Task-Oriented Dialogue

Chi Qi, Michael Hu

What is task-oriented dialogue?

An Example of Task-Oriented Dialogue

GUS:	Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go ?
Client:	I want to go to San Diego on May 28
GUS:	What time do you want to leave?
Client:	I must be in San Diego before 10 am
GUS:	Would you like P.S.A. flight 102 that arrives at 9.15 am ?
Client:	What is the next flight ?
GUS:	Air California flight 310 that leaves at 8.30am and arrives at 10.00am
Client:	I'll take the first one
GUS:	In what name should I make a reservation, please?
Client:	Dan Bobrow
GUS:	I have confirmed the following flight: P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.
GUS:	What date do you want to return on ?
Client:	On Friday in the evening.
GUS:	Would you like the flight that leaves at 7.45 pm ?
Client:	That's fine.
GUS:	I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm Thank you for calling. Goodbye

Figure 26.8 The travel domain: A transcript of an actual dialogue with the GUS system of Bobrow et al. (1977). P.S.A. and Air California were airlines of that period.

GUS remembers things said in the past.

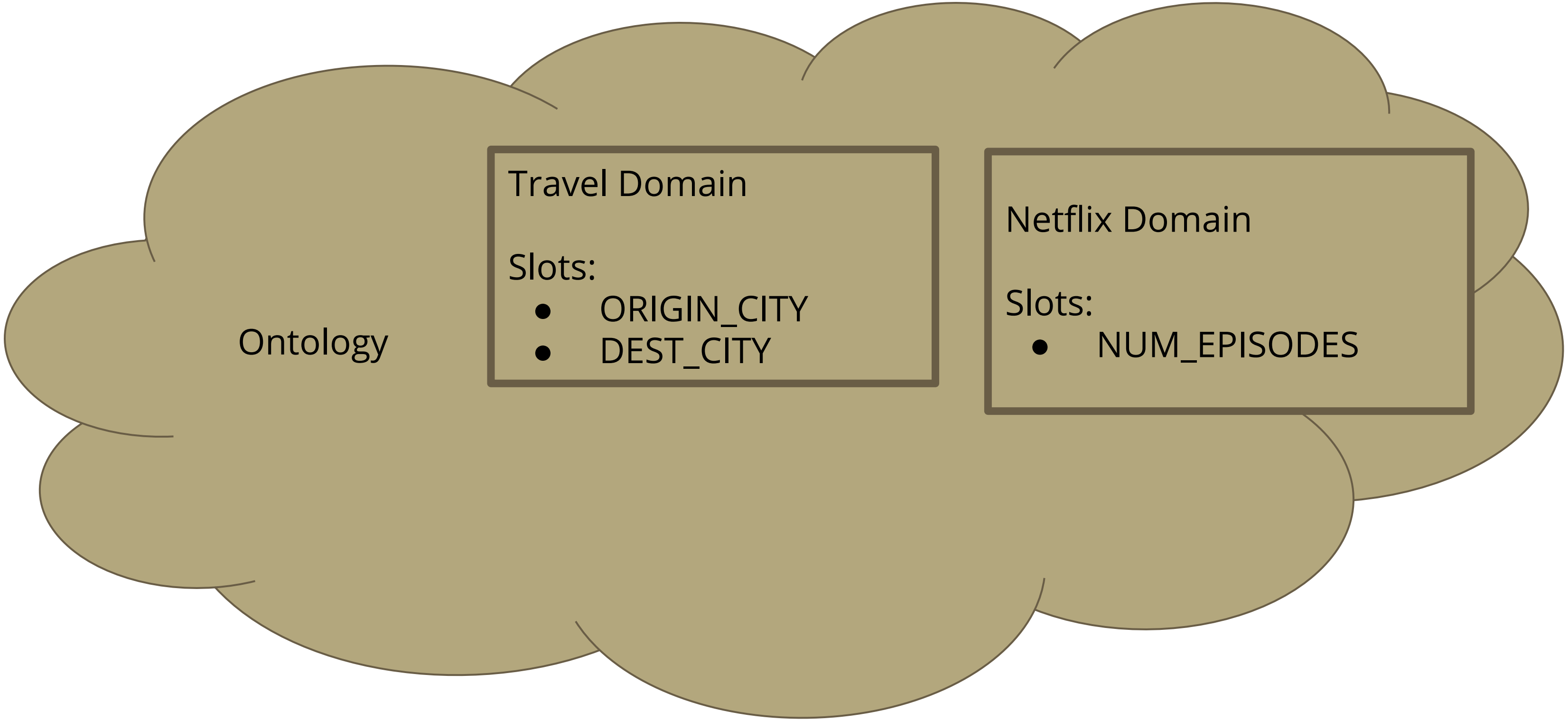
Task-Oriented Dialogue (TOD) Systems

- Help users achieve their specific goals
- Focus on understanding users, tracking states, and generating next actions.
- Minimize the number of turns: fewer turns the better

Key Terms

- **Domain ontology**: a set of knowledge structures representing the kinds of intentions the system can extract from user sentences.
- **Domain:** a domain consists of a collection of slots.
- **Slot:** each of slot can take a set of possible values.

Slot	**Type**	Question Template
ORIGIN CITY	city	“From what city are you leaving?”
DESTINATION CITY	city	“Where are you going?”
DEPARTURE TIME	time	“When would you like to leave?”
DEPARTURE DATE	date	“What day would you like to leave?”
ARRIVAL TIME	time	“When do you want to arrive?”
ARRIVAL DATE	date	“What day would you like to arrive?”



The domain ontology defines the set of actions our model can take.

The ontology file, specific all the values the three informable slots can take.

```
#######################################################
#######################################################
#  Copyright Cambridge Dialogue Systems Group, 2016 #
#######################################################
#######################################################
{
    "informable": {
        "area" : ["centre","north","west","south","east"],
        "food" : ["afghan","african","afternoon tea","asian
            oriental","australasian","australian","austrian","barbeque","basque"
            ,"belgian","bistro","brazilian","british","canapes","cantonese"
            ,"caribbean","catalan","chinese","christmas","corsica","creative"
            ,"crossover","cuban","danish","eastern
            european","english","eritrean","european","french","fusion","gastropub"
            ,"german","greek","halal","hungarian","indian","indonesian"
            ,"international","irish","italian","jamaican","japanese","korean"
            ,"kosher","latin american","lebanese","light
            bites","malaysian","mediterranean","mexican","middle eastern","modern
            american","modern eclectic","modern european","modern
            global","molecular gastronomy","moroccan","new zealand","north
            african","north american","north indian","northern
            european","panasian","persian","polish","polynesian","portuguese"
            ,"romanian","russian","scandinavian","scottish","seafood","singaporean"
            ,"south african","south indian","spanish","sri
            lankan","steakhouse","swedish","swiss","thai","the
            americas","traditional","turkish","tuscan","unusual","vegetarian"
            ,"venetian","vietnamese","welsh","world"],
        "pricerange" : ["cheap","moderate","expensive"]
    }
}
```

Natural language understanding for filling slots

“Show me morning flights from Boston to San Francisco on Tuesday”

- Task #1: Domain Classification
  - DOMAIN: AIR-TRAVEL
- Task #2: Intent Determination
  - INTENT: SHOW-FLIGHTS
- Task #3: Slot Filling
  - ORIGIN-CITY: Boston
  - ORIGIN-DATE: Tuesday
  - ORIGIN-TIME: morning
  - DEST-CITY: San Francisco

How is TOD different from other tasks?

1. Domain specificity.
    - A resulting challenge: lack of training data.
2. End goal: helping the user DO something.
    - Model must understand user & what they want
    - → Requires a deep understanding of dialogue progression
3. A focus on brevity and efficiency

Early Approaches

Approach 1: Rules-based systems.

Approach 2: Dialogue State Architecture

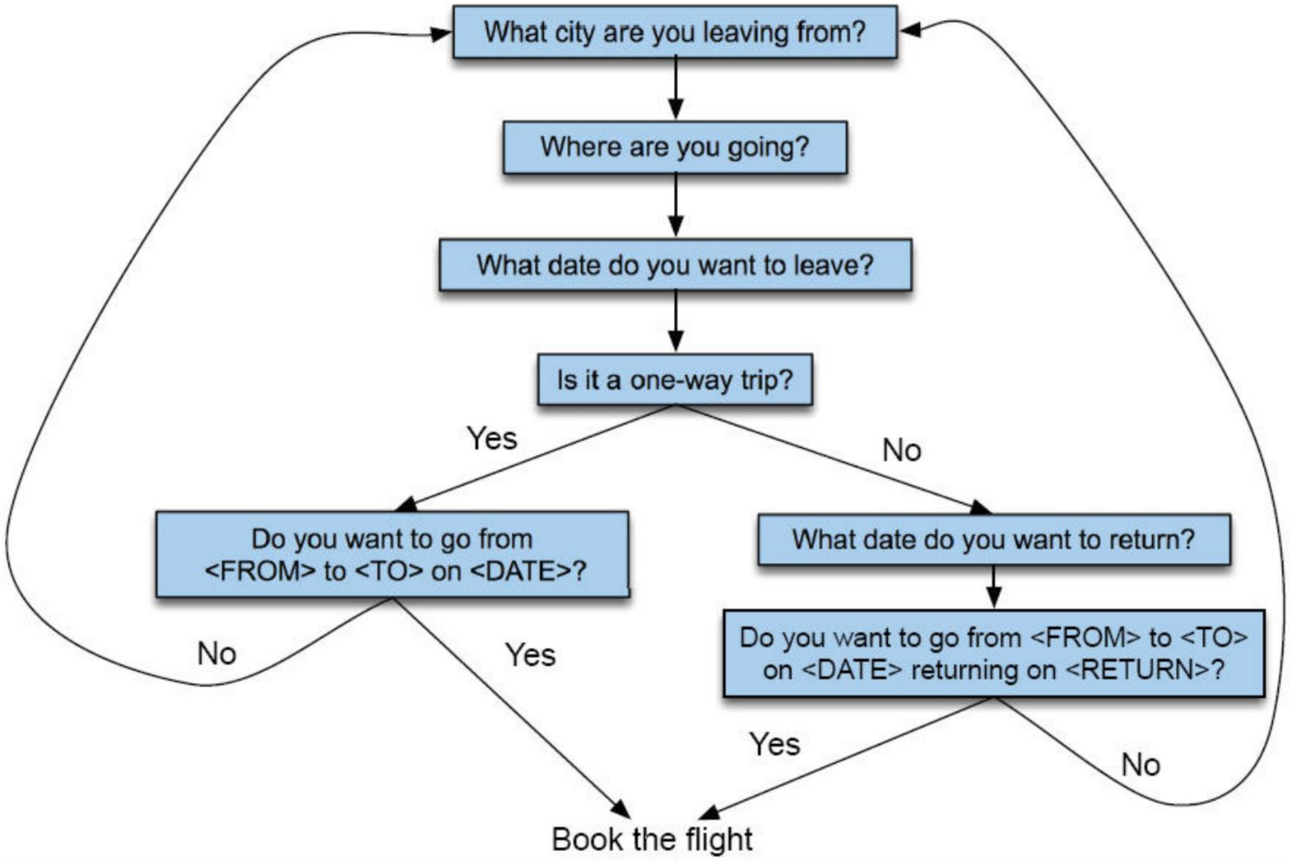
Rule-based systems

Consist of large hand-designed semantic grammars with thousands of rules.

