General Dialogue Topics

NAIST
Assistant Professor
Koichiro Yoshino

Nara Institute of Science and Technology
Augmented Human Communication Laboratory
PRESTO, Japan Science and Technology Agency

Conventional dialogue systems

I'd like to take Kintetsu-line from Ikoma

$FROM=Ikoma
$LINE=Kintetsu

Language understanding

$FROM=Ikoma
$TO_GO=???
$LINE=Kintetsu

Knowledge base, model

Dialogue management

Language generation

Where do you want to go?

1 ask $TO_GO
2 inform $NEXT
…

- **The user utterance is understood as an actual dialogue state**
  - By considering history
- **The system action is decided according to the dialogue state and its confidence**
  - Generate an utterance according to the decided system action

What NCM does

I'd like to take Kintetsu-line from Ikoma

$FROM=Ikoma
$LINE=Kintetsu

Language understanding

Knowledge base, model

$FROM=Ikoma
$TO_GO=???
$LINE=Kintetsu

Dialogue management

1 ask $TO_GO
2 inform $NEXT
…

Give-up understanding and managing

Language generation

Where do you want to go?

Speech recognition with DNN in early stage

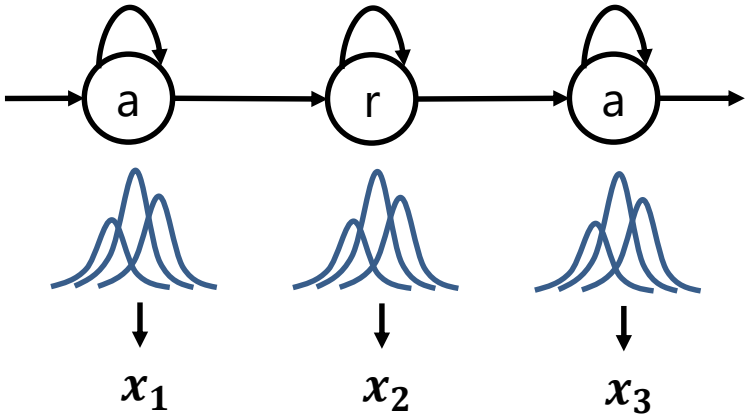
- **Conventional ASR architecture**

$$\underset{W}{\operatorname{argmax}}\, P(W|X) = \underset{W}{\operatorname{argmax}}\, P(X|W)P(W)$$

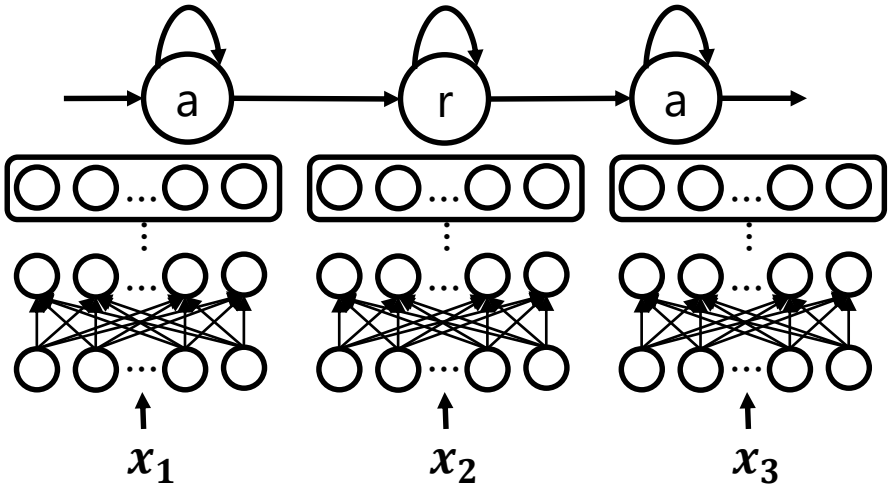
Acoustic model　Language model

W **is word sequence and** X **is speech**

GMM-HMM



DNN-HMM



What is output of the ASR?

- **N-best hypotheses of speech recognition results**
  - with posterior probabilities, which is calculated from likelihoods of acoustic model and language model

0.7	I want to take a flight to Austin
0.2	I want to take a flight to Boston
0.05	I want to take applied to Austin
	…

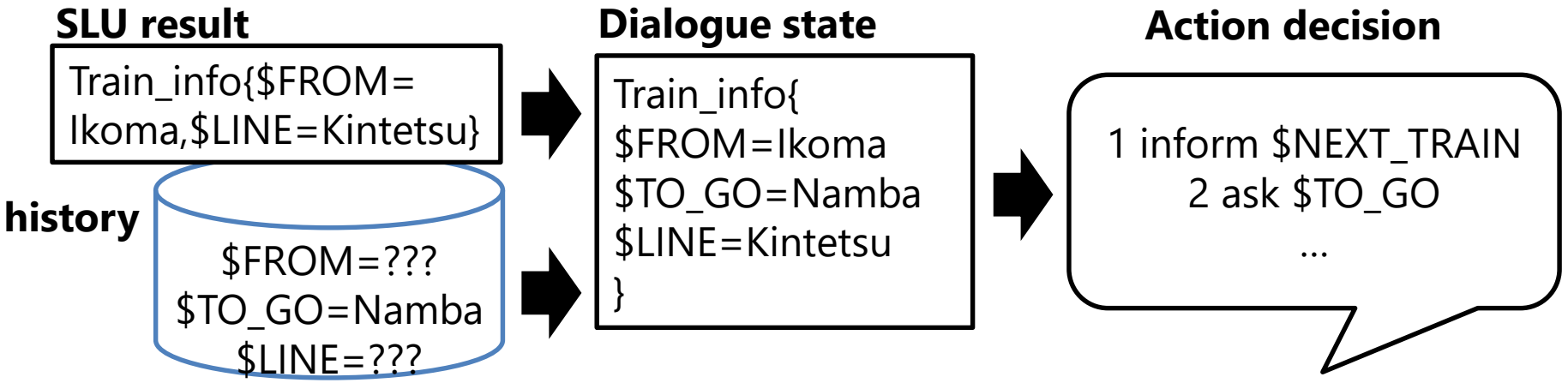
- **ASR results often contain errors**
  - Insertion, deletion, replacement, …

- **We have to consider the error in post processes (SLU, DM, …)**

Spoken language understanding (SLU) and dialogue management (DM)

