Spoken Language Processing Applications

Koichiro Yoshino

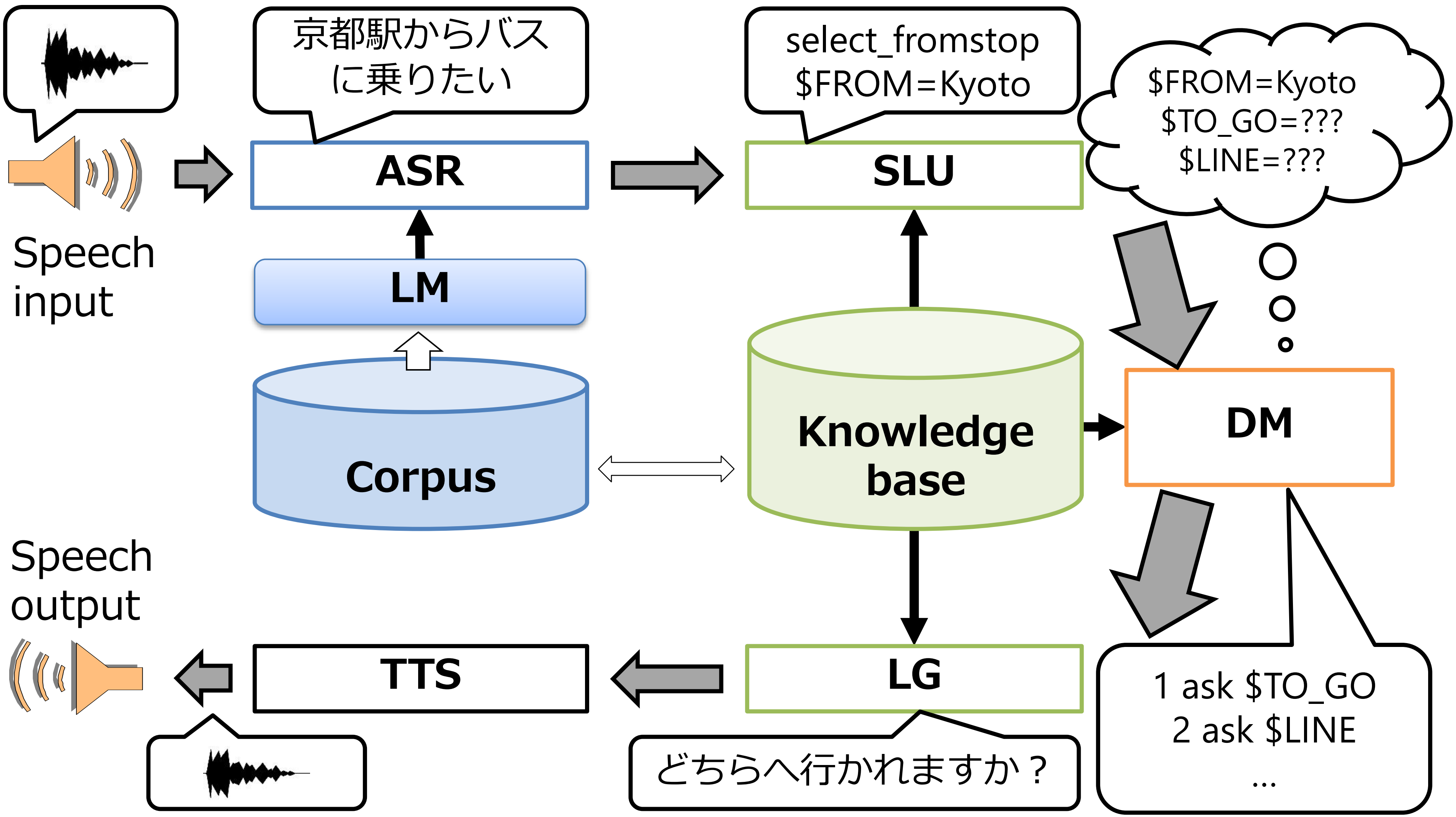
http://www.pomdp.net

Nara Institute of Science and Technology
Augmented Human Communication Laboratory

Spoken language processing applications

- **What can we do with Speech?**
  - Phone call
  - Control car-navi
- **Everyone knows we can use speech to control some devices**
  - Really uses speech?
- **What is realized by current SLP systems?**
  - Real SLP applications
  - Understand the performance of current speech recognition to build a SLP system

An architecture of SLP application



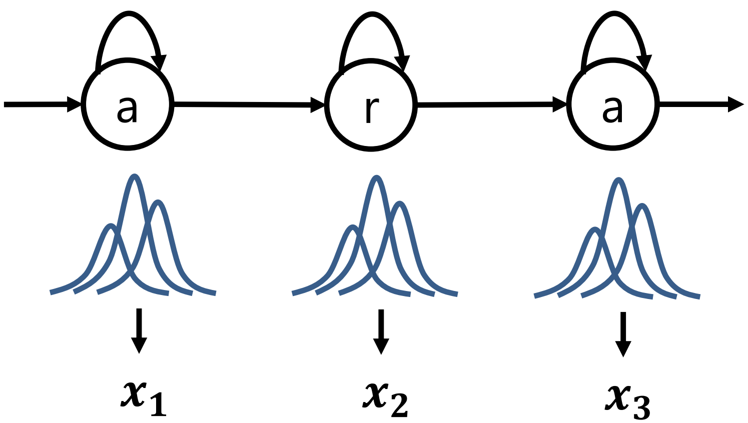
ASR and Deep Neural Network (DNN)

- **System of speech recognition**

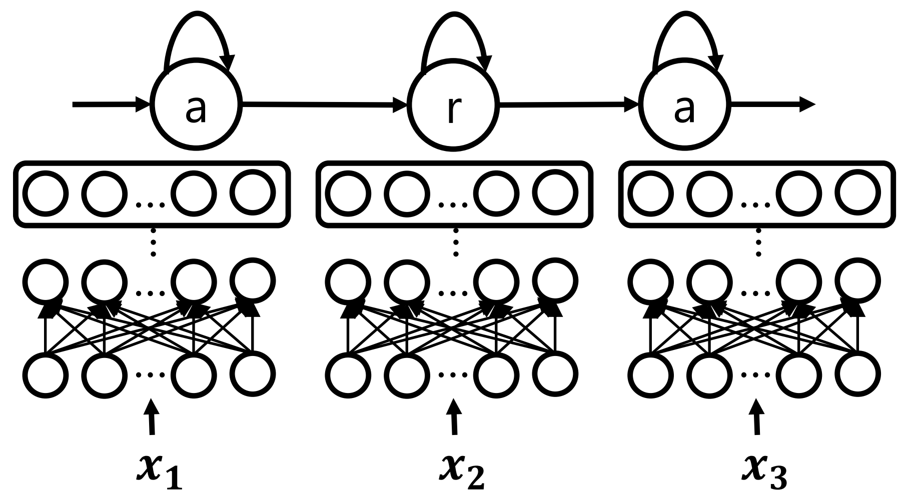
$$\underset{W}{\operatorname{argmax}}\, P(W|X) = \underset{W}{\operatorname{argmax}}\, \underbrace{P(X|W)}_{\text{Acoustic model}}\underbrace{P(W)}_{\text{Language model}}$$

W: **words** X: **speech**

GMM-HMM



DNN-HMM



State-of-the-art ASR system

- **Current systems**
  - Large-scale speech recognition on cloud cluster machines
    - Open-domain < Specified-domain
  - ASR with closed-talk microphone
    - Minutes of national congress, applications in smartphone

- **Future works**
  - Stand-alone ASR on mobile devices
    - Cloud: **Not real time**
  - Distant speech recognition
    - Beam forming, Microphone array processing

How do we design a SLP system?