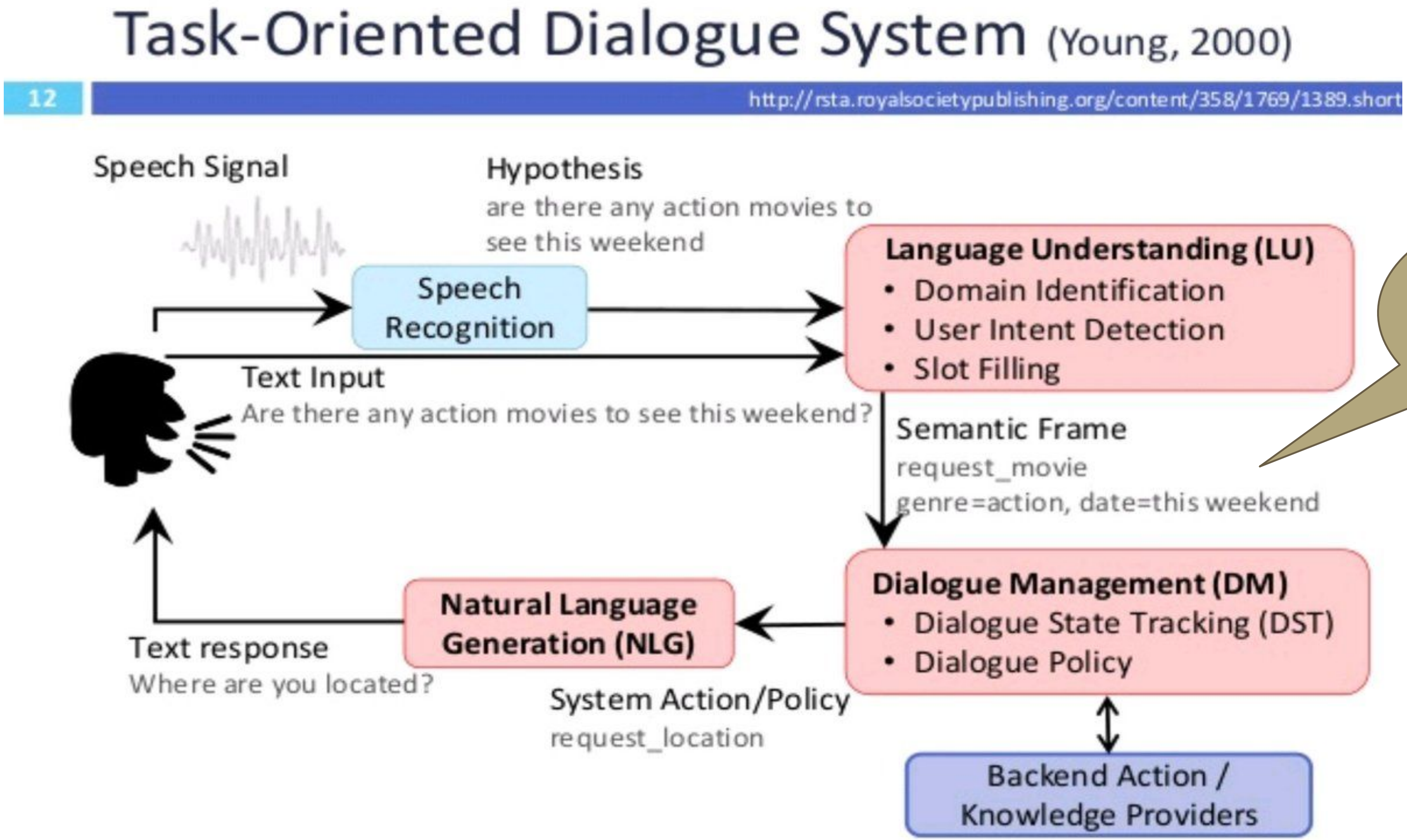
SHOW	$\rightarrow$	show me | i want | can i see|...
DEPART_TIME_RANGE	$\rightarrow$	(after|around|before) HOUR | morning | afternoon | evening
HOUR	$\rightarrow$	one|two|three|four...|twelve (AMPM)
FLIGHTS	$\rightarrow$	(a) flight | flights
AMPM	$\rightarrow$	am | pm
ORIGIN	$\rightarrow$	from CITY
DESTINATION	$\rightarrow$	to CITY
CITY	$\rightarrow$	Boston | San Francisco | Denver | Washington

E.g., Phoenix system (Ward and Issar, 1994)

Rule-based - finite state dialogue manager



The dialogue state architecture



Model is not end-to-end.

The dialogue state architecture - mostly data-driven

- **The Dialogue State Tracker** maintains the current state of the dialogue
- A more sophisticated **Dialogue Policy** compared to rule-based architecture
- A more sophisticated **Natural Language Generating** component

Data-driven vs Rule-based systems

★ Dialogue manager is more flexible and evolvable.
★ Learn from interaction - need more data, but less hand-craft rules
★ May have surprising/uncontrolled responses in unseen scenarios

Data collection

A core challenge of task oriented dialogue is getting relevant training data.

Solution: Wizard of Oz (WOZ) Data Collection

Wizard-of-Oz data collection: Users think they're talking to a computer, but they're actually talking to a human.

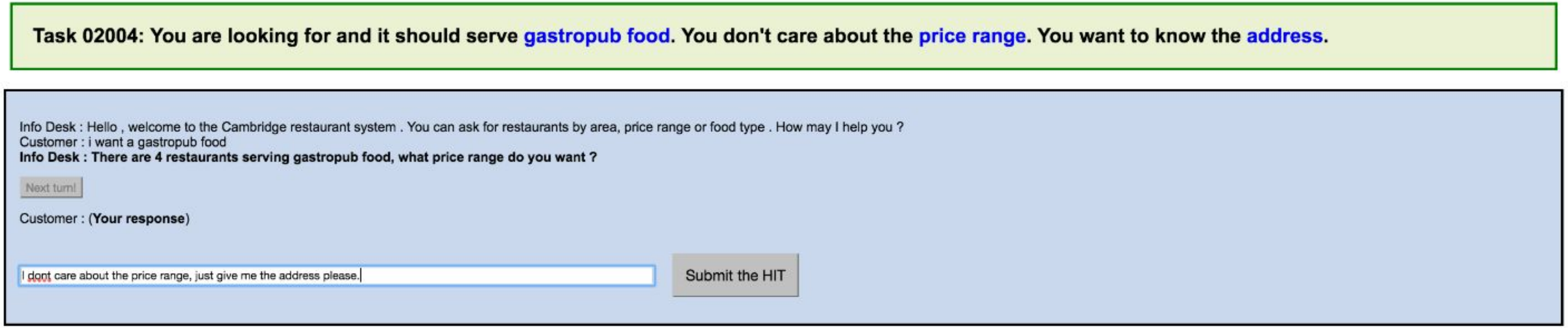
Humans pretending to be computers are called "wizards."

Circa 2016: Wen et al. needed training data relevant to restaurant selection in Cambridge, UK!

Solution: Amazon Mechanical Turk.

Data Collection

The flow: User 1 → Wizard 1 → User 2 → Wizard 2. Each person contributes 1 line to the conversation.



User Portal

Info Desk : Hello , welcome to the Cambridge restaurant system . You can ask for restaurants by area, price range or food type . How may I help you ?
Customer : i want a gastropub food
Info Desk : There are 4 restaurants serving gastropub food, what price range do you want ?
Customer : i dont care

Next turn!

Please **modifiy** the following answers based on the latest customer response:

- What does user want?

What is the **food type** the user wants? gastropub
What is the **area** the user wants? not mentioned
What is the **price range** the user wants? dont care

- What does user ask?

Is the user asking for **food type** of an offered venue? No
Is the user asking for **price range** of an offered venue? No
Is the user asking for **area** of an offered venue? No
Is the user asking for **postcode** of an offered venue? No
Is the user asking for **phone number** of an offered venue? No
Is the user asking for **address** of an offered venue? No
Is the user mentioning any **restaurant names**? No

finish

Info Desk : (**Your response**)

I would recommend backstreet bistro, a great gastropub restaurant in the centre. do you want their phone number ? end-of-dialogue? Submit the HIT

Name	Food	Area	Price Range	Phone	Address	Postcode
backstreet bistro	gastropub	centre	expensive	01223 306306	2 Sturton Street City Centre	C.B 1, 2 Q.A
royal standard	gastropub	east	expensive	01223 247877	290 Mill Road City Centre	C.B 1, 3 N.L
the cow pizza kitchen and bar	gastropub	centre	moderate	01223 308871	Corn Exchange Street	C.B 2, 3 Q.F
the slug and lettuce	gastropub	centre	expensive	--	34 - 35 Green Street	C.B 2, 3 J.U
nil	gastropub	nil	nil	nil	nil	nil

Showing 1 to 4 of 4 entries (filtered from 110 total entries) Previous 1 Next

Wizard Portal

Data Collection

Resulting training data is very domain specific: both good and bad.

1500 total dialogue turns → 680 total dialogues.

60-20-20 data split. Result: Training set of 408 dialogues.

Cost: $400. About $1 per training example.

Data collection is the main bottleneck of this task. Can we do better?

More data: Multi-domain WOZ (MultiWOZ)

- EMNLP 2018
- Setting: a tourist and a clerk
- Same collection method as Cambridge data set
- Seven domains (**Hotel, Train, Attraction, Restaurant, Taxi, Hospital, and Police**) and 16 slots (food, leave at, area, etc).
- **MultiWOZ**: the largest human-human conversational corpus with Dialogue State Tracking labels (8438 dialogues with avg 13.68 turns).

Dataset comparison

Metric	DSTC2	SFX	WOZ2.0	FRAMES	KVRET	M2M	MultiWOZ
# Dialogues	1,612	1,006	600	1,369	2,425	1,500	**8,438**
Total # turns	23,354	12,396	4,472	19,986	12,732	14,796	**113, 556**
Total # tokens	199,431	108,975	50,264	251,867	102,077	121,977	**1,490,615**
Avg. turns per dialogue	14.49	12.32	7.45	**14.60**	5.25	9.86	13.46
Avg. tokens per turn	8.54	8.79	11.24	12.60	8.02	8.24	**13.13**
Total unique tokens	986	1,473	2,142	12,043	2,842	1,008	**23689**
# Slots	8	14	4	**61**	13	14	24
# Values	212	1847	99	3871	1363	138	**4510**

Evaluation

Evaluating Task-Oriented Dialogue Systems is also a challenging task.

Human-based evaluation

- **Lab-experiments:** Users were invited to participate in the lab where they interacted with the dialogue system and subsequently filled a questionnaire [Young et al., 2010]. - very controlled, not comparable to real world
- **In-field experiments:** collecting feedback from real users of the dialogue systems - e.g., the Spoken Dialogue Challenge [Black et al., 2011]
- **Crowdsourcing:** using crowdsourcing platforms such as Amazon Mechanical Turk (AMT) - high variability of user behaviour

Difficult to set-up and to carry out:

the users need to be properly instructed, the tasks need to be prepared so that the experiment is close to real-world conditions.

Automated Evaluation metrics

- **Dialogue State Tracker performance**
    - End-to-end: Precision, Recall, F-1
    - TRADE: joint and slot accuracy
- **Dialogue Efficiency** [# turns]
- **Corpus Based Evaluation for e2e system**
    - BLEU score
    - Entity matching rate

Checkpoint 1. Questions?

- Introduction to TOD
- Early approaches
- Data collection
- Evaluation

This is a challenging task.

Seq-to-seq for TOD

Now that we've introduced TOD and its core challenges, let's motivate the 1st paper.

A new idea: seq-to-seq for TOD

(Sutskever et al., 2014) <- Machine Translation, Elisabetta & Ben

(Vinyals & Le, 2015) <- Dialogue, Xinyi & Paula

Wen et al: Let's use seq-to-seq for task-oriented dialogue!

A Network-based End-to-End Trainable Task-oriented Dialogue System

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšic, Milica Gašic, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young

I need to find an expensive restaurant that's in the south of the city.

Act: inform

There are several restaurants in the south part of town that serve expensive food. Do you have a cuisine preference?

Domain: Restaurant.
Slot {Price: Expensive, Area: south}

No I don't care about the type of cuisine.

Act: inform

Chiquito Restaurant Bar is a Mexican restaurant located in the south part of town.

Slot {Price: Expensive, Area: south; Food: Don't care}

Where is it?

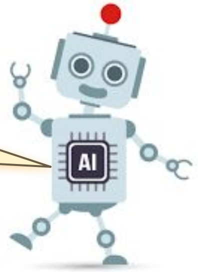
Act: request

There address is 2G Cambridge Leisure Park Cherry Hinton Road Cherry Hinton, it there anything else I can help you with?

Nope, that's what I was looking for. Thanks a lot!

Act: inform

Thank you very much for using the Cambridge Restaurant System.



Setting

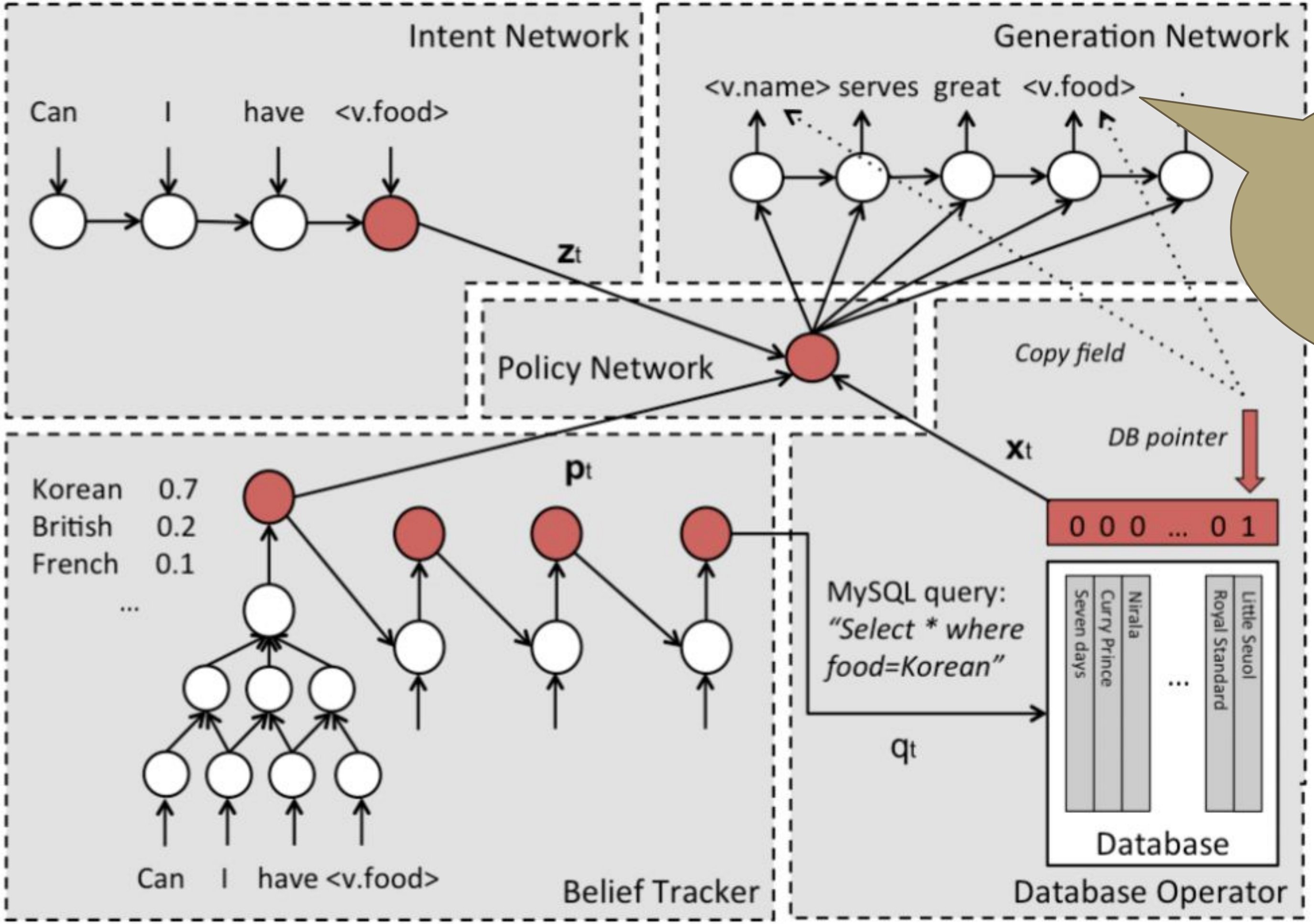
There are 110 restaurants in the DB, each with 9 attributes.