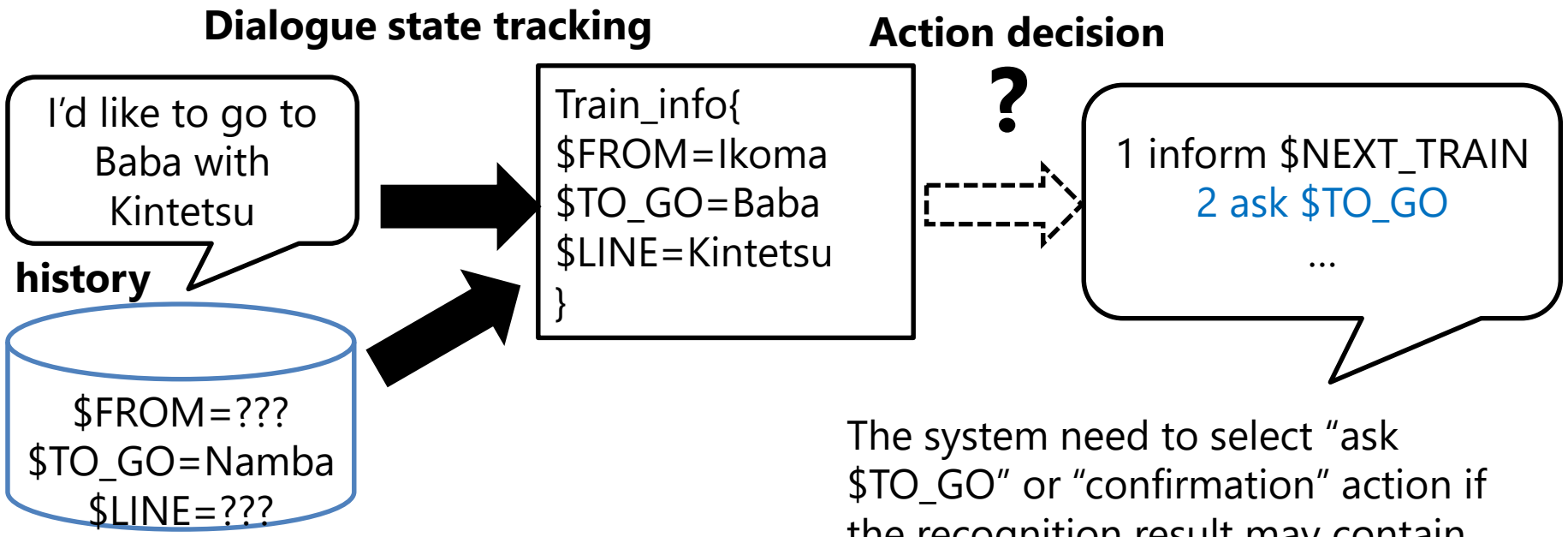
- **Language understanding**
  - Convert the user utterance into machine-readable expressions

- **Dialogue management**
  - Decide the next system action from the SLU result and dialogue history



Dialogue state tracking and action decision

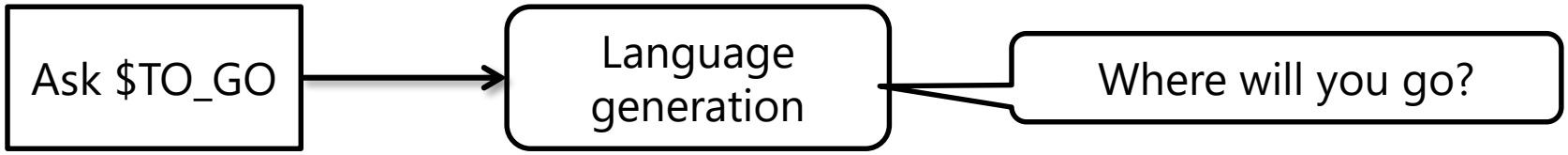
- **As mentioned before, ASR results often contain errors**
  - SLU results are probably affected by the ASR error
  - SLU module also causes error



The system need to select "ask $TO_GO" or "confirmation" action if the recognition result may contain critical errors

Language generation systems

- **Generate a sentence given a system action**



- **Difficulty of generation**
  - **Appropriateness**: Outputs contain the contents that is decided by the dialogue manager
  - **Naturalness**: Outputs is natural
  - **Understandability**: Outputs should be easy to understand
  - **Variation**: Outputs contain some variations of expression

Problems in existing systems

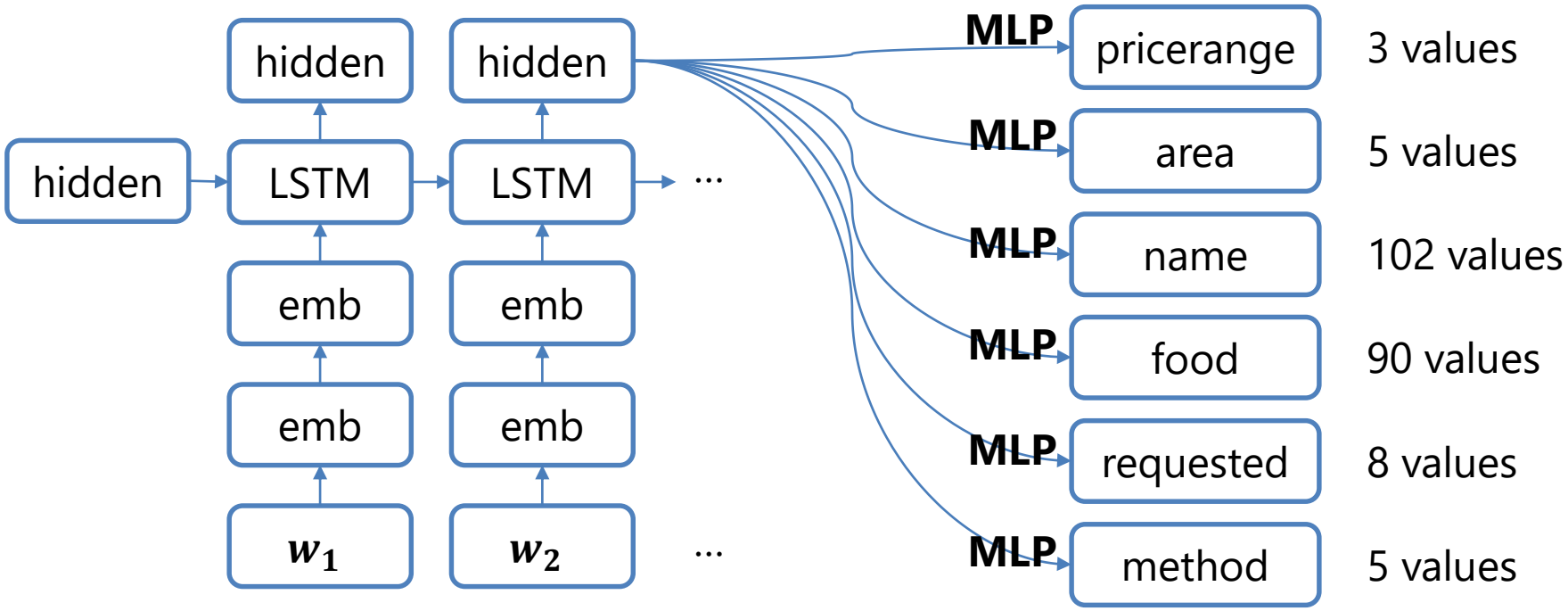
- **Turn-taking is not natural**
  - Based on voice activity detection (VAD)
- **Need to define ontology**
  - Handcrafting for any new domains
- **Dialogue strategy in a new space**
  - RL-based optimization can be used if we define states and actions
- **Controllable neural language generation**
  - Only using cross-entropy loss

Our approaches

- **Turn-taking is not natural**
  - Based on understanding results of the system
- **Need to define ontology**
  - Design of language understanding space
- **Dialogue strategy in a new space**
  - Information seeking for argumentation dialogue
- **Controllable neural language generation**
  - Use seqGAN and label aware objective