- **Assume an inputted speech**
  - System reads the speaking style of users
  - Clarify the purpose of dialogue system
- **We must not assume 100% ASR accuracy**
  - Post-process of ASR must assume ASR error
  - Adaptation of ASR system (AM, LM)
- **Compare with competing devices**
  - Text input, QR, touch panel
  - More efficient input method than speech
  - Disadvantage of other input methods

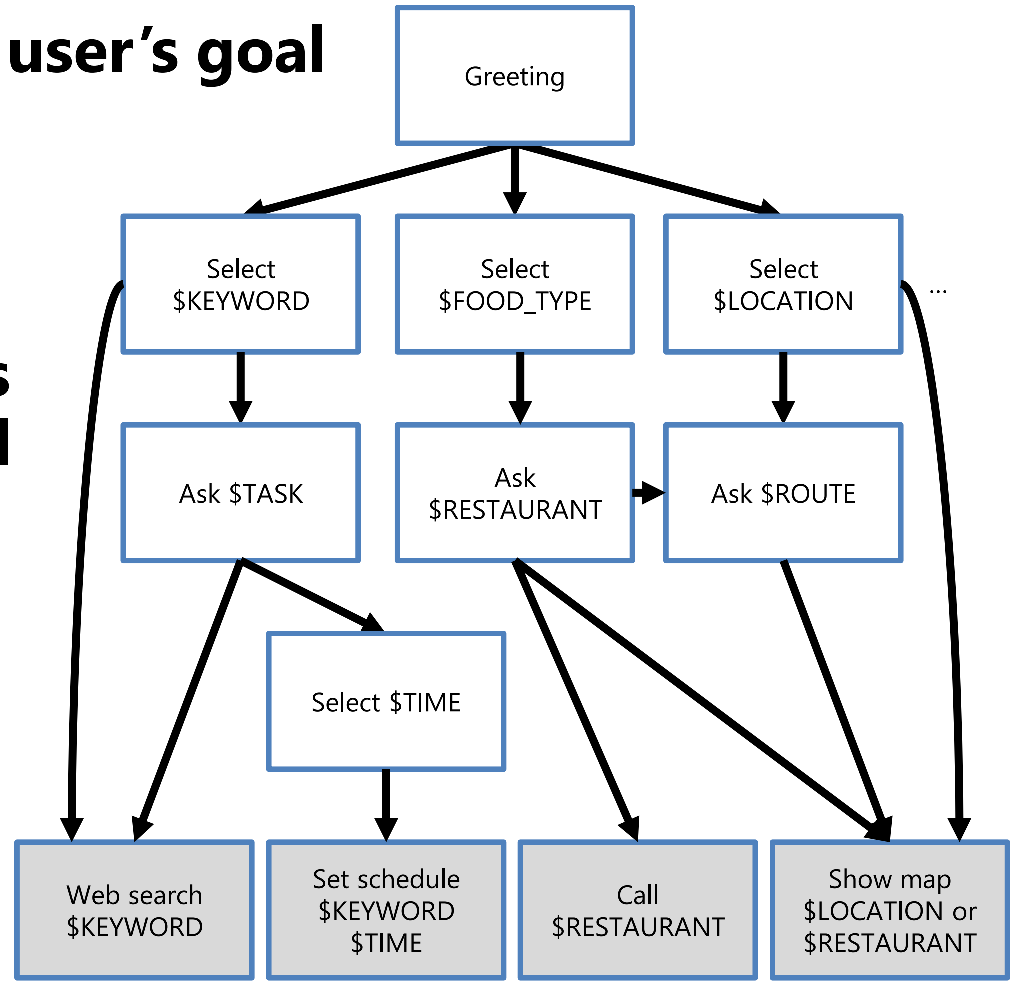
Task-oriented dialogue systems

- **System tries to reach the user's goal**
  - Ticket reservation
  - Restaurant navigation

- **Task/domain knowledges match with assumed goal**

  e.g. automaton + RDB

- **Work on well defined task/domain☺**
- **High cost to define task/domain☹**



Goal, Task, and Domain knowledge

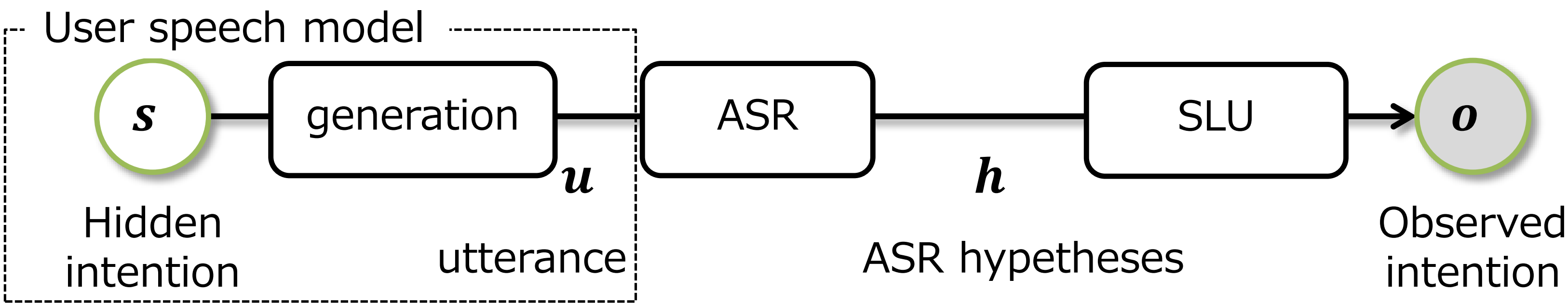
- **Goal**
  - **Purpose of dialogue shared between attendees (user and system)**
    - Bus navigation: Departure time of the next bus to Kyoto, …
    - QA system: Height of Mt. Fuji, Entrance fee of Kinkakuji-temple, …
- **Task**
  - Defined to reach a goal
    - Task-flow, question patterns, …
- **Domain knowledge**
  - Essential knowledges to realize the defined task
    - Names of bus stops, …

Kyoto bus navigation system

- **Flexible guidance generation using user model in spoken dialogue systems. Komatani et al. In Proc. ACL, pp.256—263, 2003.**
- **Real service of Kyoto city bus**
  - IVR automatically responds to users on phone call
- **Input From, To, and Bus number**
  - System responds when the next bus will arrive

- **Management: automatically generated VoiceXML**
- **vocabulary: bus stops: 652, place, building: 756**

SLU considers ASR error



- **User generates utterance from hidden intention**
- **ASR transcribes the speech to texts**
- **Spoken language understanding from text to intention**

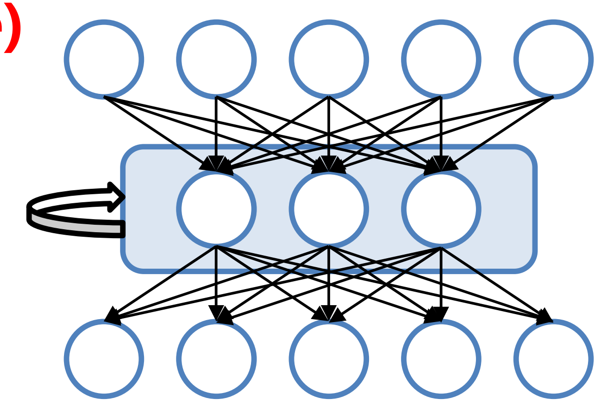
$$P(\boldsymbol{o}|\boldsymbol{s}) = \sum_{\boldsymbol{h}} P(\boldsymbol{o},\boldsymbol{h}|\boldsymbol{s}) \approx \sum_{\boldsymbol{h}} \underbrace{P(\boldsymbol{o}|\boldsymbol{h})}_{\text{SLU probability}} \underbrace{P(\boldsymbol{h}|\boldsymbol{u})}_{\text{ASR probability}}$$

Dependence to dialogue histories

- $b' = P\left(s^{t+1} \middle| o^{1:t+1}\right) \propto \underbrace{P(o' | s'_j)}_{\text{Observation}} \sum_{s_i} \underbrace{P\left(s'_j \middle| s_i, \widehat{a_k}\right)}_{\text{Transition}} \underbrace{b^t}_{\text{Current belief}}$
  - $s \in I_s$ user state
  - $a \in K$ system action
  - $o \in I_s$ **observed state**
  - $b_s = P(s | o^{1:t})$ **belief of user states (stochastic variable)**