3 Informable slots (constraints)

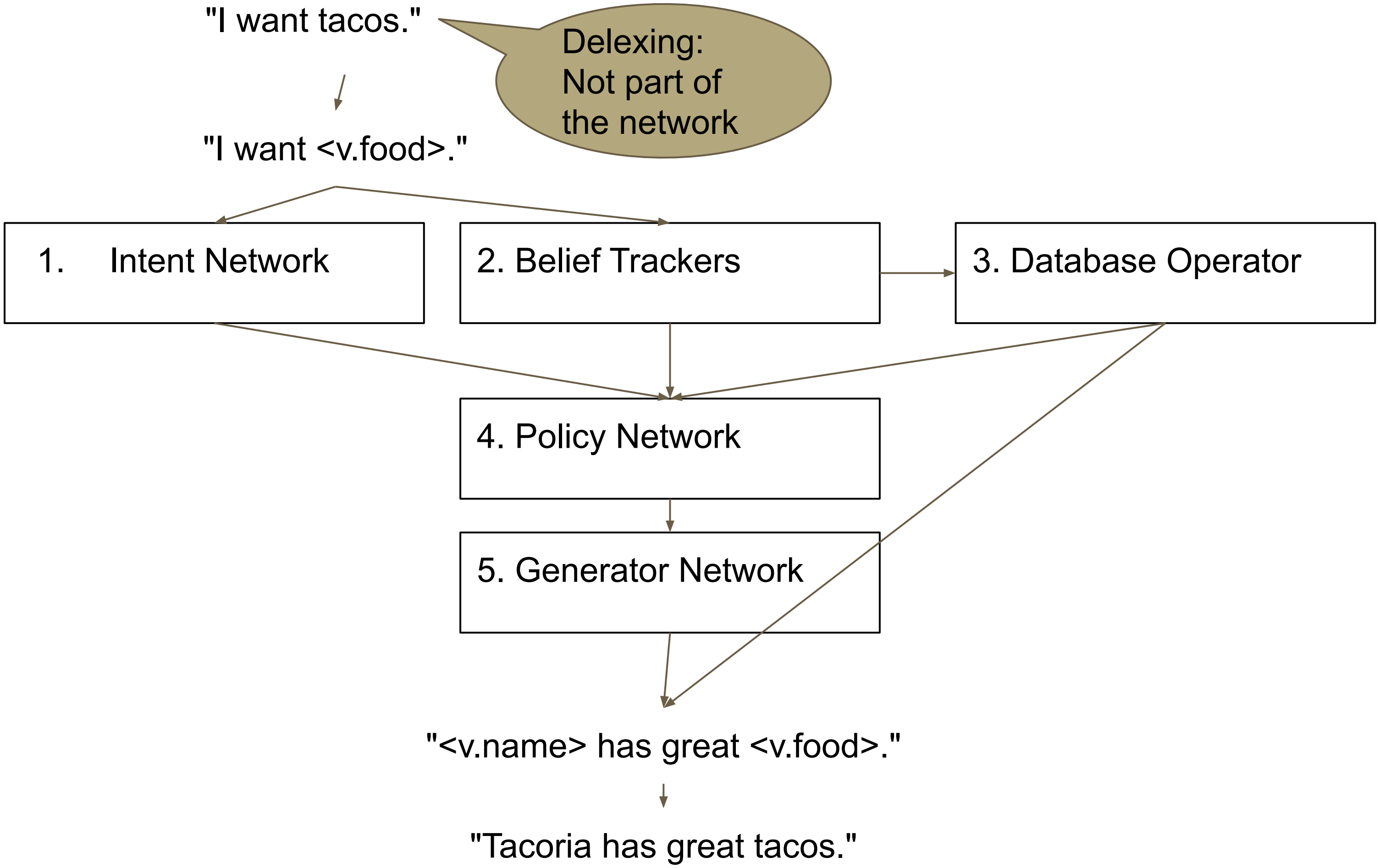
- food type, price range, area

6 Requestable slots (follow-up questions)

- address, phone number, area code
- food type, price range, area

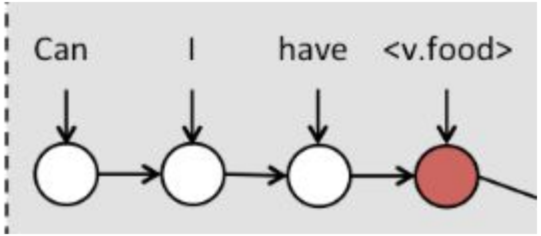


Delexicalization: replacing specific values with generic tokens. Allows for weight sharing.


"I want tacos."
Delexing:
Not part of
the network
"I want <v.food>."
1. Intent Network
2. Belief Trackers
3. Database Operator
4. Policy Network
5. Generator Network
"<v.name> has great <v.food>."
"Tacoria has great tacos."

Wen et al.'s model combines SOTA subnetworks into one big model with impressive performance.

1. Intent Network



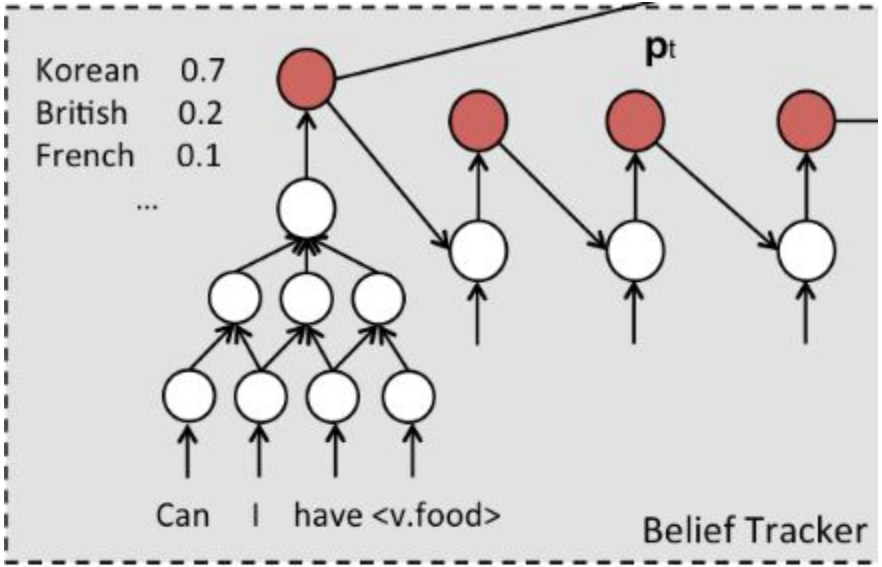
Pretty straightforward: the encoder of a classic seq-to-seq model.

Role: Natural Language Understanding

Authors tried:

- LSTM
- CNN

2. Belief Tracker



Belief Tracker

Maps input sequence to a distribution over values.
Slot-value pairs are things like price $\rightarrow$ expensive, food type $\rightarrow$ Tex-mex, etc.

Role: Dialogue State Tracker.

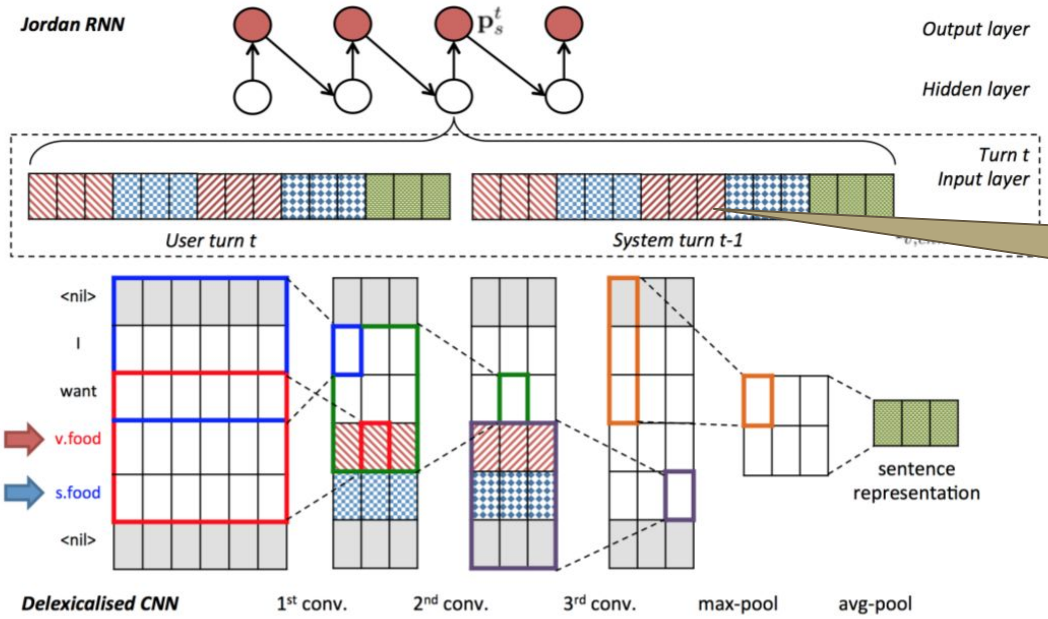
Intent network $\rightarrow$ sentence level

Belief network $\rightarrow$ conversation level

Belief Trackers

The model uses a belief tracker **per slot.**

This doesn't scale.
What if your DB
has 100 rows?



Model samples
representations
from all 3 layers.

CNN output

distribution from last turn

probability the slot hasn't been mentioned

$$\mathbf{f}_v^t = \mathbf{f}_{v,cnn}^t \oplus p_v^{t-1} \oplus p_\emptyset^{t-1} \quad (3)$$

$$g_v^t = \mathbf{w}_s \cdot \text{sigmoid}(\mathbf{W}_s \mathbf{f}_v^t + \mathbf{b}_s) + b_s' \quad (4)$$

RNN

$$p_v^t = \frac{\exp(g_v^t)}{\exp(g_{\emptyset,s}) + \sum_{v' \in V_s} \exp(g_{v'}^t)} \quad (5)$$

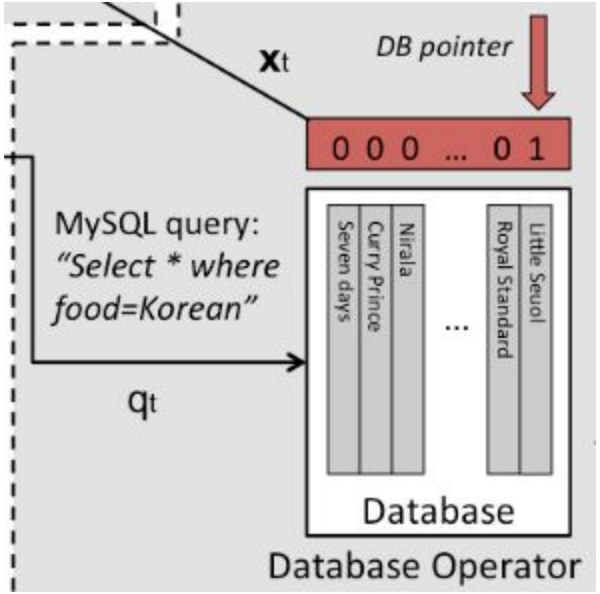
Softmax

Output: Tex-Mex 0.6, Chinese 0.1, ...