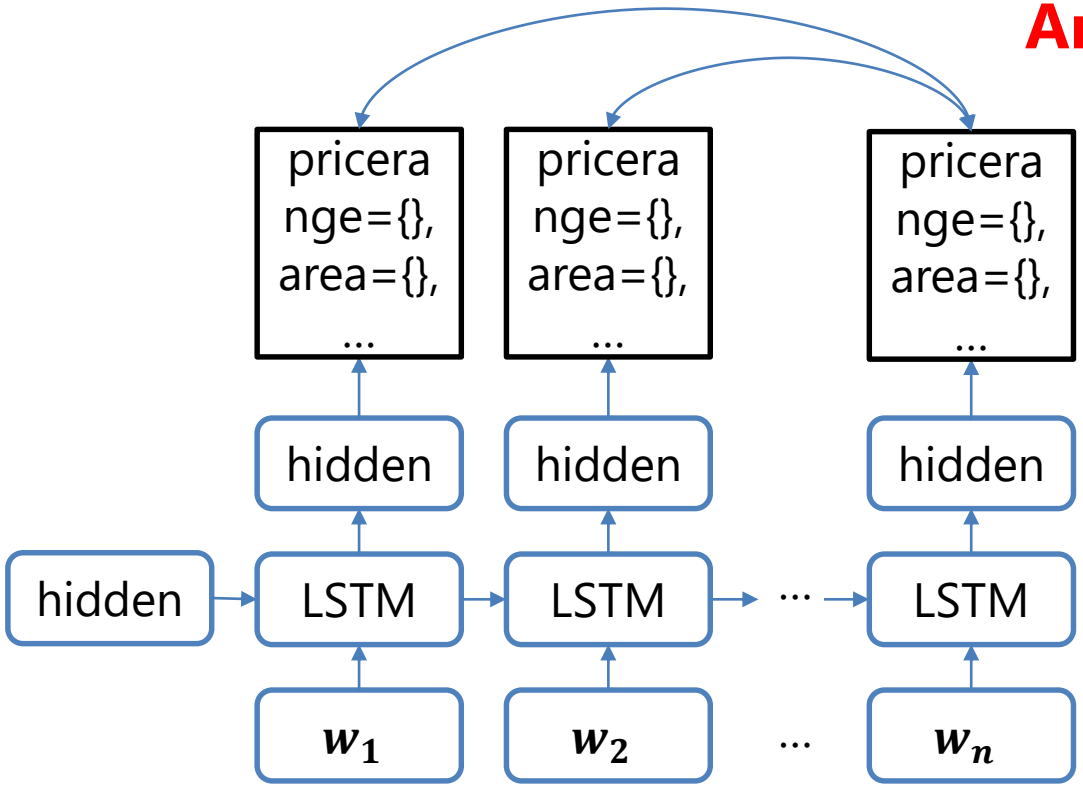
Incremental understanding system

- **Incremental system that receives a word on each time-step**
  - Multi-layer perceptron classifiers given the hidden layer of LSTM
  - Cambridge restaurant navigation system (DSTC2)



Re-labeling

- **Make a training data of turn taking by comparing DST results on any time steps with the last result**

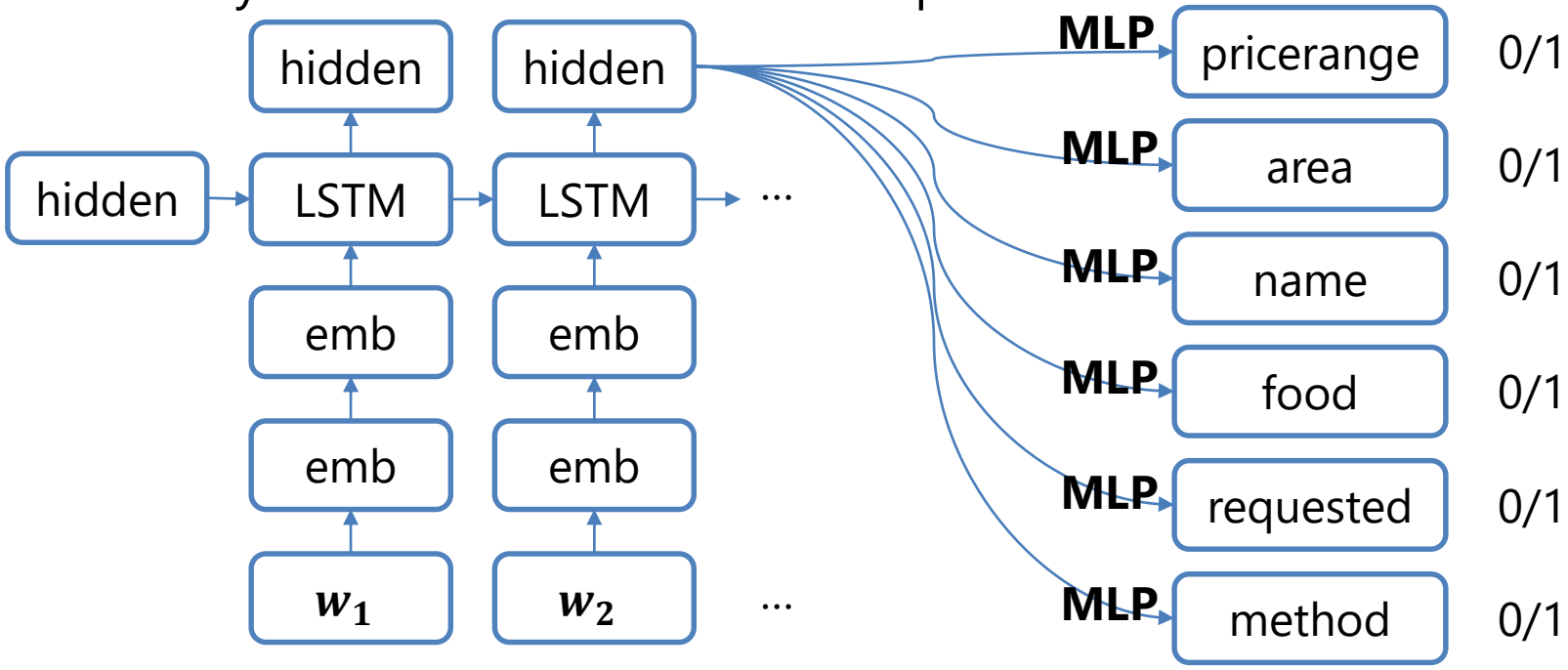


Any differences?

- Yes → the system still need to wait future words
- No → the system can take a turn at this moment!

Incremental turn taking decider

- **Comparing NLU results on between current input and the point the utterance ends**
  - No difference: 0 / Different: 1 → supervised learning
  - System can take a turn if "0" is predicted