- **Traditional state transition model → Recurrent Neural Network**

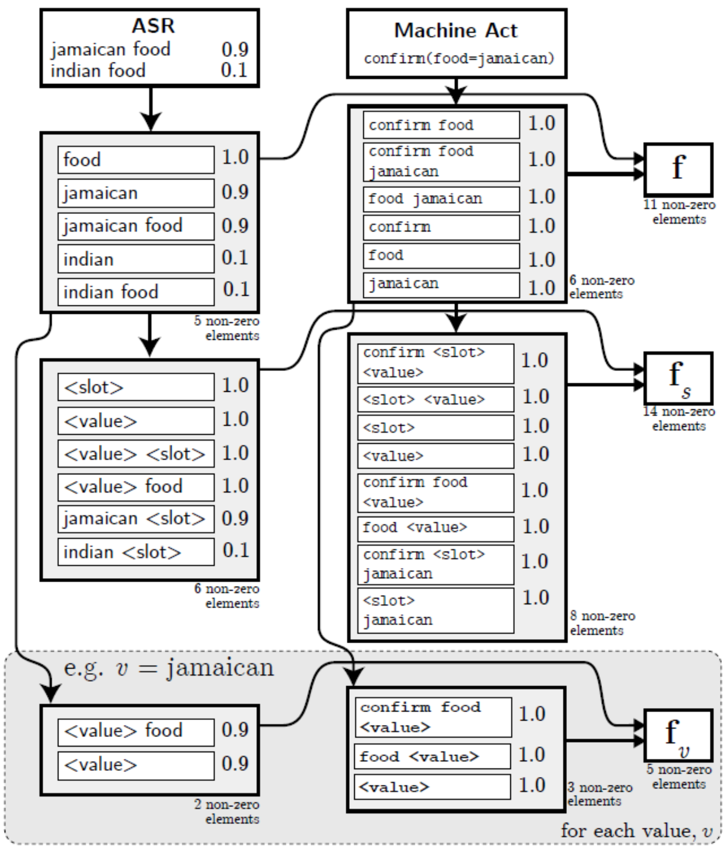
Input vector (user utterance)

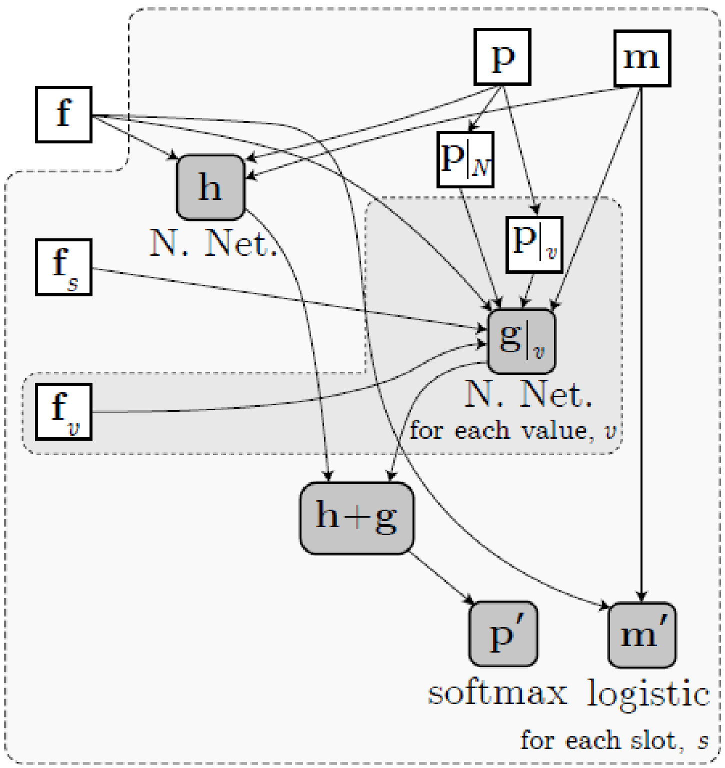


Output (belief)

RNN based SLU

- **Word-Based Dialog State Tracking with Recurrent Neural Networks. Henderson et al., In Proc. SIGDIAL, pp, 292-300,**