You can think of belief trackers as long range parsers.

3. Database Operator

$$q_t = \bigcup_{s' \in S_I} \{\operatorname*{argmax}_v \mathbf{p}^t_{s'}\}$$



Take the most likely values out of each of 3 informable slots, and write as a SQL query.

Using query results, assign a {0, 1} vector over the fields in the database. 1 = relevant.

Finally, point to an entity at random. This entity has an associated phone number, price point, etc.

Role: Dialogue Policy

4. Policy Network

A feed-forward layer. The glue holding all the gradients together.

The "conditioner"

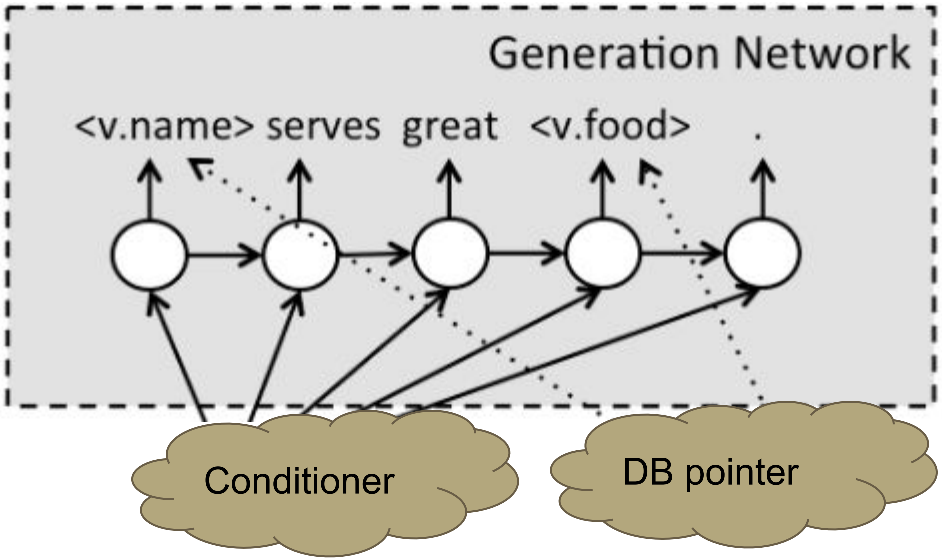
$$\mathbf{o}_t = \tanh(\mathbf{W}_{zo}\mathbf{z}_t + \mathbf{W}_{po}\hat{\mathbf{p}}_t + \mathbf{W}_{xo}\hat{\mathbf{x}}_t)$$

z_t : intent network output

$\hat{p}_t$: belief vector for each slot

$\hat{x}_t$: number of DB hits, as a one-hot vector

5. Generation Network



Generation Network

Step 1: Generate auto-regressively using an LSTM

$$P(w_{j+1}^t | w_j^t, \mathbf{h}_{j-1}^t, \mathbf{o}_t) = \mathrm{LSTM}_j(w_j^t, \mathbf{h}_{j-1}^t, \mathbf{o}_t)$$ Conditioner

Step 2: Replace delexicalized tokens with DB pointer values.

<v.name> has great <v.food>. => Tacoria has great tacos.

Role: Natural Language Generation

Optimized: The Attentive Generation Network

$$\alpha_s^{(j)} = \mathrm{softmax}\big(\mathbf{r}^\intercal \tanh(\mathbf{W}_r \cdot \mathbf{u}_t)\big) \qquad (12)$$

Compute attention weights by looking at literally all the representations that we have.

$$\hat{\mathbf{p}}_t^{(j)} = \sum_{s \in \mathbb{G}} \alpha_s^{(j)} \tanh(\mathbf{W}_{po}^s \cdot \hat{\mathbf{p}}_s^t) \qquad (11)$$

Use attention weights to recompute probability distribution.

$$\mathbf{o}_t^{(j)} = \tanh(\mathbf{W}_{zo}\mathbf{z}_t + \hat{\mathbf{p}}_t^{(j)} + \mathbf{W}_{xo}\hat{\mathbf{x}}_t) \qquad (10)$$

Recompute the conditioning vector.

Model Training

Step 1: Train belief networks using CEL between wizard labels and belief network distributions.

- Train on dialogue state.

Step 2: Train end-to-end using CEL between wizard sentences and machine predictions.

- Train on response.

```
"usr": {
    "transcript": "I need to find an expensive restauant that's in the south section of the city.",
    "slu": [
        {
            "act": "inform",
            "slots": [
                [
                    "pricerange",
                    "expensive"
                ]
            ]
        },
        {
            "act": "inform",
            "slots": [
                [
                    "area",
                    "south"
                ]
            ]
        }
    ]
},
"sys": {
    "sent": "There are several restaurants in the south part of town that serve expensive food. Do you have a cuisine preference?",
```

Decoding

Beam search with beam size 10.

$$m_t^* = \underset{m_t}{\operatorname{argmax}}\{\log p(m_t|\theta, u_t)/J_t\}$$

Highest per-token log probability

$$m_t^* = \underset{m_t}{\operatorname{argmax}}\{\log p(m_t|\theta, u_t)/J_t - \lambda \log p(m_t)/J_t + \gamma R_t\}$$

"Weighted" decoding

Reward heuristic:
More reward if the model generates an address when an address is requested

Use a separate language model to predict probability of generating each word

**n-grams go up to trigram.

Evaluation: Belief Trackers

Tracker type	Informable			Requestable		
	Prec.	Recall	F-1	Prec.	Recall	F-1
cnn	99.77%	96.09%	97.89%	98.66%	93.79%	96.16%
ngram	99.34%	94.42%	96.82%	98.56%	90.14%	94.16%

Conclusion: Belief trackers learn how to parse commands into a distribution over slot values.

Precision: % time requested slot value returned.
Recall: % of info returned that was actually requested.

Evaluation: Models

Quantitative metrics: BLEU, entity match rate, and success rate.

- BLEU: computed on delexicalized forms
- Entity match rate: % recommendations of correct type:
  - E.g. You ask for tacos, and the model recommends Tacoria
- Success rate: % time entity matches, **and** all follow-up questions are answered.

Qualitative metrics, out of 5: comprehension, naturalness

Evaluation: Models

Encoder	Tracker	Decoder	Match(%)	Success(%)	T5-BLEU	T1-BLEU
Baseline						
lstm	-	lstm	-	-	0.1650	0.1718
lstm	turn recurrence	lstm	-	-	0.1813	0.1861
Variant	No requestable trackers					
lstm	rnn-cnn, w/o req.	lstm	89.70	30.60	0.1769	0.1799
cnn	rnn-cnn	lstm	88.82	58.52	0.2354	0.2429
Full model w/ different decoding strategy						
lstm	rnn-cnn	lstm	86.34	75.16	0.2184	0.2313
lstm	rnn-cnn	+ weighted	86.04	78.40	0.2222	0.2280
lstm	rnn-cnn	+ att.	90.88	80.02	0.2286	0.2388
lstm	rnn-cnn	+ att. + weighted	90.88	83.82	0.2304	0.2369

No DB access

Top performer

Note: A low BLEU score is okay, as long as success rate is high.
We measure success and BLEU using delexicalized forms.

The model learns something!

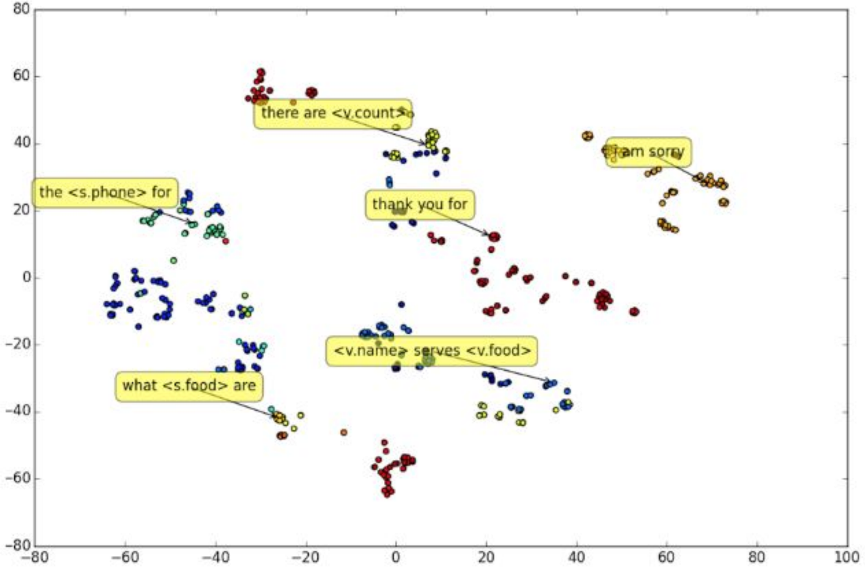


Figure 3: The action vector embedding $\mathbf{o}_t$ generated by the NN model w/o attention. Each cluster is labelled with the first three words the embedding generated.

Clusters generated with t-SNE. t-SNE: a probabilistic cousin of PCA.

Table 3: Human assessment of the NN system. The rating for comprehension/naturalness are both out of 5.

Metric	NN
Success	98%
Comprehension	4.11
Naturalness	4.05
# of dialogues:	245

Table 4: A comparison of the NN system with a rule-based modular system (*HDC*).

Metric	NDM	HDC	Tie
Subj. Success	96.95%	95.12%	-
Avg. # of Turn	3.95	4.54	-
Comparisons(%)			
Naturalness	46.95*	25.61	27.44
Comprehension	45.12*	21.95	32.93
Preference	50.00*	24.39	25.61
Performance	43.90*	25.61	30.49

Limitations

The model:

- Cannot not handle noisy dialogue.
- Cannot ask user for clarification.
- Gives only 1 recommendation at a time, by construction.
- Cannot generalize. (A limitation our 2nd paper tries to address!)

Takeaways

Wen et al.'s model:

1. Demonstrates that a seq-to-seq approach can work for the task-oriented dialogue task.
2. Composes several SOTA models end-to-end.
3. Defines a procedure for generating data.

A successful proof of concept!

How many different components are there in a task-oriented dialogue system? How are each component and the full system evaluated?

How many different components are there in a task-oriented dialogue system? How are each component and the full system evaluated?

There are 4 components.

1. Natural language understanding, evaluated end-to-end.
2. Dialogue state tracking, evaluated with F1 score.
3. Dialogue policy, evaluated end-to-end.
4. Natural language generation, evaluated end-to-end.

The full model is evaluated on BLEU, entity match rate, success rate, and qualitative metrics such as naturalness.

Checkpoint 2. Questions?

Generalization in TOD

How can we make our model more robust to unseen slot values?

How can we reduce dependence on an ontology?

Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, Pascale Fung

Dialogue State Tracking (DST)

- A key subtask in dialogue management - Representation of the system's belief of the user's goal(s) at any time during the dialogue given the dialogue history.

Challenges:

- ❖ How to define the state space?
- ❖ How to tractably maintain the dialogue state?
- ❖ Which actions to take for each state?
- ❖ Multi-domain, Multi-turn Conversations?

Dialogue State Tracking (DST)

	Utterance	Food	Area
S	Hello, How may I help you?	Persian	South
U	I need a Persian restaurant in the south part of town.		
S	What kind of food would you like?	Persian	South
U	Persian.		
S	I'm sorry but there is no restaurant serving persian food	Portuguese	South
U	How about Portuguese food?		
S	Peking restaurant is a nice place in the south of town.	Portuguese	South
U	Is that Portuguese?		
S	Nandos is a nice place in the south of town serving tasty Portuguese food.	Portuguese	South
U	Alright. Whats the phone number?		
S	The phone number of nandos is 01223 327908 .	Portuguese	South
U	And the address?		
S	Sure, nandos is on Cambridge Leisure Park Clifton Way.	Portuguese	South
U	Thank you good bye.		

An example of multi-domain DST

I'm looking for a cheap pizza restaurant in the city center.

Sure. There is D'angelo Pizzeria nearby. How many guests?

Restaurant:
{Price: Cheap,
Type: Pizza,
Area: Center}

An example of multi-domain DST

I'm looking for a cheap pizza restaurant in the city center.

Restaurant:
{Price: Cheap,
Type: Pizza,
Area: Center}

Sure. There is D'angelo Pizzeria nearby. How many guests?

Three people Wednesday at 11am please. Please make sure there's NO PINEAPPLE on the pizza!

Restaurant:
{People: Three, Day: Wednesday, Time: 11am, Others: No pineapple}

Booked! QWERT is your reservation code.

An example of multi-domain DST

I'm looking for a cheap pizza restaurant in the city center.

Restaurant: {Price: Cheap, Type: Pizza, Area: Center}

Sure. There is D'angelo Pizzeria nearby. How many guests?

Three people Wednesday at 11am please. Please make sure there's NO PINEAPPLE on the pizza!

Restaurant: {People: Three, Day: Wednesday, Time: 11am, Others: No pineapple}

Booked! QWERT is your reservation code.

Also looking for some architectural attractions close to the restaurant.

Attraction: {Area: Center, Type: Architectural}

All Saints Church is famous. Would you like to head there?

