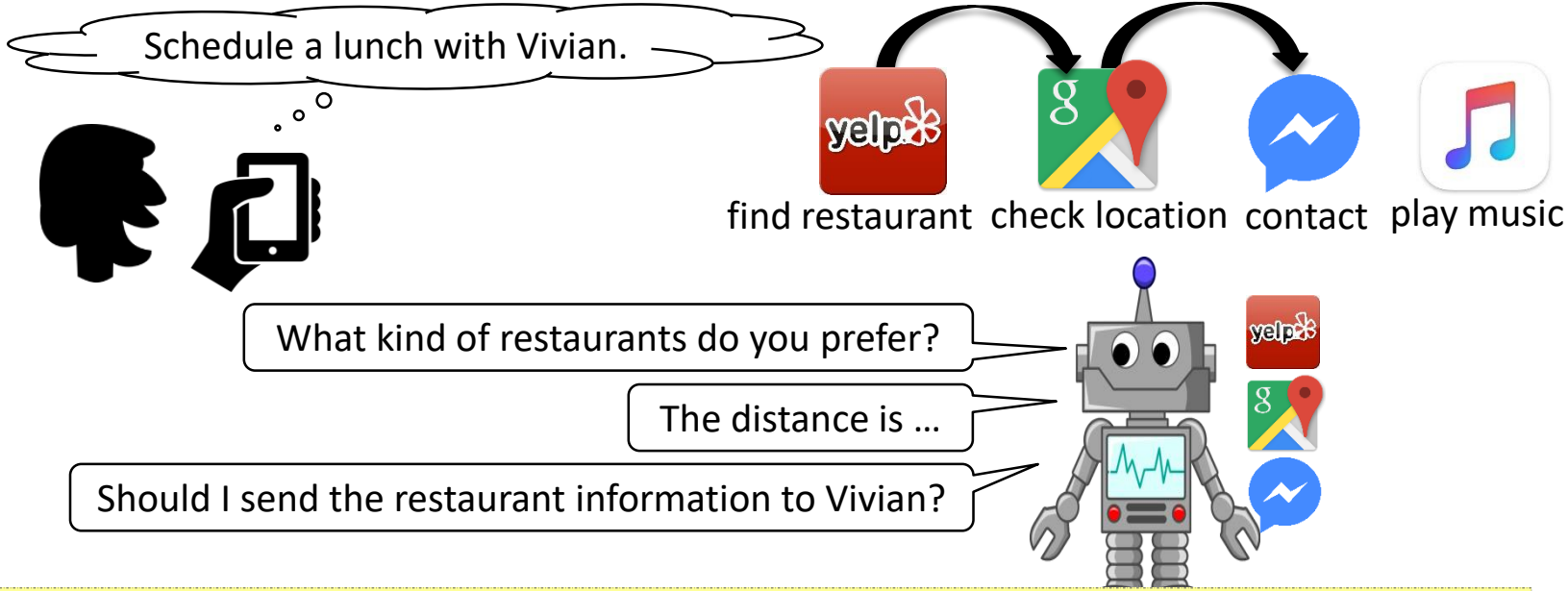
① **User preference**
- ✓ Some people prefer “Message” to “Email”
- ✓ Some people prefer “Ping” to “Text”

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; Sun et al., 2016)

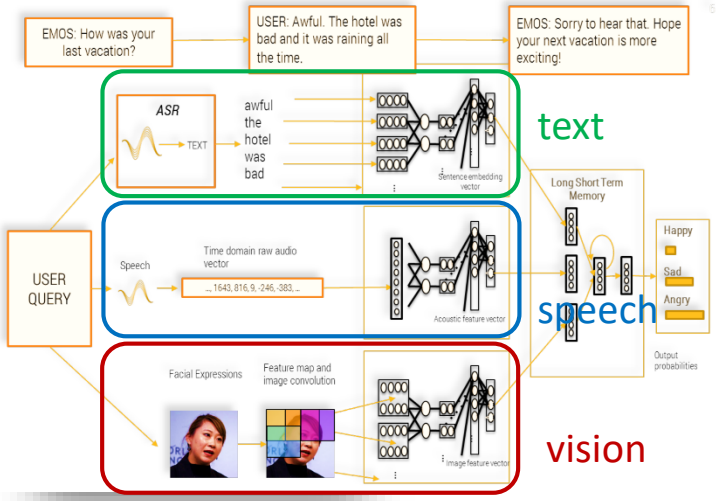
- High-level intention may span several domains



Users interact via high-level descriptions and the system learns how to plan the dialogues

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

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



Emotion Recognizer

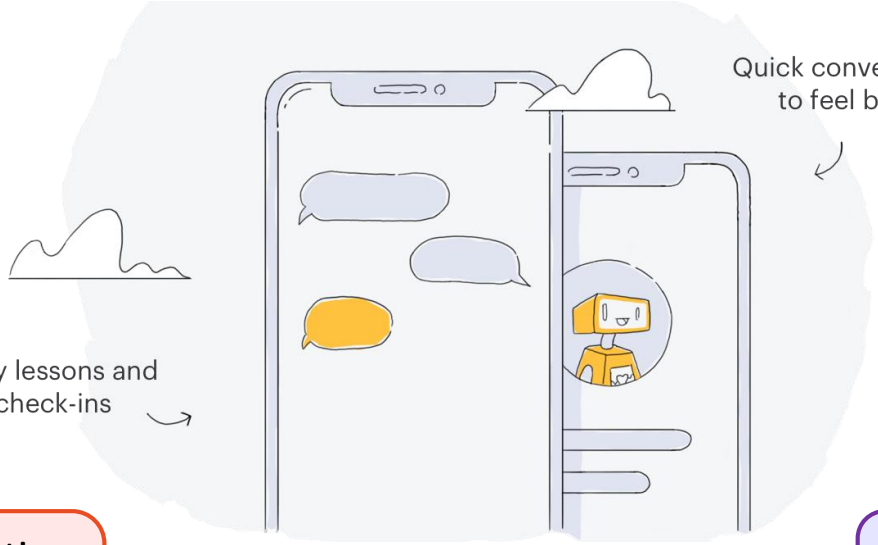
```
Face recognition output                                  (index):1728
                                                         (index):1729
 {
  "recognition": "Race: Asian Confidence: 65.42750000000001 Smiling:
3.95896 Gender: Female Confidence: 88.9369",
  "race": "Asian",
  "race_confidence": "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

Challenges and 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 requires high accuracy for working good as a whole

Her (2013)

What can machines achieve now or in the future?

Thanks for Your Attention!

Q & A

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