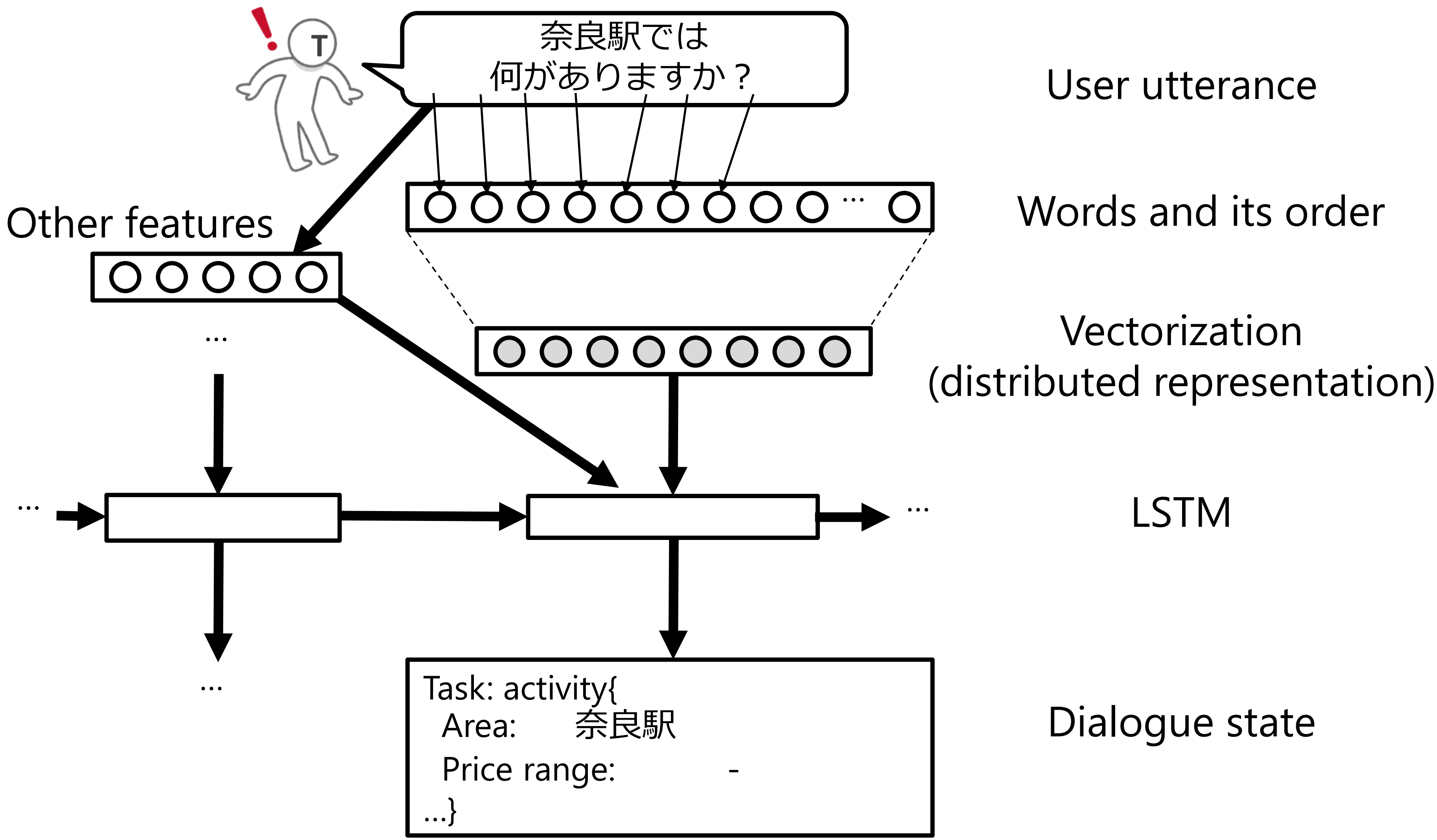


図は論文より引用

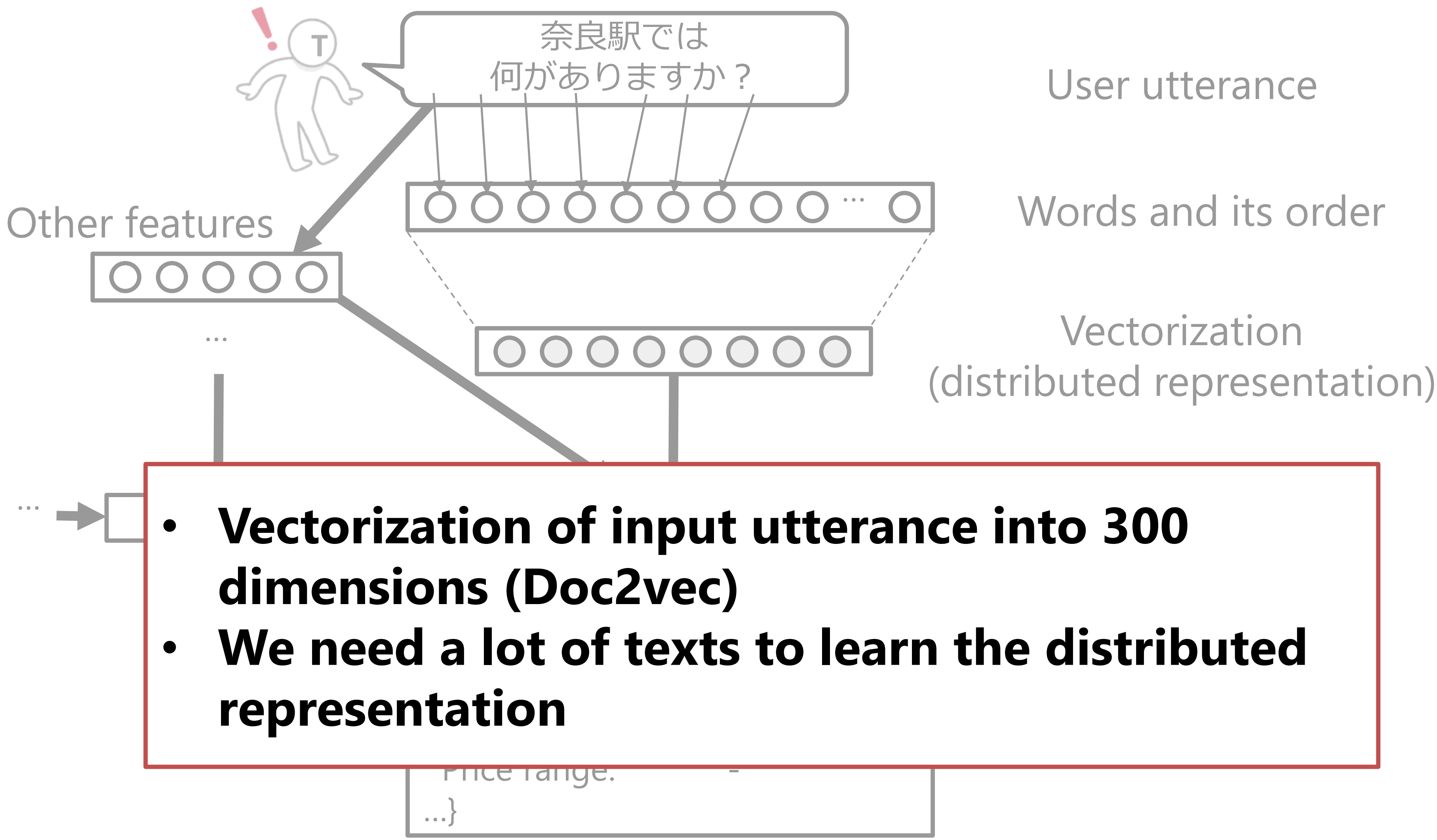
Recurrent Neural Network → Long Short Term Memory Neural Network

- **LSTM can keep more distant information than RNN**
- **Dialogue State Tracking using Long Short Term Memory Neural Networks. Yoshino et al., In Proc. IWSDS, 2016.**
- **Context Sensitive Spoken Language Understanding using Role Dependent LSTM layers. Hori et al., In Proc. NIPS-WS, 2015.**
- **Incremental LSTM-based Dialog State Tracker. Zuka et al., In Proc. ASRU, 2015.**

LSTM-based SLU



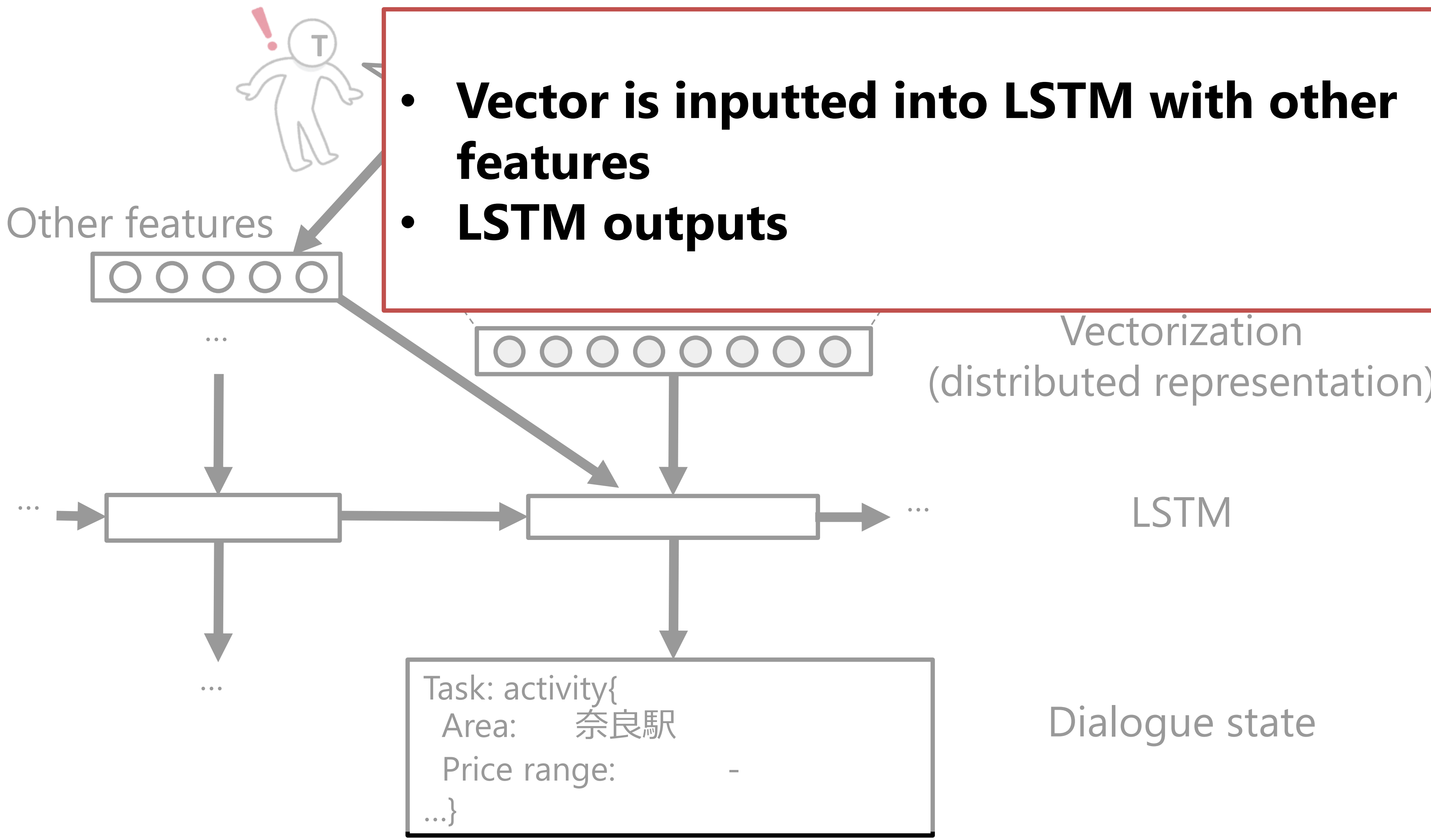
LSTM-based SLU



- **Vectorization of input utterance into 300 dimensions (Doc2vec)**
- **We need a lot of texts to learn the distributed representation**

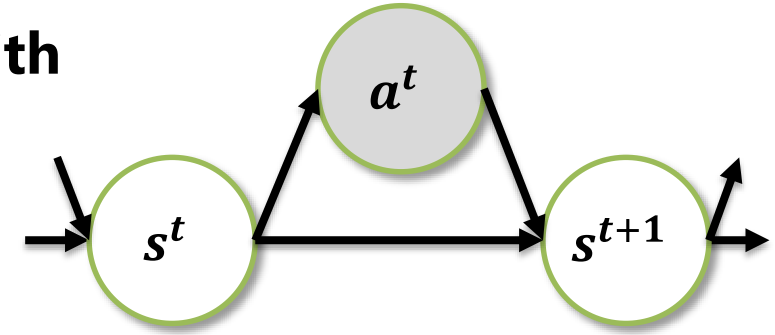
LSTM-based SLU

- **Vector is inputted into LSTM with other features**
- **LSTM outputs**



Action selection given SLU results

- s^t: **user action in turn** t
  - actions: Select $FROM, Select $TO_GO …
  - histories: $FROM=神保町駅, $LINE=半蔵門線
- a^t: **system action in turn** t
  - next actions: Ask $TO_GO, Ask $LINE, Confirm …
- **User action can be represented with** $P(s^{t+1} | s^t, a^t)$ **Markov property**
  - **Apply reinforcement learning**