An example of multi-domain DST

I'm looking for a cheap pizza restaurant in the city center.

Restaurant: {Price: Cheap, Type: Pizza, Area: Center}

Sure. There is D'angelo Pizzeria nearby. How many guests?

Three people Wednesday at 11am please. Please make sure there's NO PINEAPPLE on the pizza!

Restaurant: {People: Three, Day: Wednesday, Time: 11am, Others: No pineapple}

Booked! QWERT is your reservation code.

Also looking for some architectural attractions close to the restaurant.

Attraction: {Area: Center, Type: Architectural}

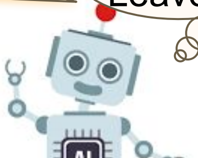
All Saints Church is famous. Would you like to head there?

Yes help me book a taxi between the restaurant and the church.

Taxi: {Destination: All Saints Cathedral, Departure: D'angelo, Leave at: 1:30 pm}

What time do you need the taxi?

Around 1:30 pm please.



Ontology-based DST

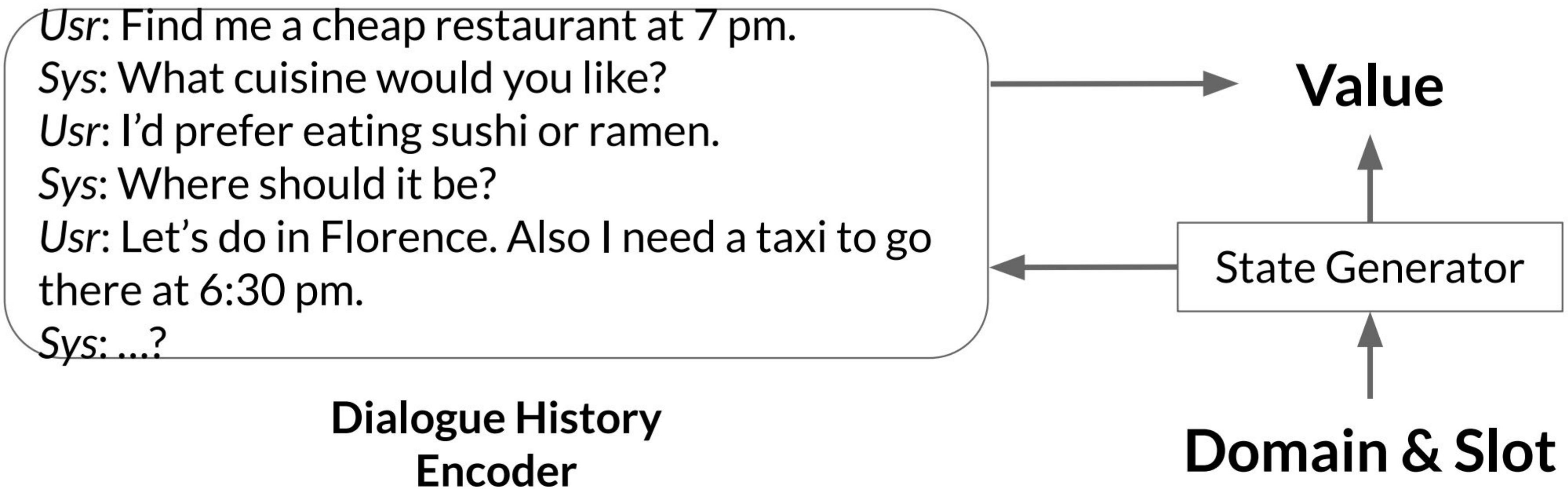
- Given system response and current user utterance, each slot in each domain is predicted to be one of the **predefined values** in **ontology** (e.g., the belief tracker in Wen et al. 2016).

Challenges:

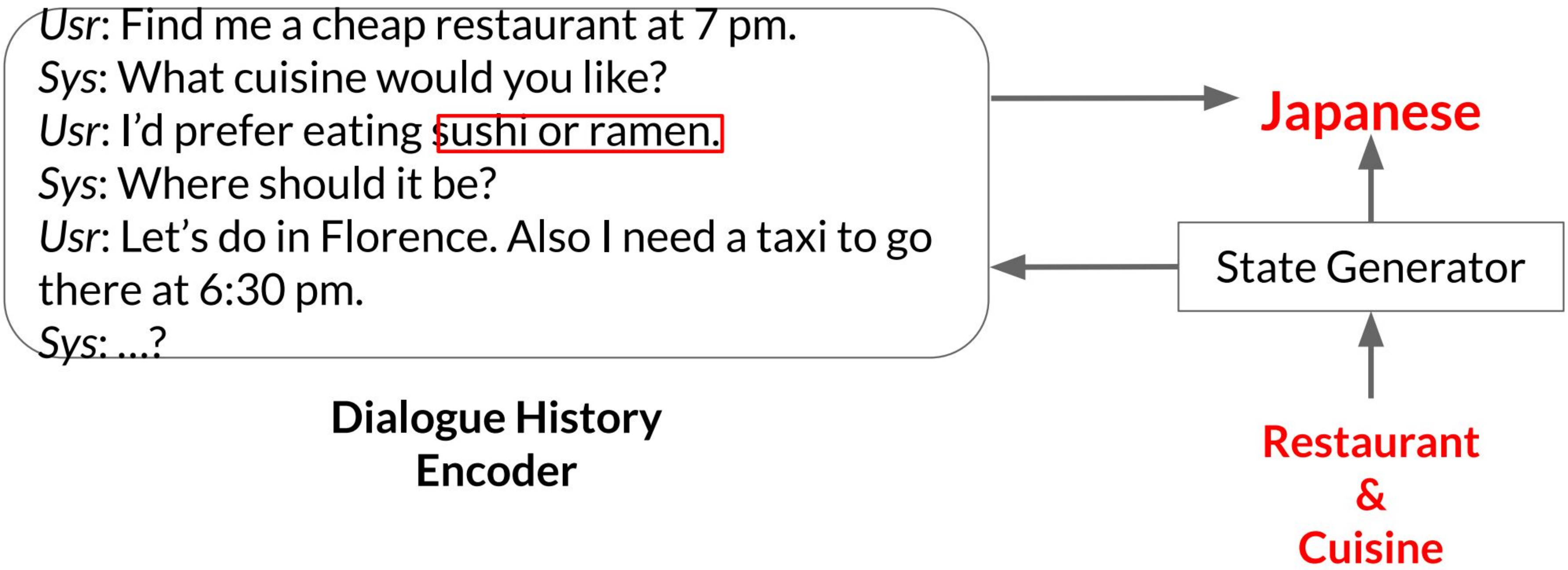
❖ Ontology is hard to obtain in real scenarios
❖ Need to track lots of slot values
❖ Cannot track unseen slot values
❖ Missing domain sharing capacities

DST without ontology?