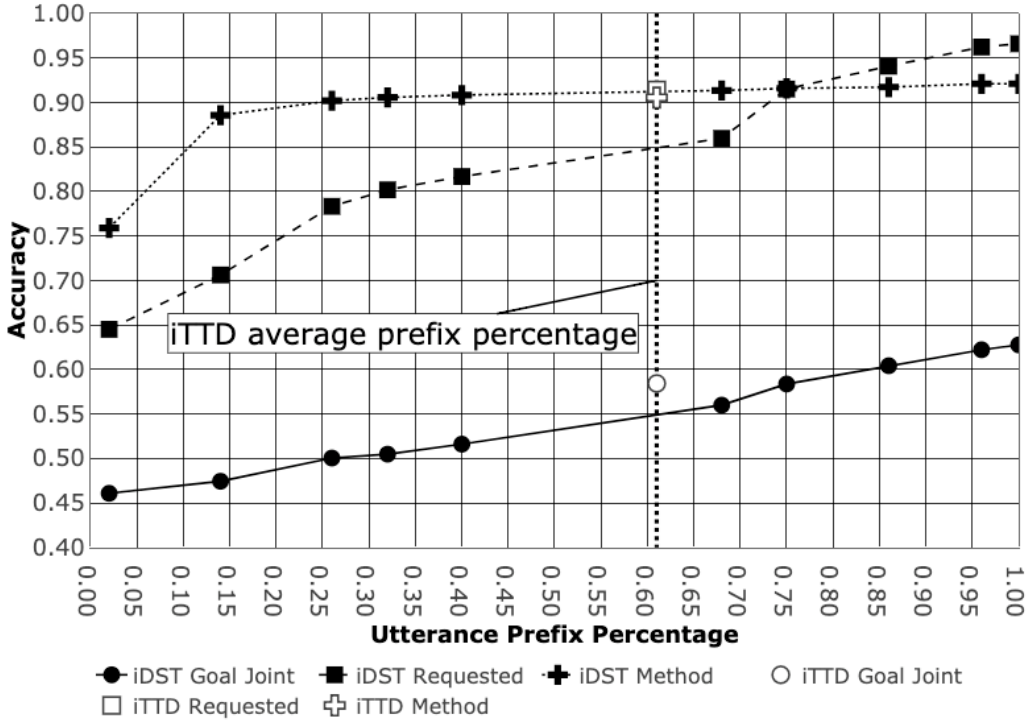


Results on DSTC2 dataset

Model	Dev Goal Acc.	Dev Goal L2	Dev Method Acc.	Dev Method L2	Dev Requested Acc.	Dev Requested L2	Test Goal Acc.	Test Goal L2	Test Method Acc.	Test Method L2	Test Requested Acc.	Test Requested L2
LecTrack [9]	0.63	0.74	0.90	0.19	0.96	0.08	0.62	0.75	0.92	0.15	0.96	0.07
iDST_ASR(r = 1.0)	0.64	0.53	0.90	0.17	0.96	0.07	0.63	0.56	0.92	0.13	0.97	0.06
iDST_TRA(r = 1.0)	0.87	0.23	0.94	0.10	0.99	0.02	0.82	0.30	0.94	0.09	0.99	0.02
iDST_ASR(r = 0.6)	0.57	0.61	**0.89**	0.18	0.86	0.23	0.56	0.62	0.91	0.14	0.86	0.21
iTTD_ASR(d = 0.85)	**0.59**	0.60	0.88	0.19	**0.91**	0.16	**0.58**	0.61	0.91	0.15	0.91	0.15
iDST_TRA(r = 0.6)	0.77	0.34	**0.93**	0.11	0.88	0.18	0.73	0.39	**0.94**	0.10	0.88	0.18
iTTD_TRA(d = 0.85)	**0.80**	0.31	0.92	0.12	**0.91**	0.15	**0.76**	0.37	0.93	0.11	**0.91**	0.15

- **DST accuracy itself was improved by the incremental process**
  - 80-97% for each slots, if we use transcription
- **r=0.6: the system interrupt at 60% utterance**
- **d=0.85: the system interrupt if iTTD conf. is bigger than 0.85**
- **Comparable scores to waiting any words by users**

Analysis



- **Adaptive turn taking can manage both NLU accuracy and interrupting**

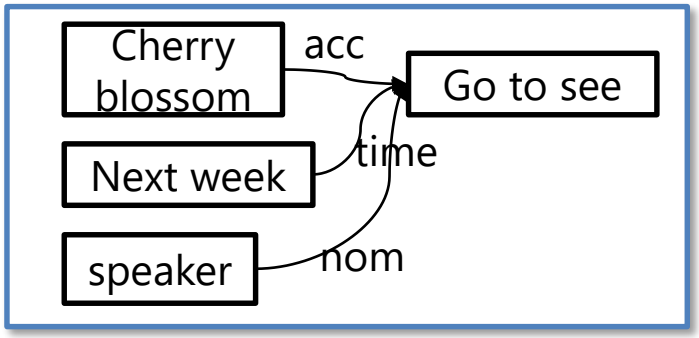
On-going project: Language understanding based on events

- **Ontology-based NLU space requires handcrafting to define**
  - Each domain requires own ontology
  - Generation also requires handcrafting

Frame (slot-value)

Act: Request
Type: Chinese restaurant
Price_range: don't care
Count: 2
Kids_allowed: NULL

Event (predicate-argument)



- **Idea: using event (P-A) as understanding space**
  - Will work for any domain that parsers can work
  - Coverage is limited
    - Difference between "go to see" and "visit to see"

New dialogue domain: argumentation

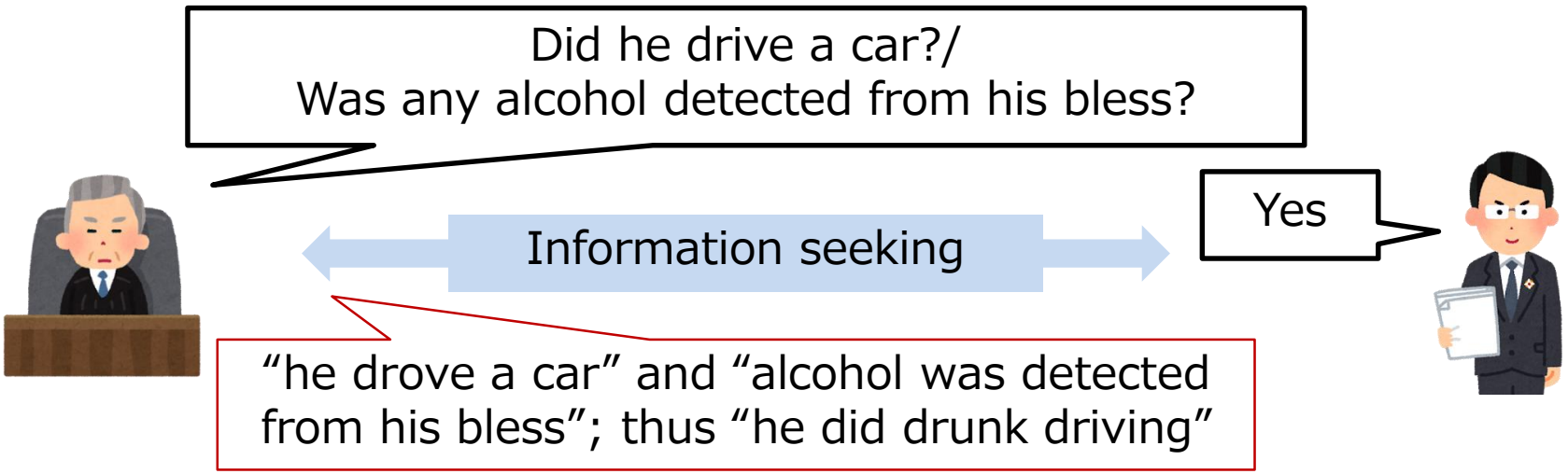
- **Argumentation**

"he drove a car" and "alcohol was detected from his breath"; thus "he did drunk driving"

- **Claim:**
  - he did drunk driving
- **Supportive facts:**
  - he drove a car
  - alcohol was detected from his breath

Information seeking for argumentation

- **Collecting supportive facts through a dialogue**



Collecting rational arguments

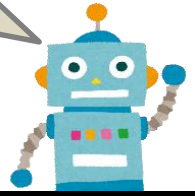
Presented at AAAI2019 DeepDIAL-WS

- **A lot of possible questions**
  - Policy is trained to decide
    “which action will the system ask on which situation”

Information seeking based on Markov decision process

He did drunk driving, because alcohol was detected from his breath

claim



system

- Alcohol was detected from his breath

Collected facts

Reitionality 0.1

Calculated by weighted abduction [Ovchinnikova+, 14]

- **He drove a car**

What he know

answerer

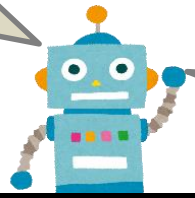
Information seeking based on Markov decision process

He did drunk driving, because alcohol was detected from his breath

claim

Reitionality 0.1

Did he drove a car?



system

- Alcohol was detected from his breath

Collected facts

Yes

- **He drove a car**

What he know

answerer

Information seeking based on Markov decision process

He did drunk driving, because alcohol was detected from his breath

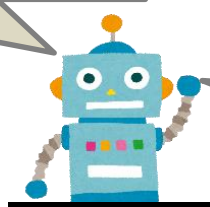
claim

Reitionality 0.1

Did he drove a car?