Dialogue management with RL

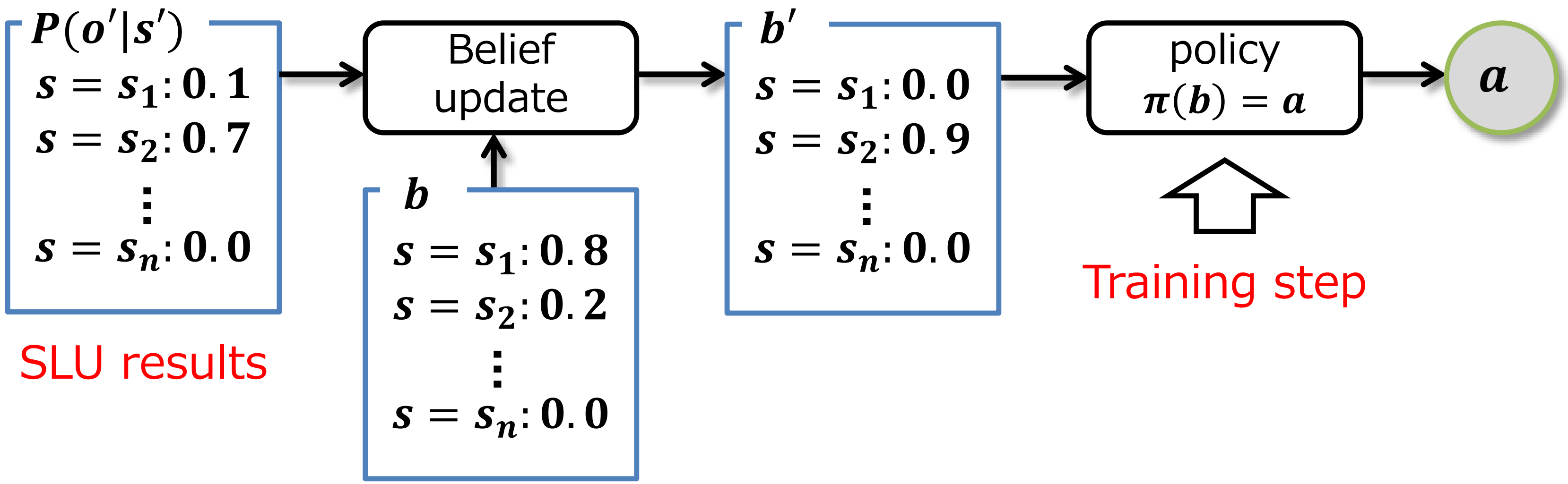
- $s \in I_s$ user state
- $a \in K$ system action
- $R(s, a)$ reward gives reward and penalty
- $\pi(s) = a$ policy learn in RL
- ε learning rate
- γ discount factor

- Select a policy function that maximize the value function $V^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^k R(s^{t+k}, a^{t+k})$
- Q-value is used in Q-learning to get the optimal policy
  - $Q(s^t, a^t)$ $\underset{update}{\longleftarrow} (1 - \varepsilon) Q(s^t, a^t) + \varepsilon \left(R(s^t, a^t) + \gamma \max_{a^{t+1}} Q(s^{t+1}, a^{t+1}) \right)$

Action selection for ambiguous SLU results

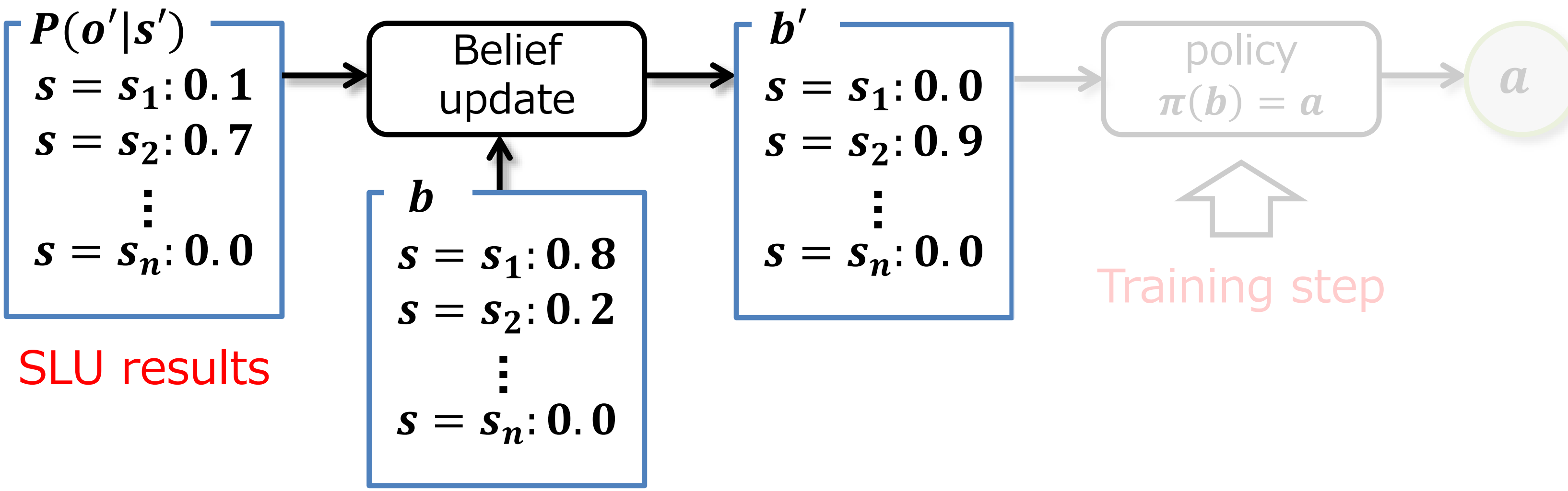
- **ASR result must include errors**
- **SLU results are given as stochastic variable**
  - Application decides an action given a variable
  - Not $\boldsymbol{s}$, given $\boldsymbol{b_s}$
- **Decision making with Partially Observable Markov Decision Process (POMDP)**
- **Learn the optimal policy $\pi^*(b) = a$ in partially observable situation**
  - **One of the major problem of SDS**
  - **Dialogue data for training is limited**

POMDP based DM



- $s \in I_s$ **user state**
- $a \in K$ **system action**
- $o \in I_s$ **observed state**
- $b_i = P(s_i|o^{1:t})$ **belief of $s = s_i$ (stochastic variable)**
- $R(s, a)$ **reward**
- $\pi(b) = a$ **policy**