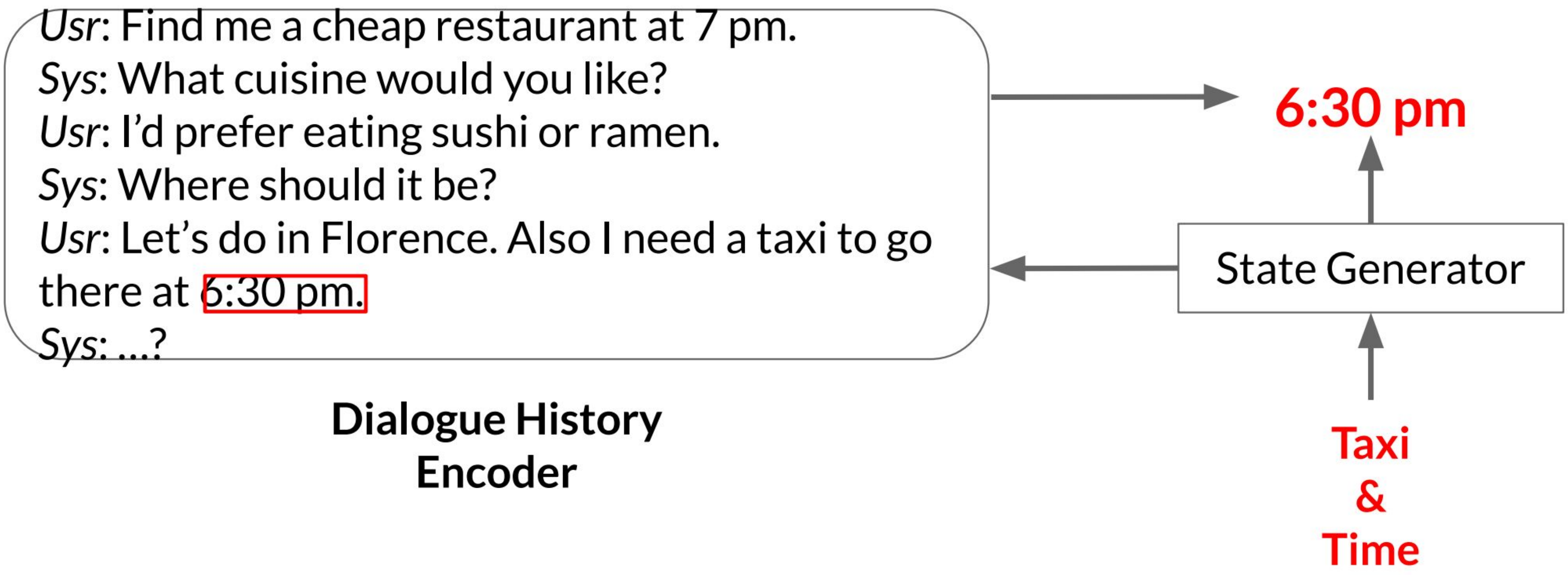
DST without ontology intuition



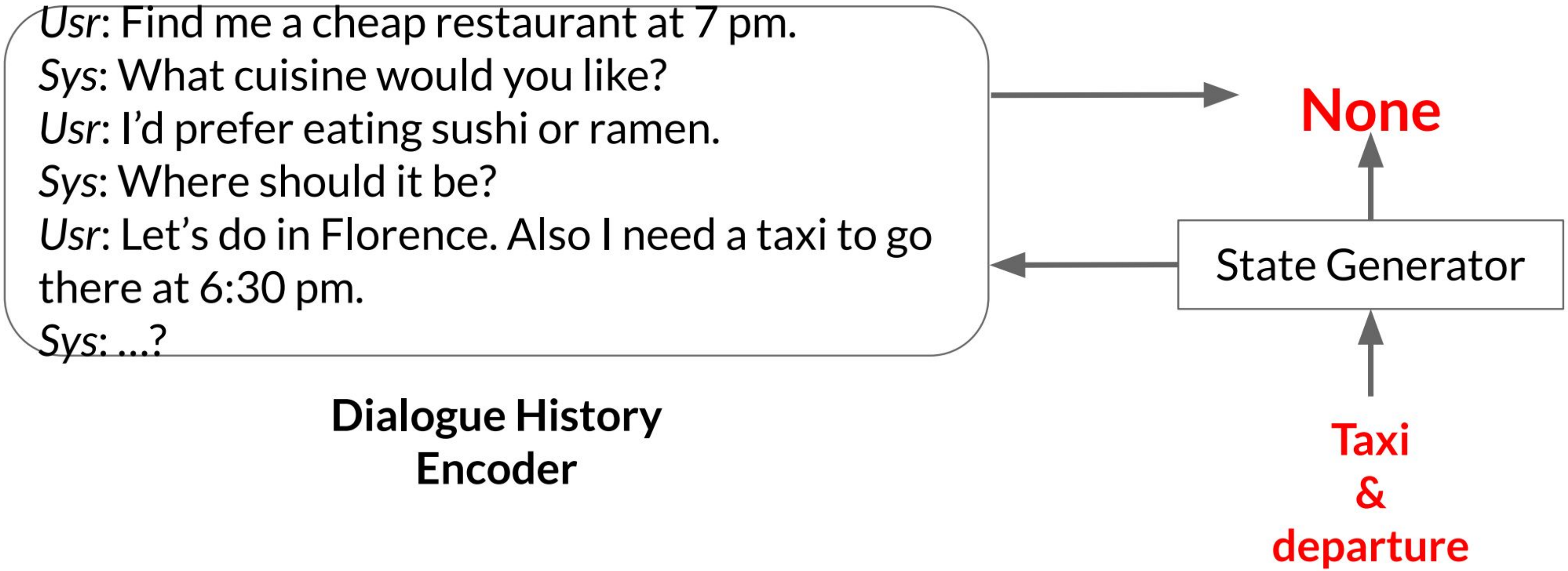
DST without ontology intuition



DST without ontology intuition

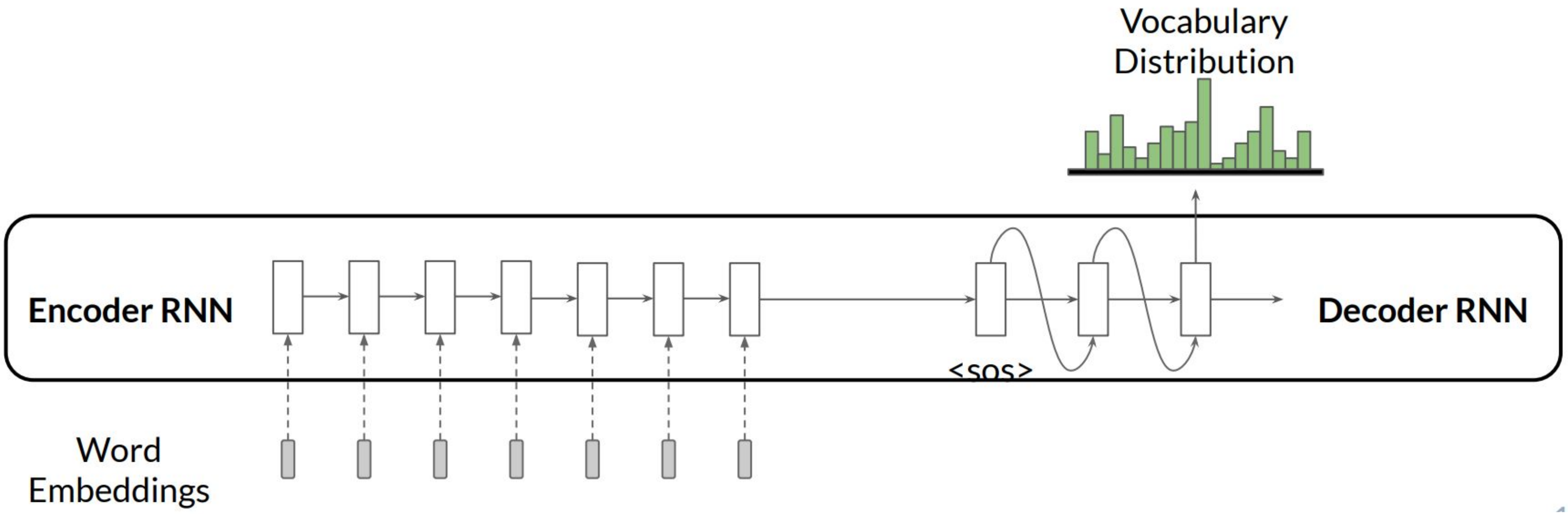


DST without ontology intuition



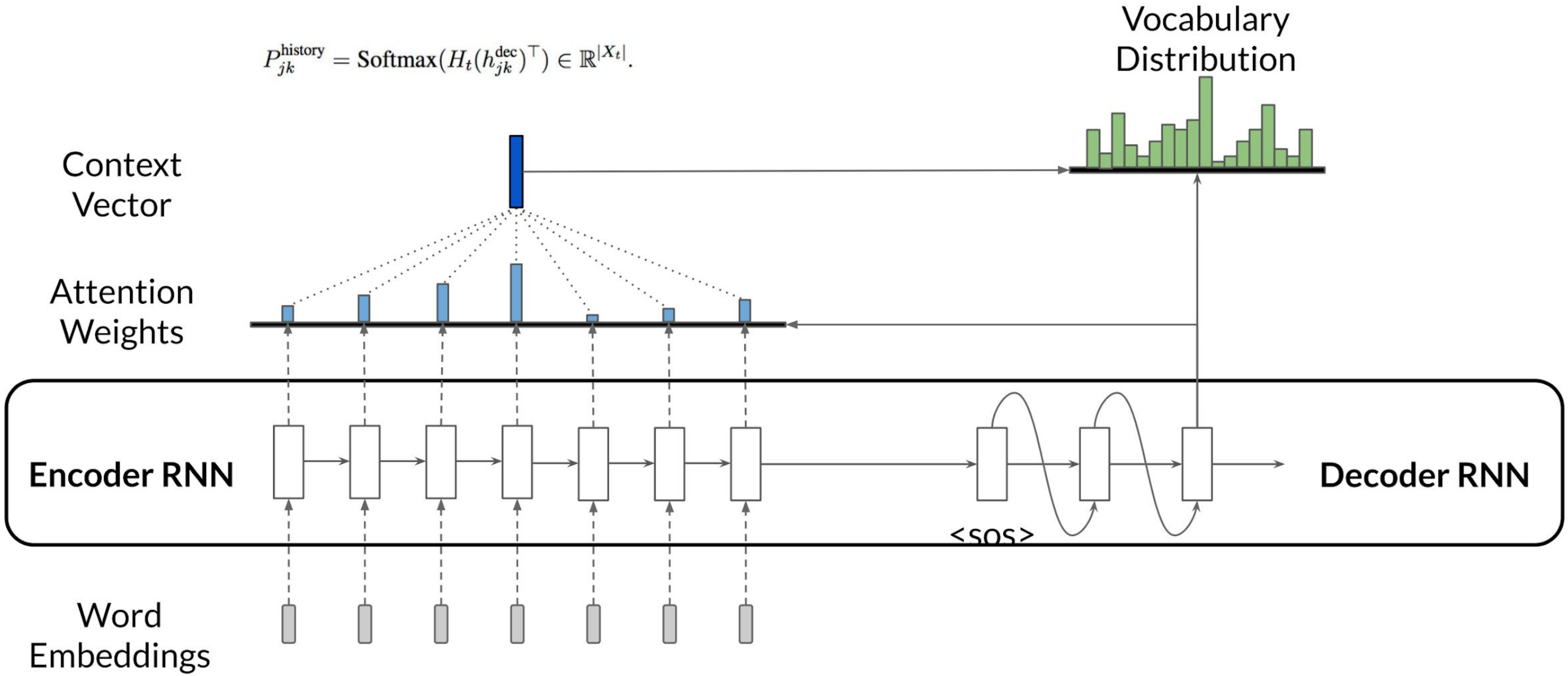
Sequence-to-Sequence (Seq2Seq)

$$P_{jk}^{\text{vocab}} = \text{Softmax}(E(h_{jk}^{\text{dec}})^\top) \in \mathbb{R}^{|V|},$$



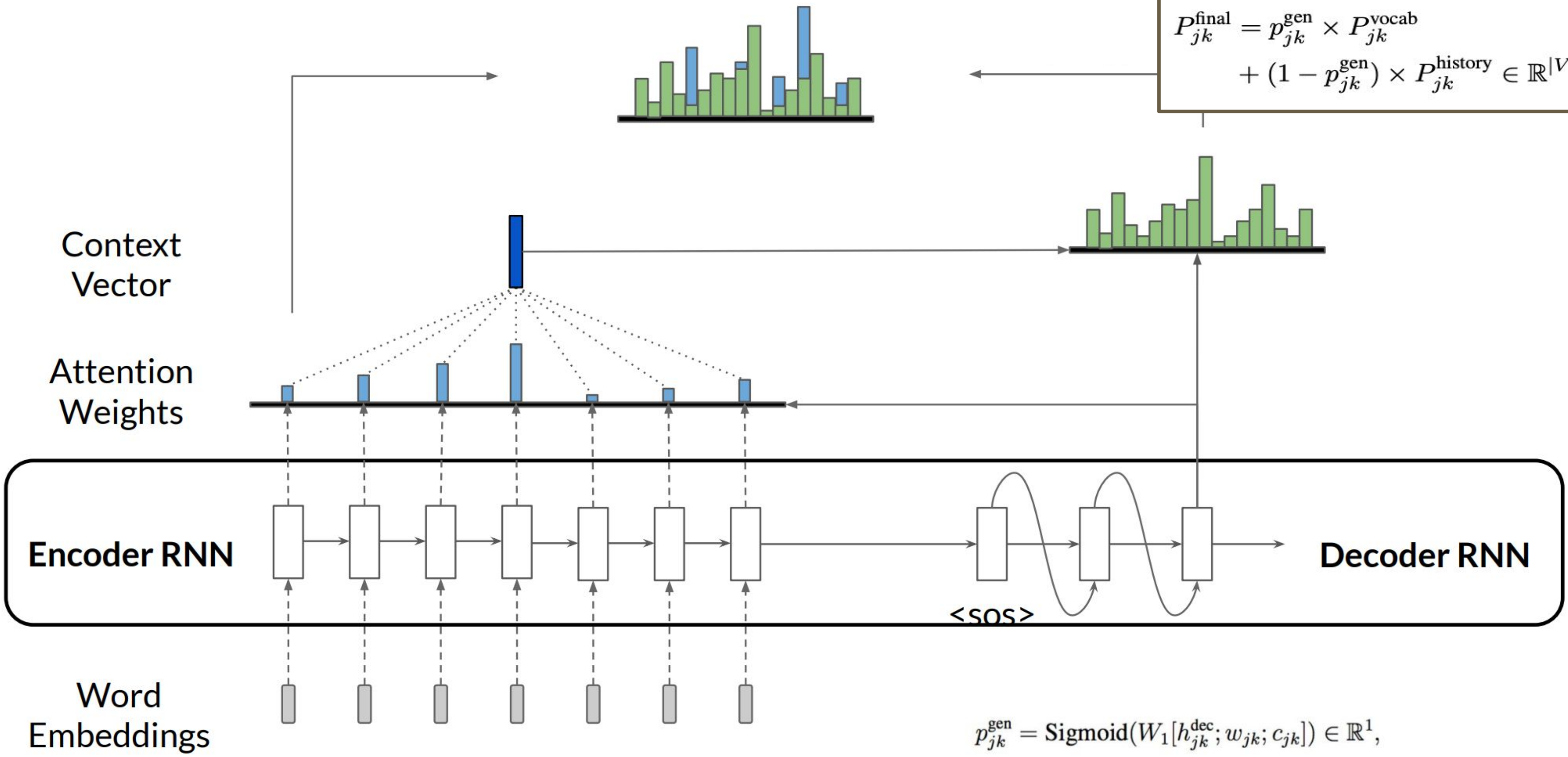
Seq2Seq with Attention

$$P_{jk}^{\text{history}} = \text{Softmax}(H_t(h_{jk}^{\text{dec}})^\top) \in \mathbb{R}^{|X_t|}.$$



Seq2Seq with Soft Copy Mechanism (See et al. 2017)

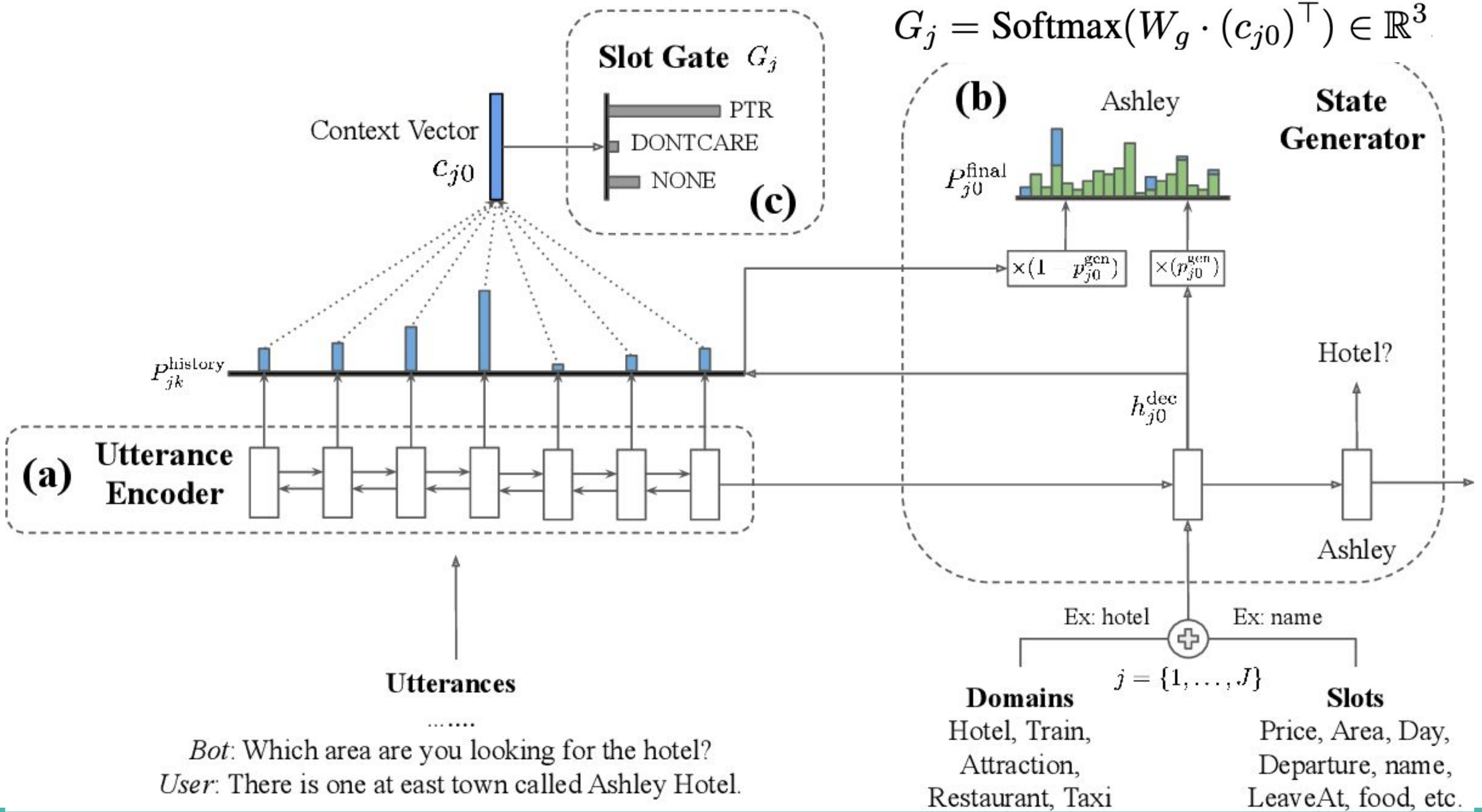
$$P^{\text{final}}_{jk} = p^{\text{gen}}_{jk} \times P^{\text{vocab}}_{jk} + (1 - p^{\text{gen}}_{jk}) \times P^{\text{history}}_{jk} \in \mathbb{R}^{|V|}.$$



$$p^{\text{gen}}_{jk} = \text{Sigmoid}(W_1[h^{\text{dec}}_{jk}; w_{jk}; c_{jk}]) \in \mathbb{R}^1,$$