- Alcohol was detected from his breath

Collected facts



system

Yes

- **He drove a car**

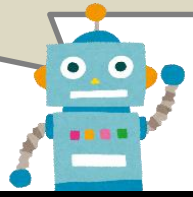
What he know

answerer

He did drunk driving, because alcohol was detected from his breath, and he drove a car

Raitionality **1.0**

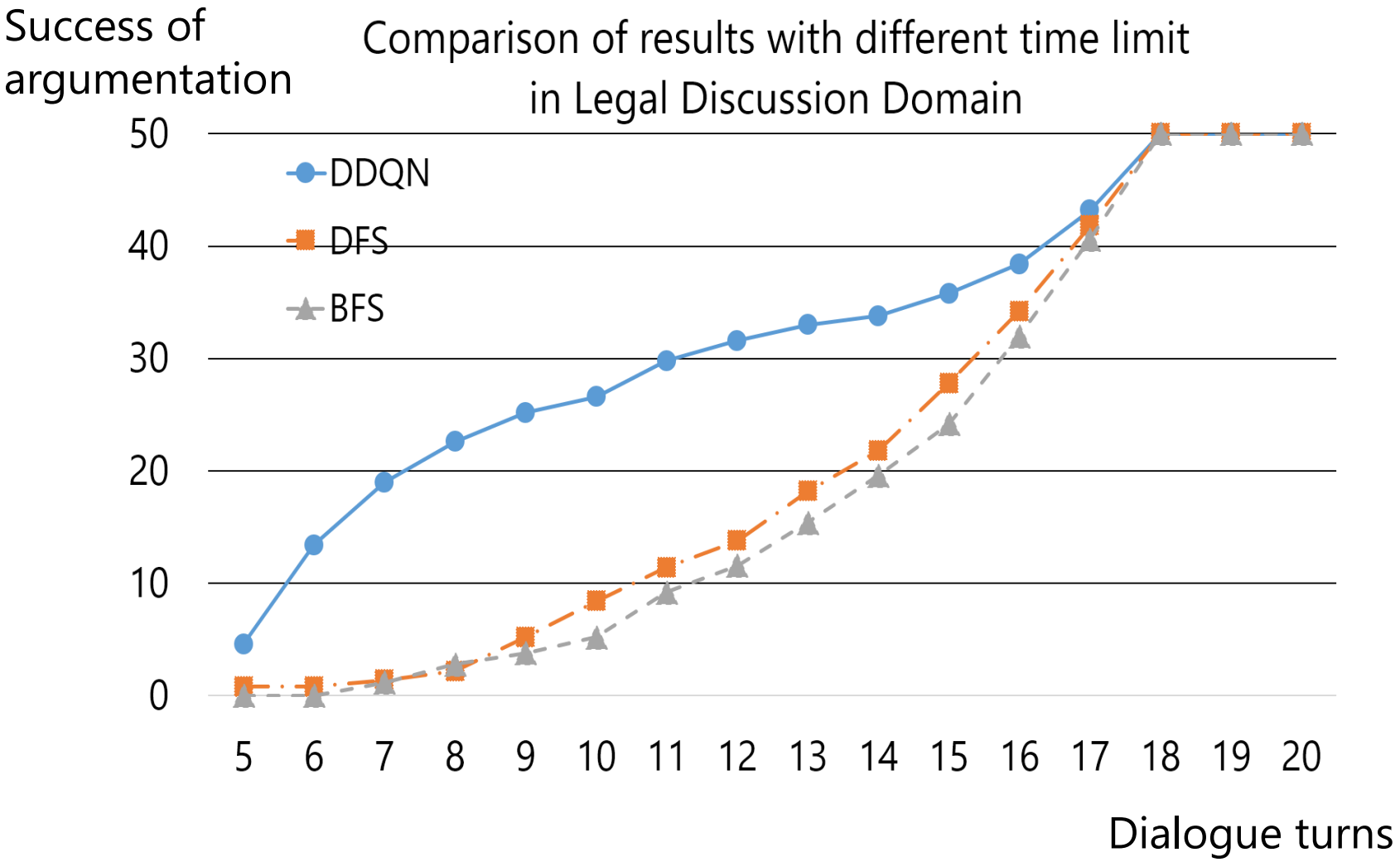
- Alcohol was detected from his breath
- **He drove a car**



system

reward: $r(s_t, a_t, s_{t+1})$

Strategy optimized by deep reinforcement learning



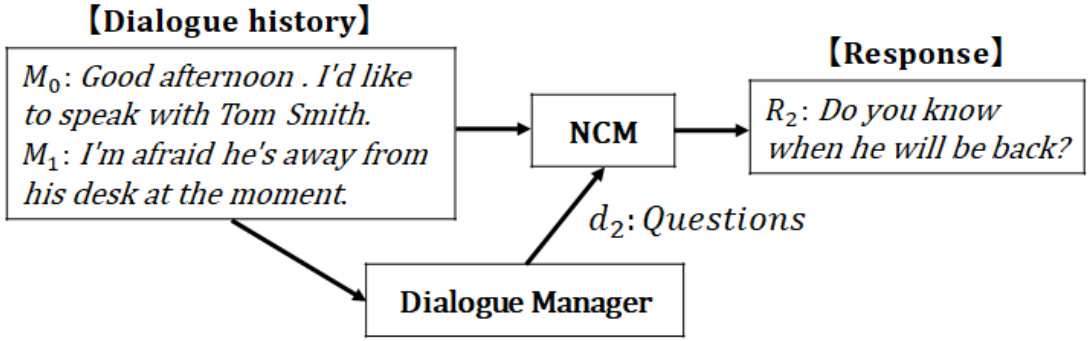
Dialogue example

- **Dialogue on twenty angry men dataset**
  - System tries to construct a rational argument to claim “the boy is not guilty”
  - System can collect facts from the dialogue partner

Step	Speaker	Utterance	Rationality
1	Sys	Did not the woman across the street see the murder committed?	
	Ans	**The woman across the street did not see the murder committed.**	0.1
2	Sys	Did not the old man hear the kid yell, “I’m gonna kill you”?	
	Ans	**The old man did not hear the kid yell, “I’m gonna kill you”.**	0.4
3	Sys	Is the old man a liar?	
	Ans	I do not know.	0.4
4	Sys	Did not the boy crouch down and stab a taller man in the chest?	
	Ans	I do not know.	0.4
5	Sys	Did not the woman across the street see the boy stab his father?	
	Ans	I do not know.	0.4
6	Sys	Did not the boy admit buying the knife?	
	Ans	**The boy did not admit buying the knife.**	0.7

Language generation for dialogue systems

- **Contents to be contained in the generation results are decided by dialogue manager**