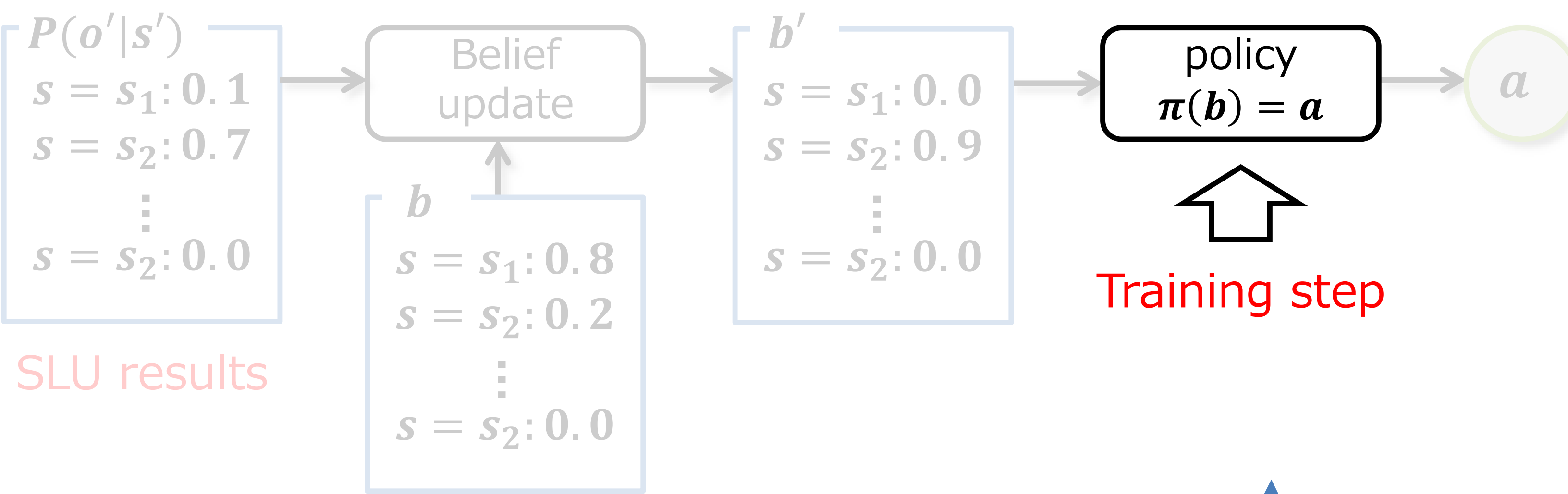
Belief update of POMDP



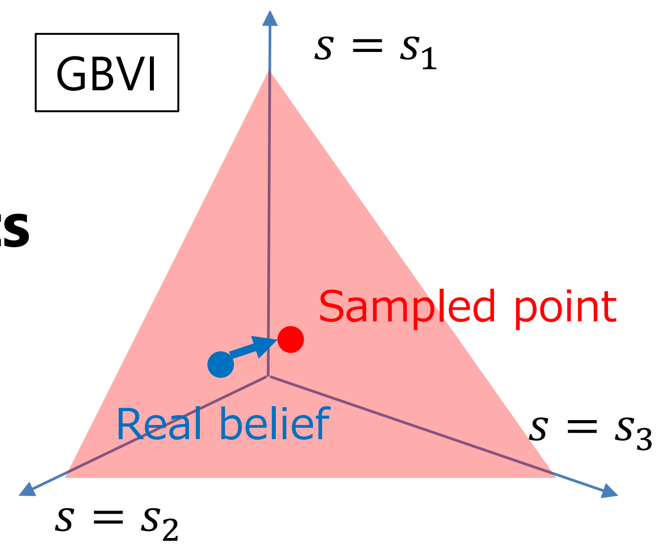
- $b' = P\left(s^{t+1} \middle| o^{1:t+1}\right) \propto \underbrace{P(o'|s'_j)}_{\text{observation}} \sum_{s_i} \underbrace{P\left(s'_j \middle| s_i, \widehat{a_k}\right)}_{\text{transition}} \underbrace{b^t}_{\text{current belief}}$

- **Update belief**
  - Update belief is sent to the policy to decide the next action

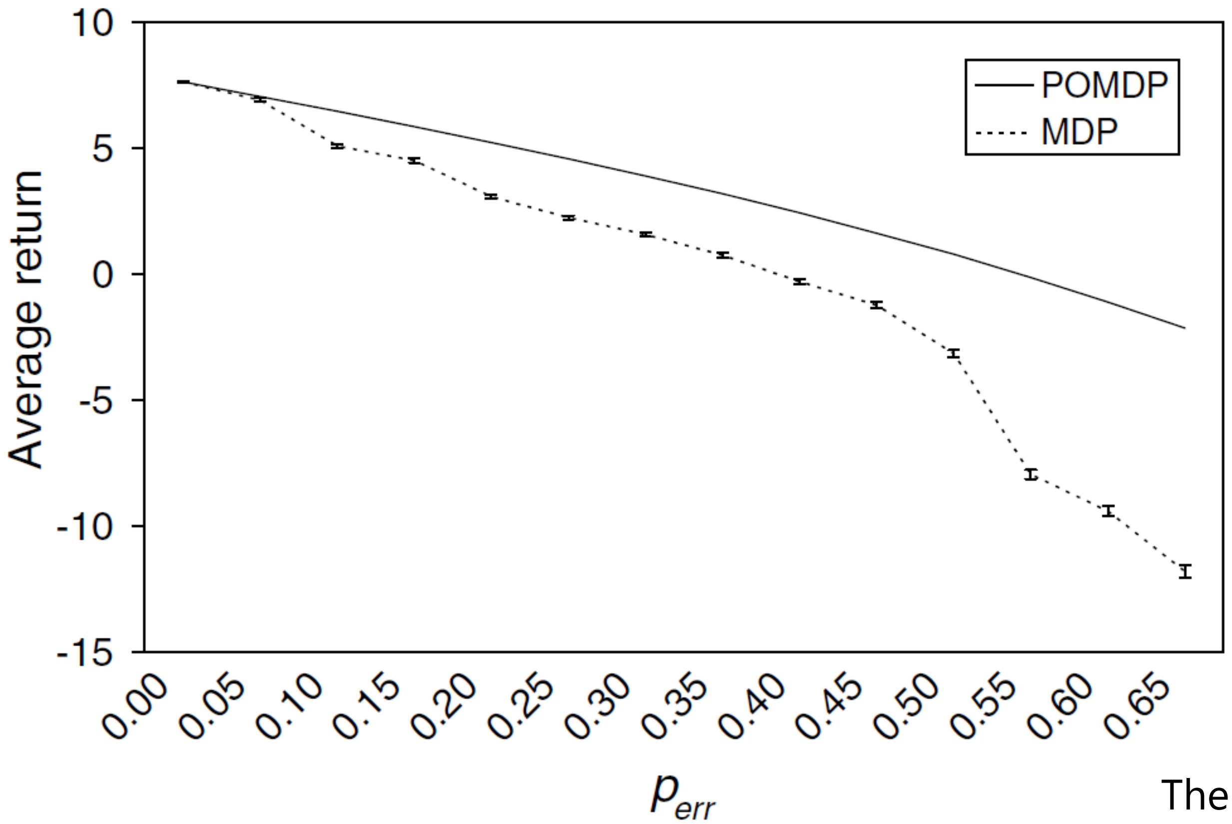
(Typical) POMDP learning



- **Sample several points on belief**
- **Maximize $Q(b, a)$ on sampled points**
- **Train policy in simulated dialogue with a simulator**



MDP → POMDP



The figure is cited from the paper

- **POMDP works robustly in much error cases**
  - Partially observable Markov decision processes for spoken dialog systems. Williams et al., Computer Speech & Language, 393—422, Vol.22, No.1, 2007.

Problems when we apply POMDP to SDS

- **Not enough data to learn optimal $\pi^*()$**
  - **Efficient training method is required**

1. **Hybrid of rule and POMDP**
2. **Efficient sampling**
3. **Efficient calculation of Q-function**