Transferable Dialogue State Generator (TRADE)

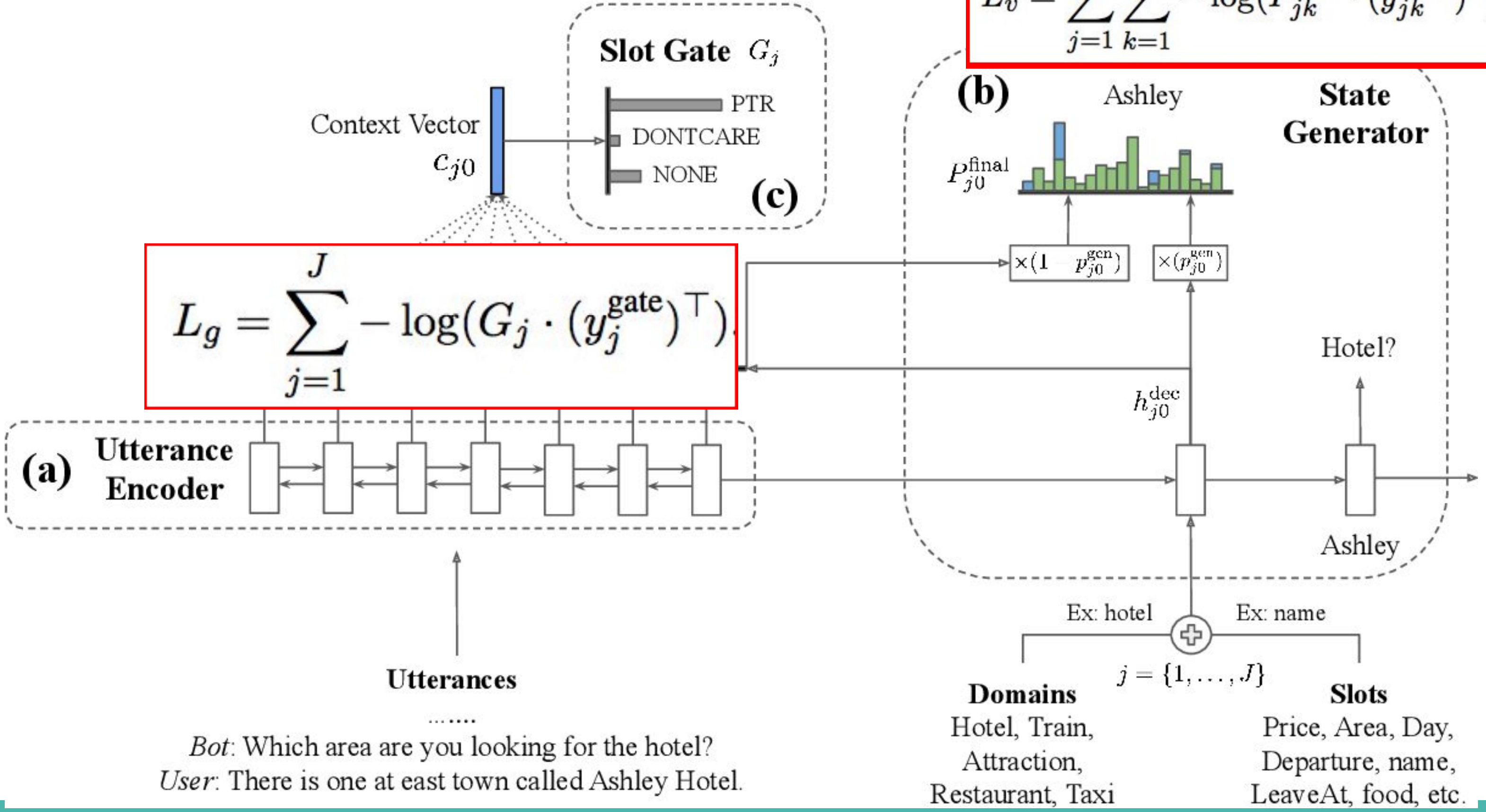
$$G_j = \text{Softmax}(W_g \cdot (c_{j0})^\top) \in \mathbb{R}^3$$



TRADE Optimization

$$L_v = \sum_{j=1}^{J} \sum_{k=1}^{|Y_j|} -\log(P_{jk}^{\text{final}} \cdot (y_{jk}^{\text{value}})^\top).$$

$$L_g = \sum_{j=1}^{J} -\log(G_j \cdot (y_j^{\text{gate}})^\top).$$



The dataset info of MultiWOZ - 30 (domain, slot) pairs

	Hotel	Train	Attraction	Restaurant	Taxi
Slots	price, type, parking, stay, day, people, area, stars, internet, name	destination, departure, day, arrive by, leave at, people	area, name, type	food, price, area, name, time, day, people	destination, departure, arrive by, leave by
Train	3381	3103	2717	3813	1654
Valid	416	484	401	438	207
Test	394	494	395	437	195

Multi-domain DST Evaluation metrics

- Joint goal accuracy
  - Compares the predicted dialogue states to the ground truth B_t at each dialogue turn t
  - Correct output iff all the predicted values exactly match B_t
- Slot accuracy
  - Individually compares each (domain, slot, value) triplet to its ground truth label

Results

TRADE - highest performance on joint goal accuracy

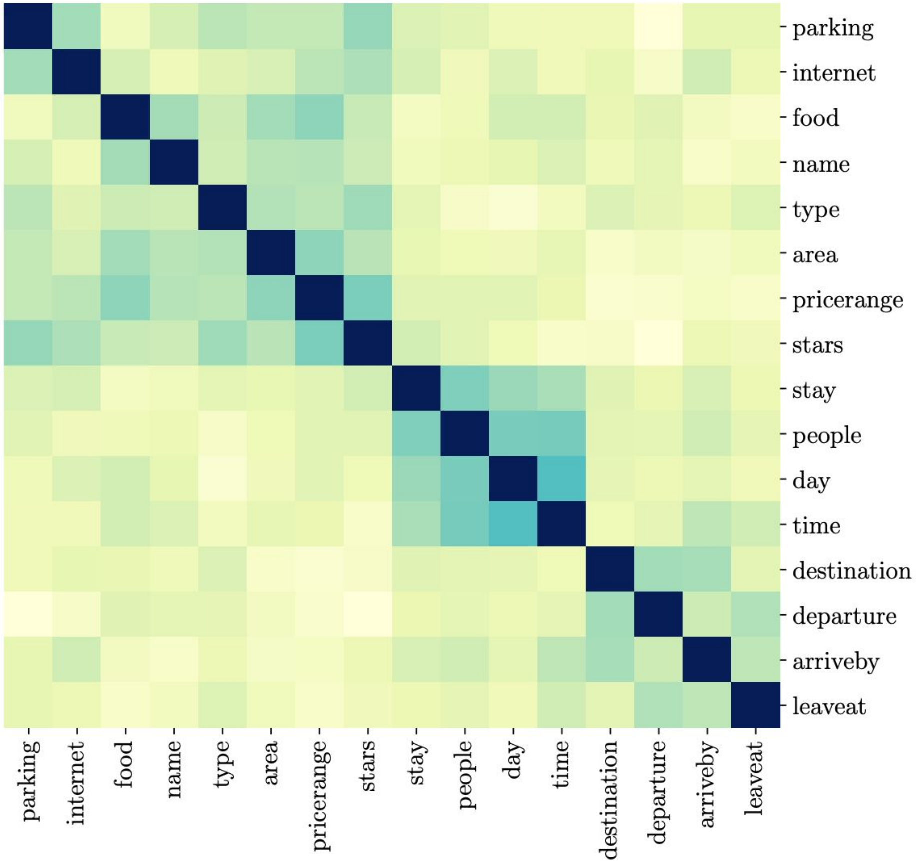
Potential limitations of other models:

- MDBT, GLAD, and GCE all need a predefined domain ontology
- SpanPtr uses index-based copying

	MultiWOZ		MultiWOZ (Only Restaurant)	
	Joint	*Slot*	*Joint*	*Slot*
MDBT	15.57	89.53	17.98	54.99
GLAD	35.57	95.44	53.23	$\boxed{96.54}$
GCE	36.27	$\boxed{98.42}$	60.93	95.85
SpanPtr	30.28	93.85	49.12	87.89
TRADE	**48.62**	96.92	**65.35**	93.28

Embeddings cosine similarity visualization

- The rows and columns are all the possible slots in MultiWOZ.
- Slots that share similar values or have correlated values learn similar embeddings.



Unseen Domain DST - zero shot

- Zero-shot setting:
  - No training data in the new domain
  - Generate target values given the context X, target domain D, and target slot S without using any training samples
    - [e.g., train - departure -> taxi -departure].
  - Extremely challenging if the target slot has never been trained.

Zero-shot experiments on an unseen domain

- Trained Single column is the results achieved by training on 100% single-domain data as a reference.
- Taxi domain reaches good performance >60%

	Trained Single		Zero-Shot	
	Joint	*Slot*	*Joint*	*Slot*
Hotel	55.52	92.66	13.70	65.32
Train	77.71	95.30	22.37	49.31
Attraction	71.64	88.97	19.87	55.53
Restaurant	65.35	93.28	11.52	53.43
Taxi	76.13	89.53	**60.58**	73.92

Unseen Domain DST - few shot

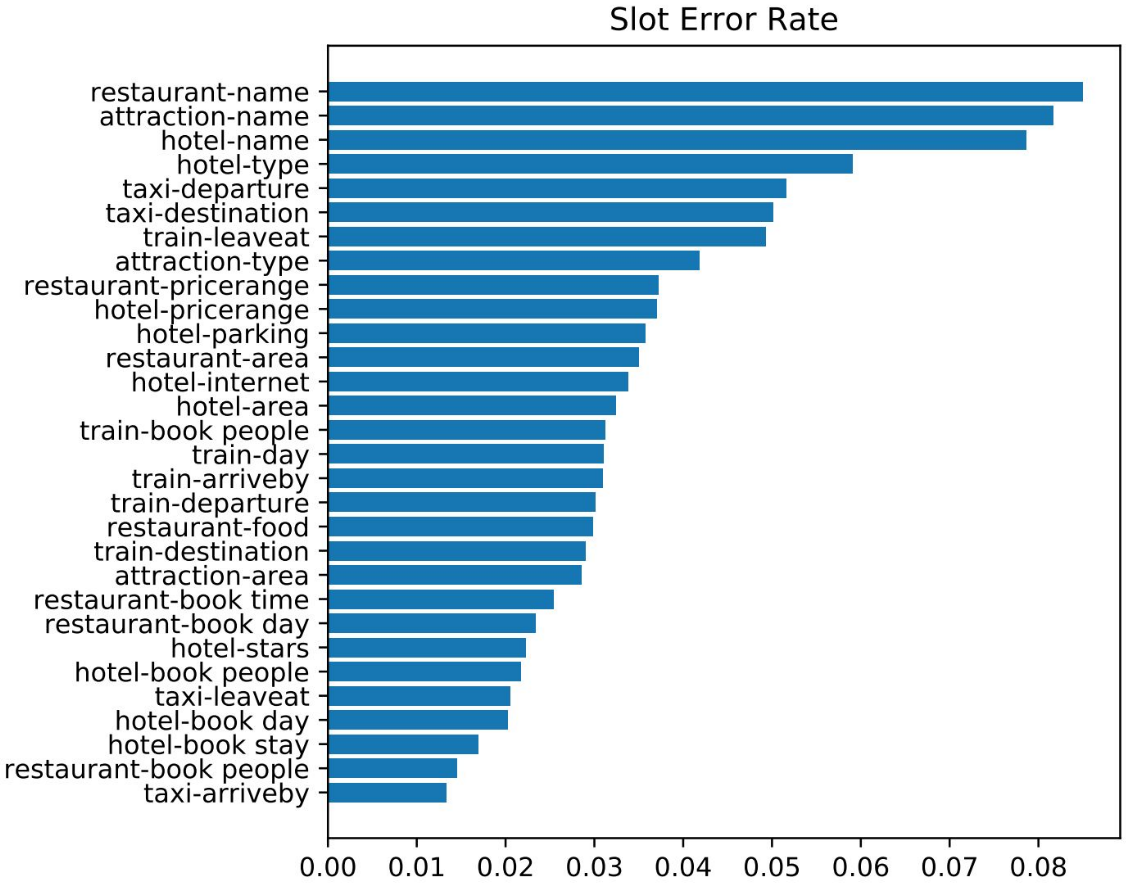
- Expanding DST for Few-shot setting:
  - 1% of the original training data in the unseen domain is available (around 20 to 30 dialogues)
  - Employ two continual learning techniques - elastic weight consolidation(EWC) and gradient episodic memory(GEM) to fine-tune the model.
  - EWC loss - $L_{ewc}(\Theta) = L(\Theta) + \sum_i \frac{\lambda}{2} F_i(\Theta_i - \Theta_{S,i})^2,$
  - GEM training process - $\text{Minimize}_{\Theta}\ L(\Theta)$