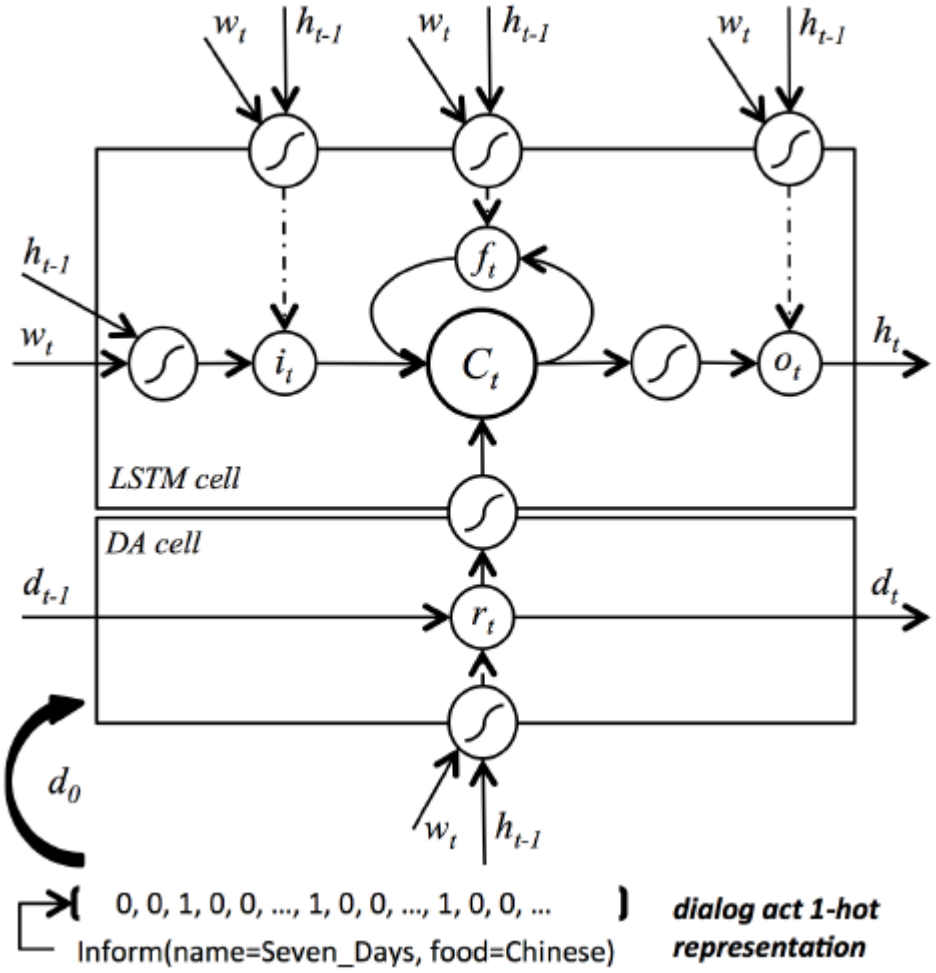


- **There are some works to generate sentences given an action**
  - Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. Wen et al., In Proc. EMNLP, 2015.
  - Dusek et al., A context-aware natural language generator for dialogue systems. In Proc. SIGDIAL 2016.

SC-LSTM by Wen et al., 2015

recurrent hidden layer

embedding of a word

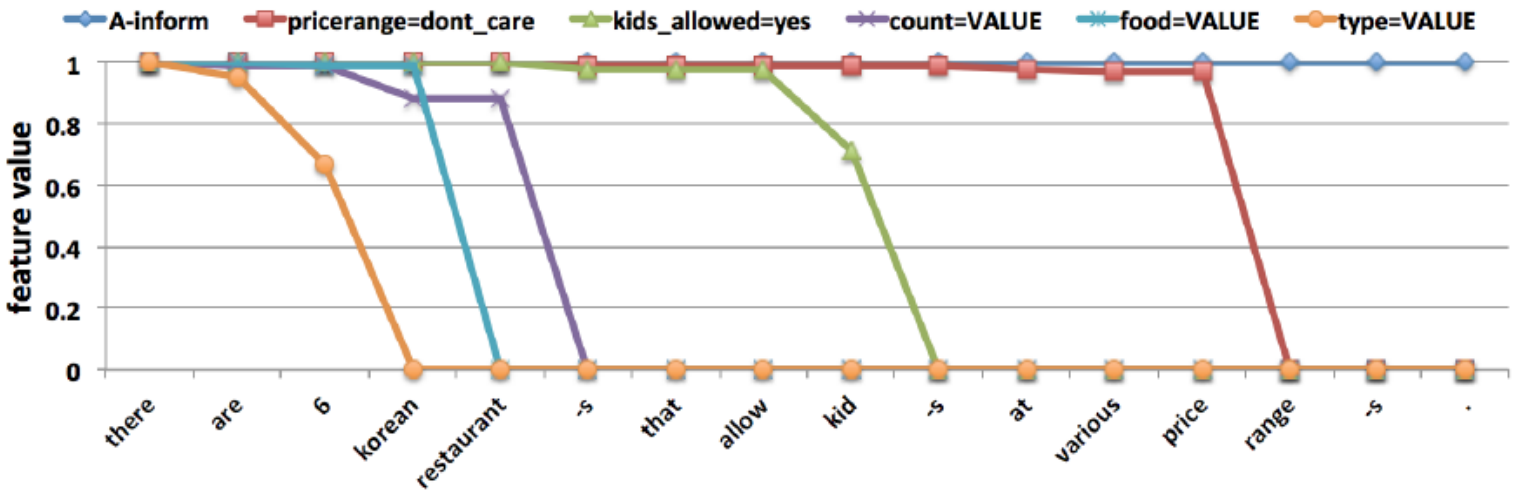


1-hot dialog act and slot values

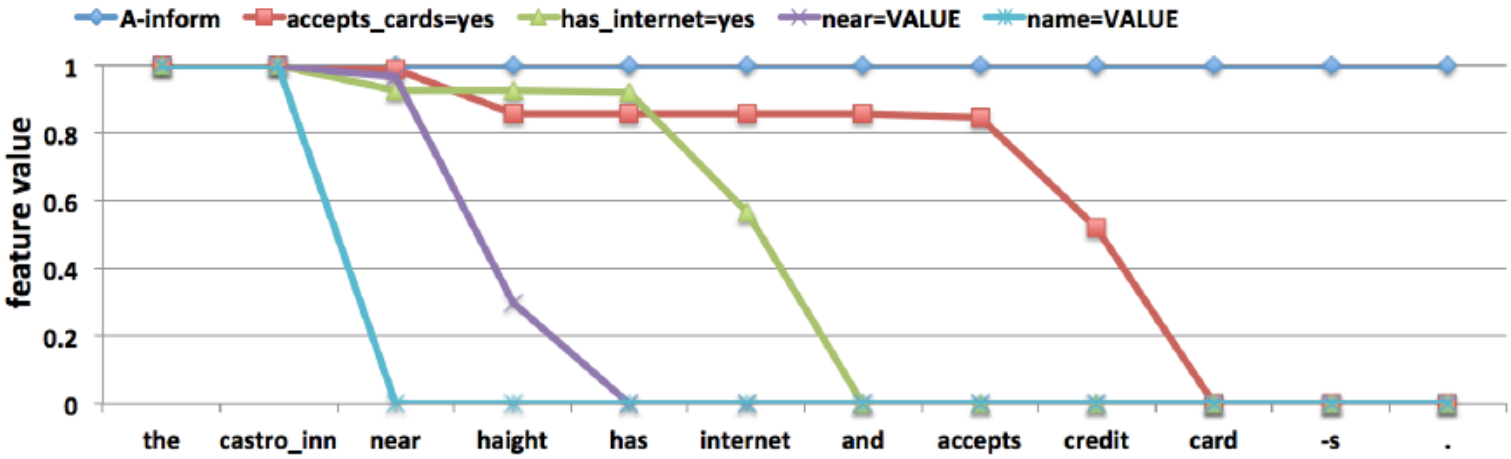
How to say (=LM)