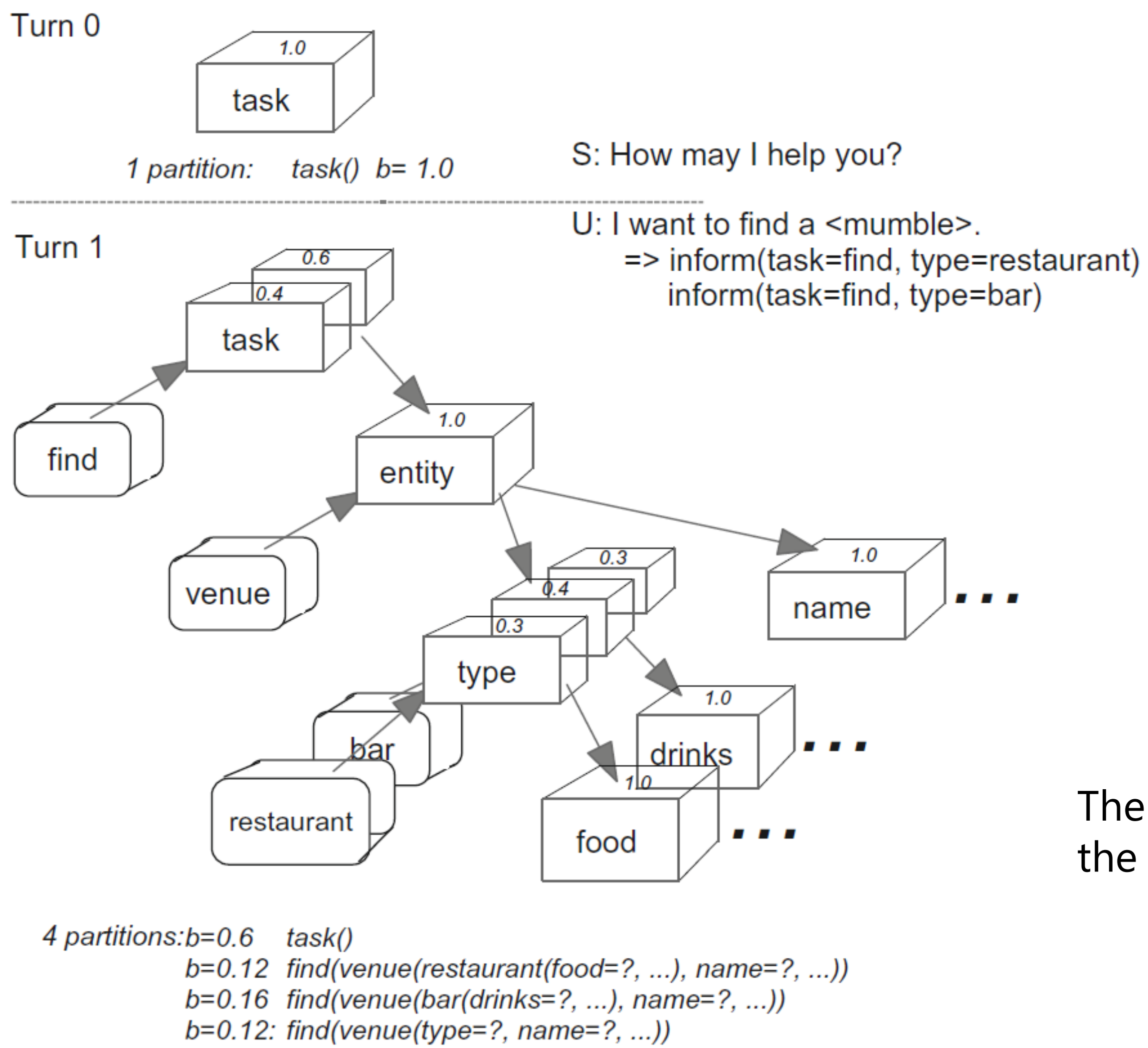
Hybrid of rule and POMDP

- **The hidden information state model: a practical framework for POMDP-based spoken dialogue management**
  Young et al., Computer Speech & Language, Vol.24, No.2, pp.150-174, 2010.

- **Statistical dialogue management using intention dependency graph. Yoshino et al., In Proc. IJCNLP, pp.962-966, 2013.**

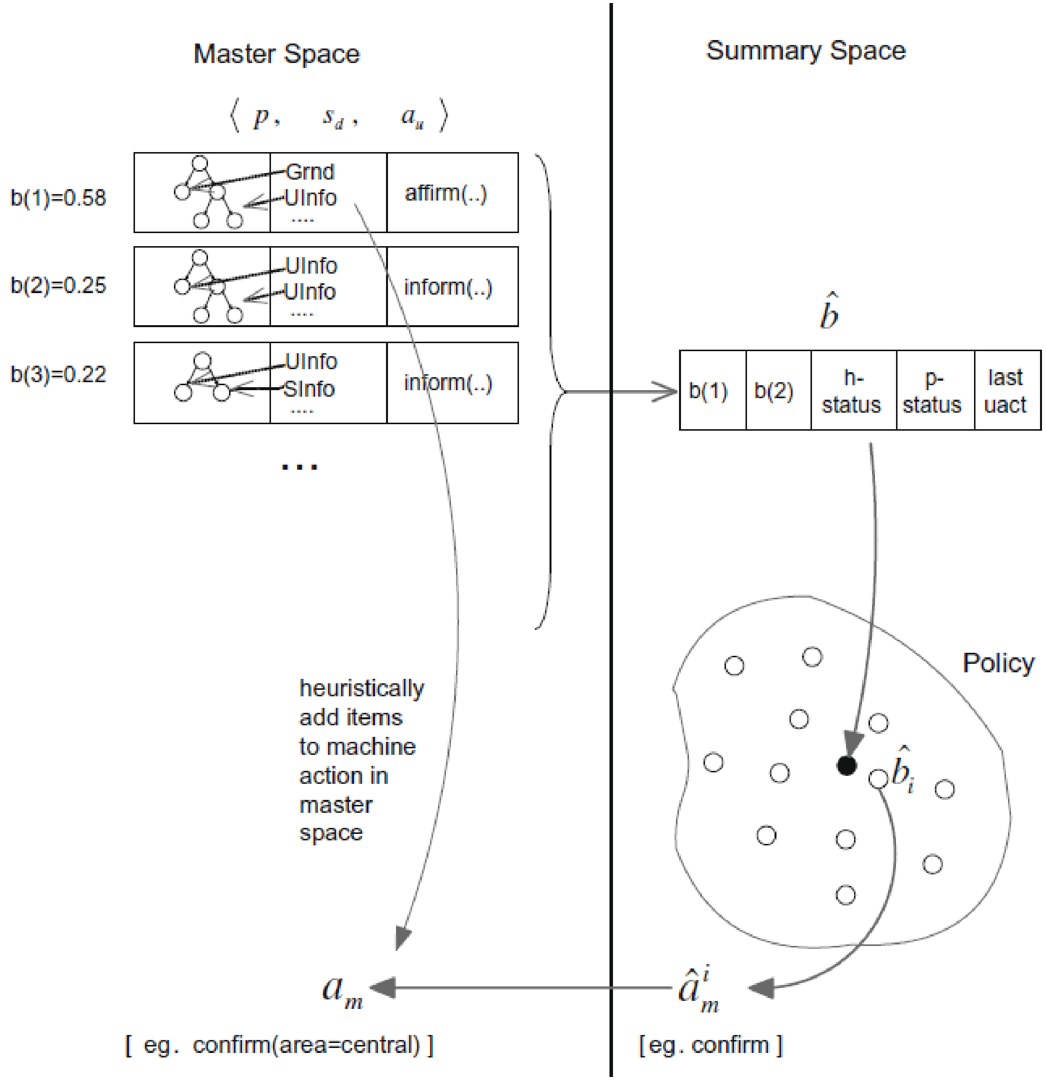
- **Handcrafted rules are used as restrictions of search space**

Hidden Information State Model



The figure is cited from the paper

Hidden Information State Model



The figure is cited from the paper

Intention Dependency Graph

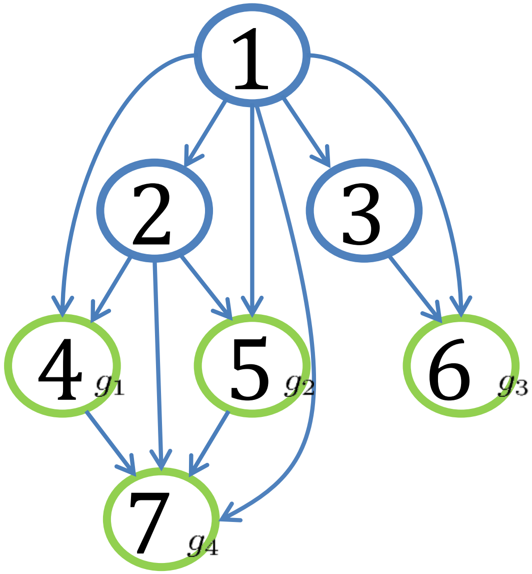
- **Define state transition probabilities in pre-defined task structures**

1. ROOT[] (=no specified request)

2. PLAY_MUSIC[artist=null, album=null]
3. CONTROL_VOLUME[value=null]

4. PLAY_MUSIC[artist=$artist_name, album=null]
5. PLAY_MUSIC[artist=null, album=$album_name]
6. CONTROL_VOLUME[value=$up_or_down]

7. PLAY_MUSIC[artist=$artist_name, album=$album_name]

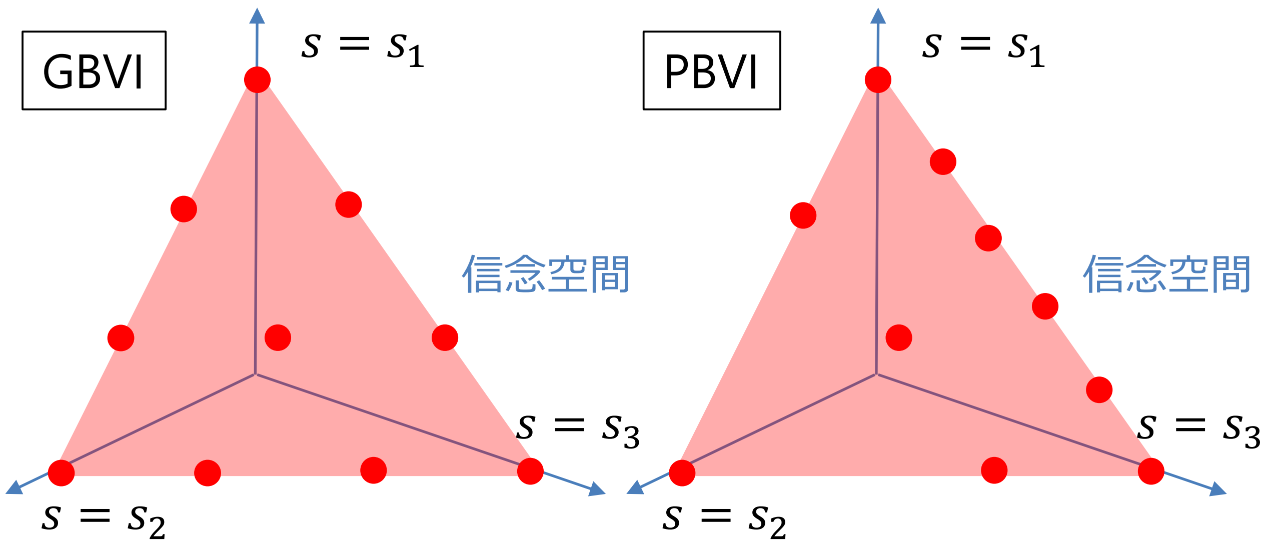


Benefits to use rules and task structures

- **To launch a new system**
  - A rule-base system is used to collect data
  - The rule-based manager can be shifted to statistical one
- **Weight unobserved states, sequences**
  - It is impossible to cover all possible situations by training data (dialogue data)
- **Adaptation for unobserved states, new domains**