    $\text{Subject to } L(\Theta, K) \leq L(\Theta_S, K),$

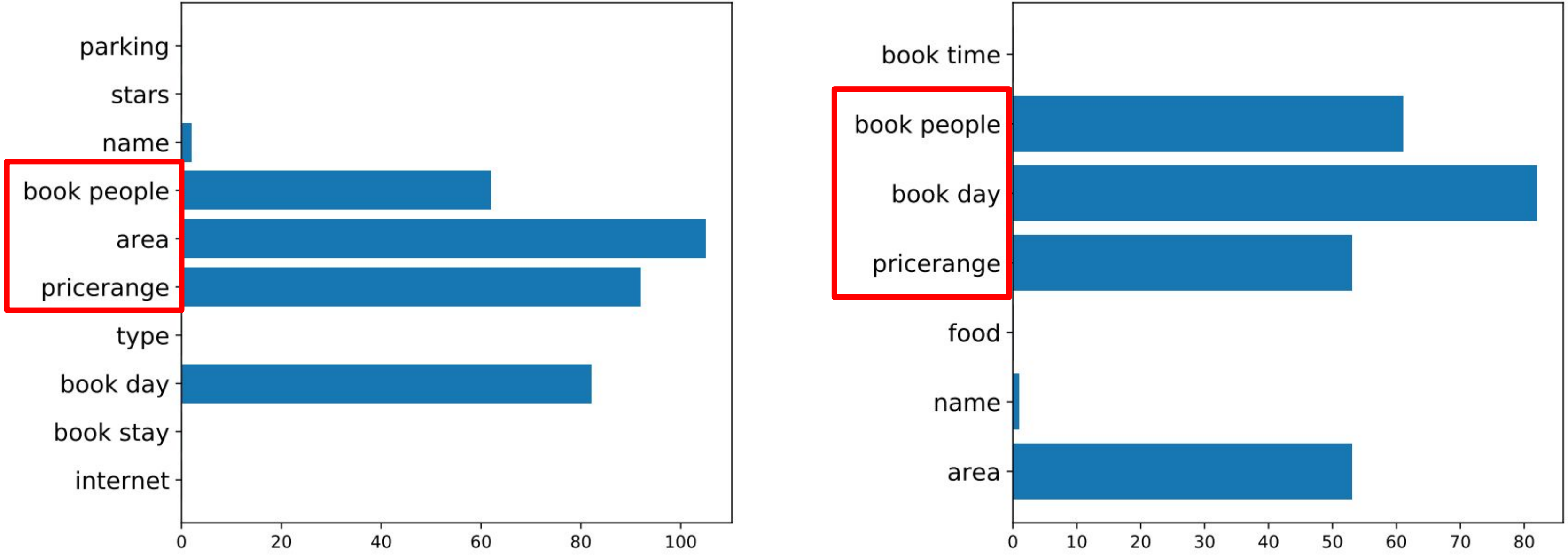
Domain expansion experiments by excluding one domain and fine-tuning on that domain

Evaluation on 4 Domains		**Joint** *Except Hotel*	**Slot**	**Joint** *Except Train*	**Slot**	**Joint** *Except Attraction*	**Slot**	**Joint** *Except Restaurant*	**Slot**	**Joint** *Except Taxi*	**Slot**
Base Model (BM) training on 4 domains		58.98	96.75	55.26	96.76	55.02	97.03	54.69	96.64	49.87	96.77
Fine-tuning BM on 1% new domain	*Naive*	36.08	93.48	23.25	90.32	40.05	95.54	32.85	91.69	46.10	96.34
	EWC	40.82	94.16	28.02	91.49	45.37	84.94	34.45	92.53	**46.88**	96.44
	GEM	**53.54**	**96.27**	**50.69**	**96.42**	**50.51**	**96.66**	**45.91**	**95.58**	46.43	**96.45**
Evaluation on New Domain		*Hotel*		*Train*		*Attraction*		*Restaurant*		*Taxi*	
Training 1% New Domain		19.53	77.33	44.24	85.66	**35.88**	**68.60**	32.72	82.39	60.38	72.82
Fine-tuning BM on 1% new domain	*Naive*	19.13	75.22	**59.83**	**90.63**	29.39	60.73	**42.42**	**86.82**	**63.81**	**79.81**
	EWC	19.35	76.25	58.10	90.33	32.28	62.43	40.93	85.80	63.61	79.65
	GEM	**19.73**	**77.92**	54.31	89.55	34.73	64.37	39.24	86.05	63.16	79.27

Error Analysis -Slots error rate



Zero-shot DST error analysis



(a) Hotel (b) Restaurant

TRADE Conclusion

- A copy-augmented generative model
- Can conduct multi-domain DST without ontology
- Enables zero-shot, and few-shot DST in an unseen domain with limited performance

Slide reference: **Chien-Sheng(Jason) Wu**
TRADE: Transferable Multi-Domain State Generator for Task-Oriented Dialogue Systems (ACL 2019)

Further challenges

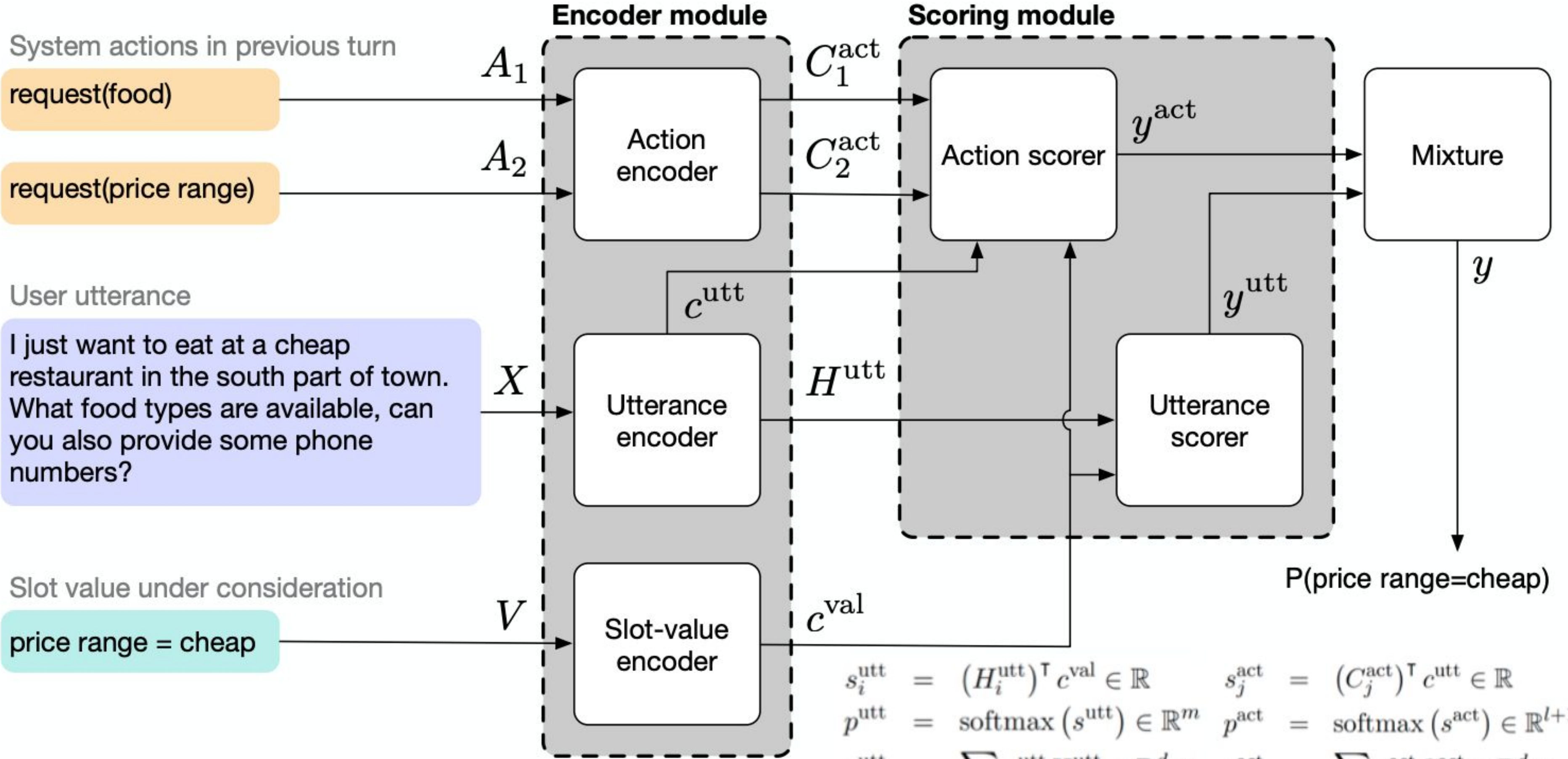
- The scale of the task-oriented corpora
- The noise and uncertainty in speech recognition
- The ambiguity when understanding human language
- The need to integrate third-party services and dialogue context in the decision-making
- The ability to generate natural and engaging responses

Q2: If we compare the dialogue state tracking models in these two papers, what is the biggest advance of the second paper (Wu et al, 2019)?

Q2: If we compare the dialogue state tracking models in these two papers, what is the biggest advance of the second paper (Wu et al, 2019)?

The biggest advance: predicts slot values directly, without pre-defining an ontology. The model is able to share parameters across different domains for multi-domain tasks.

Bonus paper - Global-Locally Self-Attentive Dialogue State Tracker (GLAD)



$$
\begin{aligned}
s_i^{\mathrm{utt}} &= \left(H_i^{\mathrm{utt}}\right)^{\intercal} c^{\mathrm{val}} \in \mathbb{R} & s_j^{\mathrm{act}} &= \left(C_j^{\mathrm{act}}\right)^{\intercal} c^{\mathrm{utt}} \in \mathbb{R} \\
p^{\mathrm{utt}} &= \mathrm{softmax}\left(s^{\mathrm{utt}}\right) \in \mathbb{R}^{m} & p^{\mathrm{act}} &= \mathrm{softmax}\left(s^{\mathrm{act}}\right) \in \mathbb{R}^{l+1} \\
q^{\mathrm{utt}} &= \sum_i p_i^{\mathrm{utt}} H_i^{\mathrm{utt}} \in \mathbb{R}^{d_{\mathrm{rnn}}} & q^{\mathrm{act}} &= \sum_j p_j^{\mathrm{act}} C_j^{\mathrm{act}} \in \mathbb{R}^{d_{\mathrm{rnn}}}
\end{aligned}
$$

$$
y = \sigma\left(y^{\mathrm{utt}} + w_y y^{\mathrm{act}}\right) \in \mathbb{R} \qquad y^{\mathrm{utt}} = W^{\mathrm{utt}} q^{\mathrm{utt}} + b^{\mathrm{utt}} \in \mathbb{R} \qquad y^{\mathrm{act}} = \left(q^{\mathrm{act}}\right)^{\intercal} c^{\mathrm{val}} \in \mathbb{R}
$$

References:

Bobrow, Daniel G., et al. "GUS, a frame-driven dialog system." *Artificial intelligence* 8.2 (1977): 155-173

Budzianowski, Paweł, et al. "Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling." *arXiv preprint arXiv:1810.00278* (2018).

Deriu, Jan, et al. "Survey on evaluation methods for dialogue systems." *arXiv preprint arXiv:1905.04071* (2019).

Schmitt, Alexander, and Stefan Ultes. "Interaction quality: assessing the quality of ongoing spoken dialog interaction by experts—and how it relates to user satisfaction." Speech Communication 74 (2015): 12-36.

Walker, Marilyn, Candace Kamm, and Diane Litman. "Towards developing general models of usability with PARADISE." Natural Language Engineering 6.3-4 (2000): 363-377.

Wen, Tsung-Hsien, et al. "A network-based end-to-end trainable task-oriented dialogue system." *arXiv preprint arXiv:1604.04562* (2016).

Wu, Chien-Sheng, et al. "Transferable multi-domain state generator for task-oriented dialogue systems." *arXiv preprint arXiv:1905.08743* (2019).

Young, Steve, et al. "The hidden information state model: A practical framework for POMDP-based spoken dialogue management." *Computer Speech & Language* 24.2 (2010): 150-174..