What to say (contents)

Sample generations by SC-LSTM



(a) An example realisation from SF restaurant domain

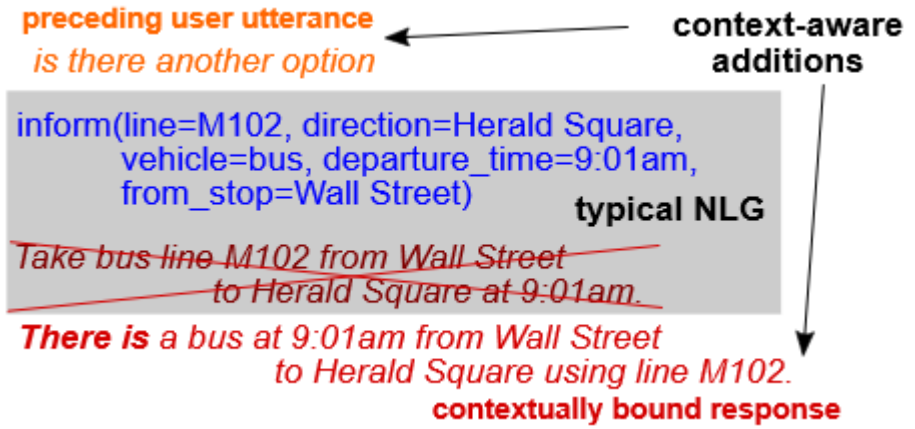


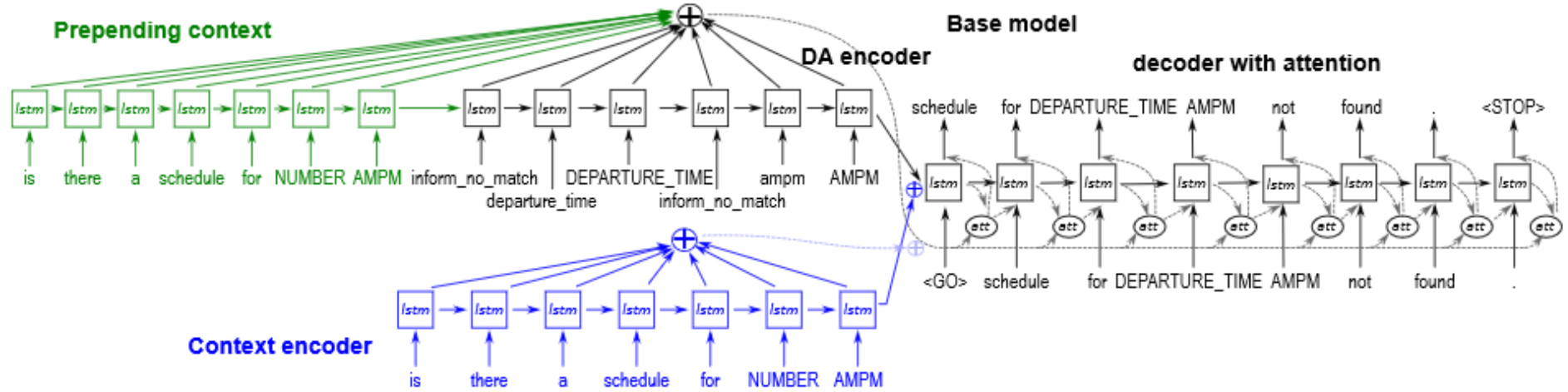
(b) An example realisation from SF hotel domain

Figure 3: Examples showing how the SC-LSTM controls the DA features flowing into the network via

Context-aware NLG

- **Sequence-to-sequence modeling of generation**
  - Change the response according to the dialogue context



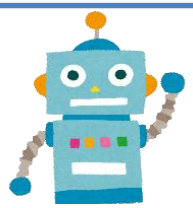


What the problem?

- **Existing generation systems are trained by softmax-cross entropy-loss to words**
  - No guarantee to contain given information by system action

Input

Candidates: 2
Area: Düsseldorf
Pets: allow



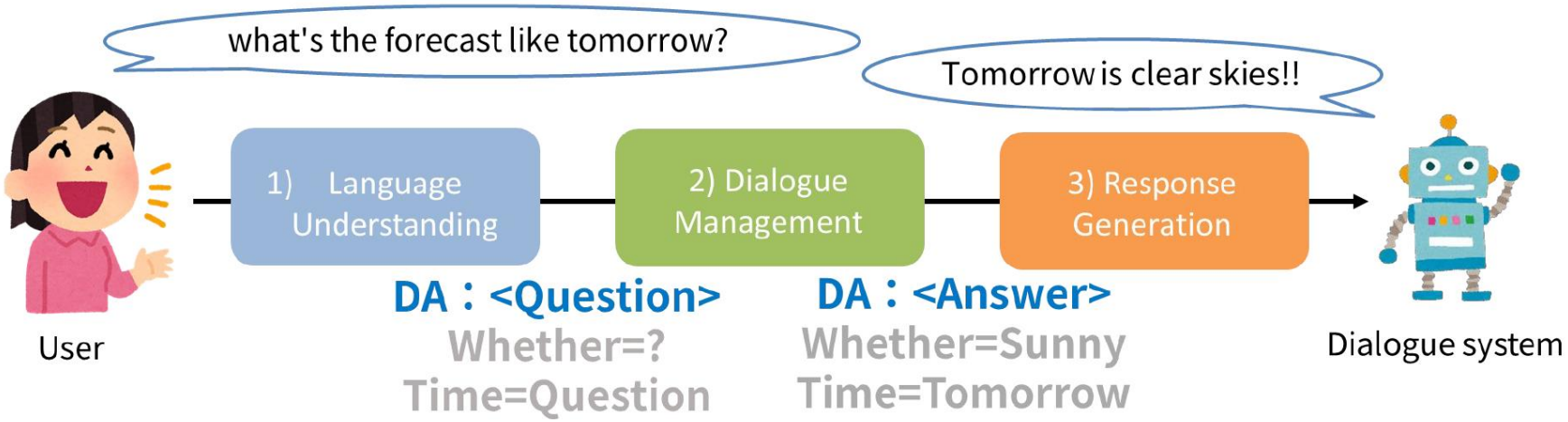
Generation

There are 2 hotels that allow pets ?

Training data

There are 2 hotels in Düsseldorf that allows pets

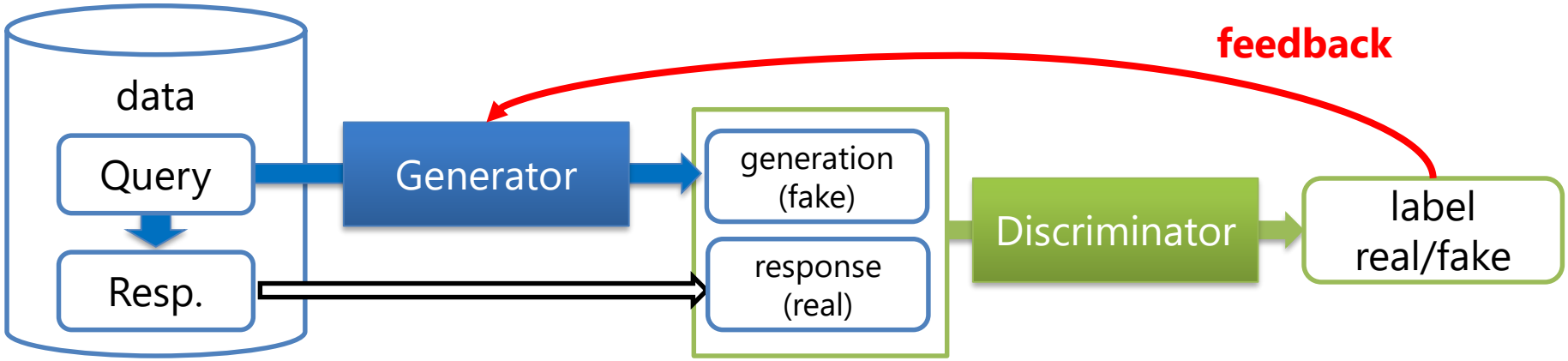
Controlling generation results with condition



- **We built a generation system based on**
  - generative adversarial network (Seq-GAN) and
  - label-aware objective
- **We only controlled by dialogue acts of the system**
  - The system itself is NCM

Generation based on SeqGAN

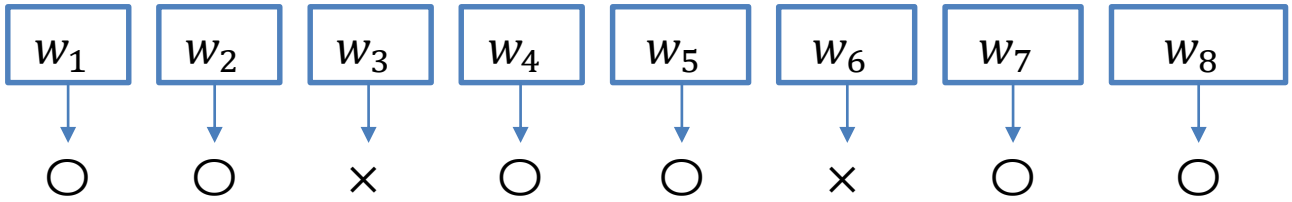
- **Sequential generative adversarial network is a technique to evaluate whole of sentence (not word-by-word)**



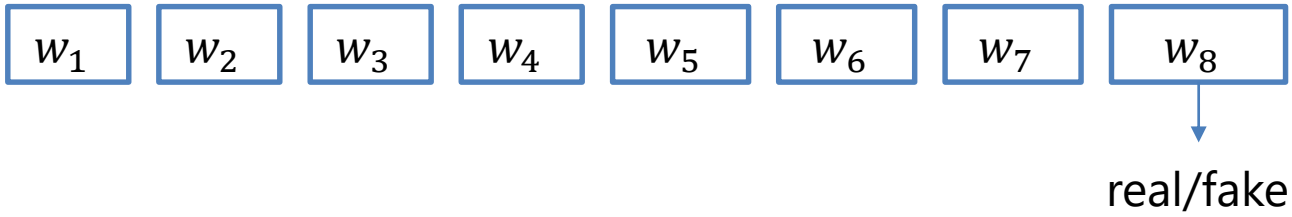
- **Discriminator predicts two classes (real/fake)**
- **Generated receives reward to the generation sequence**
  - Reinforcement learning is used (n-step delayed reward)

Cross-entropy loss and SeqGAN

- **Generation model based on softmax-cross entropy calculates loss for each word**

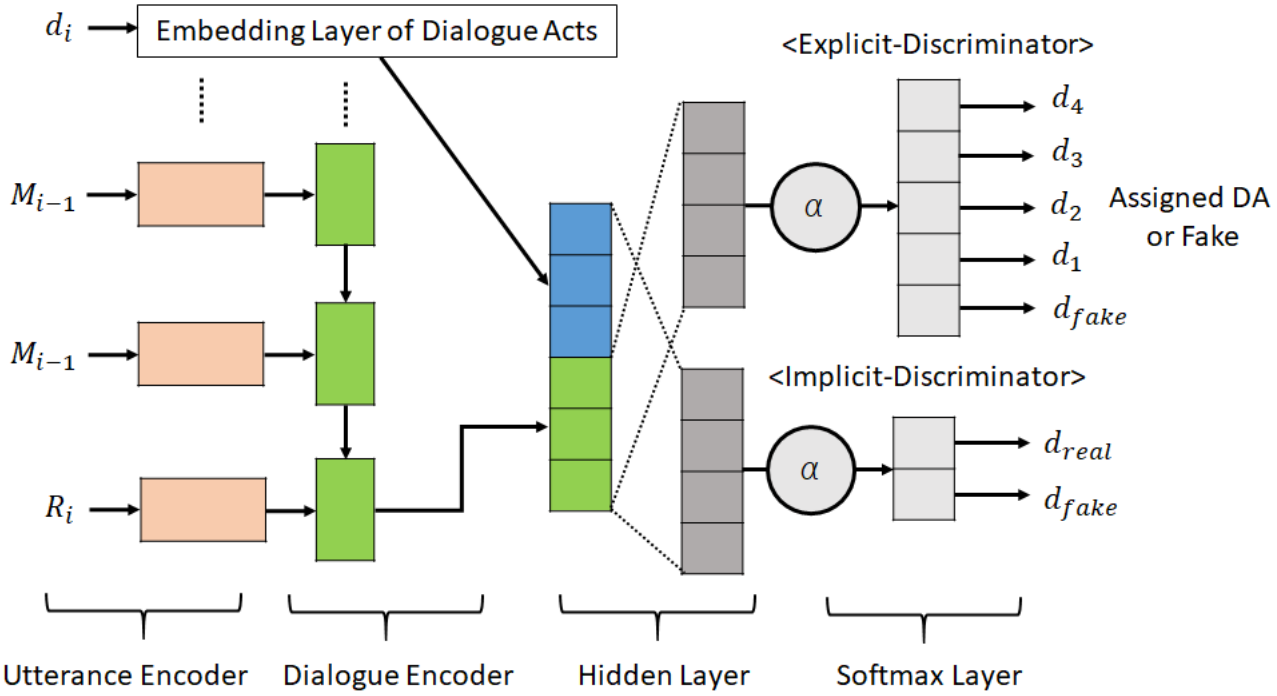


- **SeqGAN only calculate feedback at the end of sequence**



Label aware objective

- **SeqGAN only distinguish real or fake**
  - We extended the discriminator to multi-class classification to know the generation result is based on input or not



Naturalness and controllability (human)

Naturalness	Natural	Not natural
NCM-w/condition	0.49	0.51
Adversarial-Implicit	0.57	0.43
Adversarial-Explicit	0.58	0.42

Controllability	Acc.	F-1
NCM-w/condition	0.743	0.759
Adversarial-Implicit	0.706	0.681
Adversarial-Explicit	**0.797**	**0.787**

- **Both naturalness and controllability were improved**
  - Explicit penalty to the condition improved the controllability
  - Adversarial learning improved naturalness

Summary

- **We introduced basic architecture of spoken dialogue systems, and tackled several problems of existing systems**
  - Turn-taking is not natural
  - Need to define ontology
  - Dialogue strategy in a new space
  - Controllable neural language generation