Efficient sampling

- **Grid-based value iteration is not efficient**



- **GBVI: select belief points with equable grids**
- **PBVI: bias belief points according to the already observed states**
  - s_1 and s_3 are confusable states

Efficient calculation of Q-function

- **Learning of POMDP = maximization of $\boldsymbol{Q(b, a)}$**
  - Define Q calculation for all possible pairs of $\boldsymbol{b}$ and $\boldsymbol{a}$
- **Calculate a similarity of $(\boldsymbol{b_i}, \boldsymbol{a_i})$ and $(\boldsymbol{b_k}, \boldsymbol{a_k})$ to calculate new $\boldsymbol{Q(b_k, a_k)}$ with known $\boldsymbol{Q(b_i, a_i)}$**
  - Define a kernel function to calculate the similarity
- **Bayesian update of dialogue state: A POMDP framework for spoken dialogue systems. Thomson et al., Computer Speech & Language, vol. 24, no. 4, pp. 562–588, 2010.**

Question answering systems

- **There are several shared tasks in 2000s**
  - NTCIR etc…
- **IBM Watson**
  - Collect massive good question-answer pairs
    - Won in Jeopardy! TV show
- **Several systems that use inference are research in recent**

- **Major spoken dialogue systems are consist of**
  - **task oriented system for typical tasks**
  - **question answering system**
  - **search engine of Web**

Important points to construct a task oriented system

- **Good task design: make it clear the user's intention**
  - Users often encounter a situation that they don't know what to say
    - To be a good first-food clerk
- **Necessary and sufficient task structure**
  - Roughly does not achieve anything
  - Detailed structure disturb a learning of management
- **Fall-back of system**
  - E.g. call Web search when anything does not hit

Important points to construct a non-task oriented system

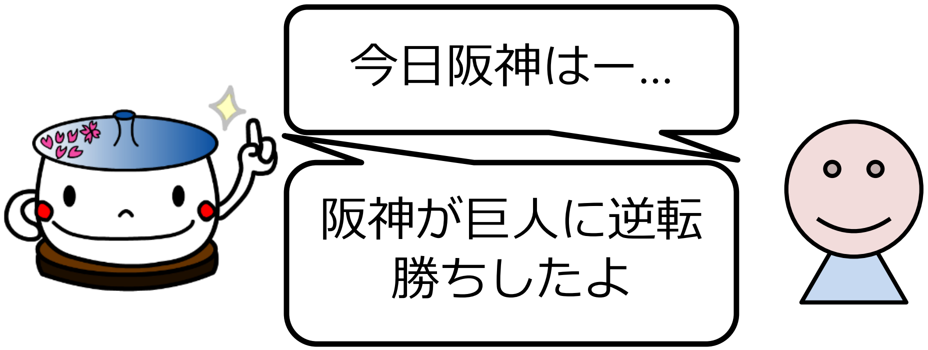
- **Conversational System for Information Navigation based on POMDP with User Focus Tracking. Yoshino et al., Computer Speech & Language, Vol.34, Issue.1, pp.275--291, 2015.**

- **Good task design: design that works well even if the user intention is ambiguous**
  - Proactive contact form the system to clarify the user's intention
  - Design of information extraction that works for ambiguous queries
- **Classify user intentions to several classes**
  - On a limited situation
- **Introduce a new observation state that matches to the system concept**
  - Information navigation: user focus, topic of the dialogue
  - Emotional system: emotion state

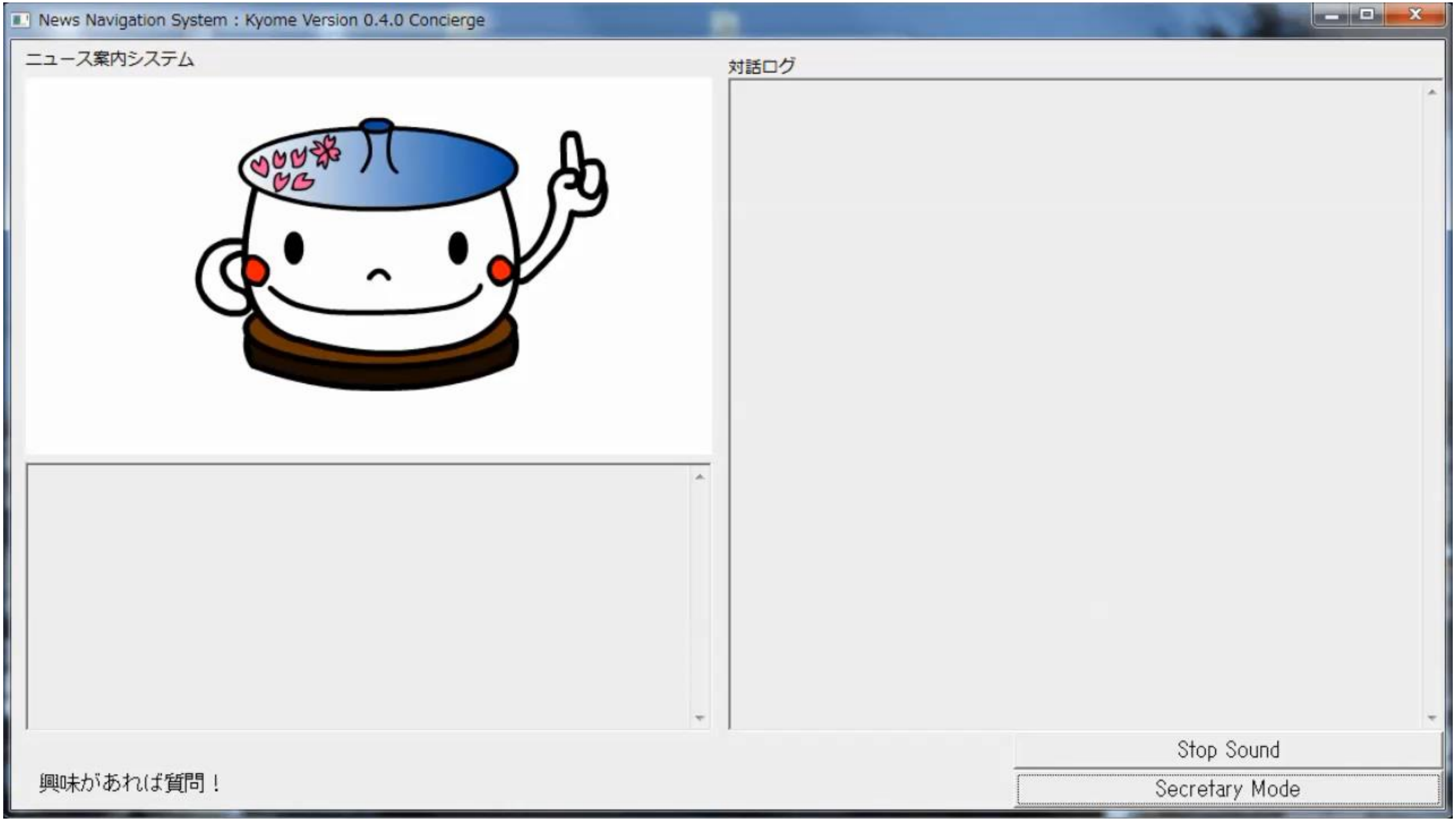
Information navigation task

- **Navigate contents written in knowledge source (doc)**
  - News texts updated day-by-day
  - Limit a domain (baseball, football, economics, etc…)
    - Use automatically extracted domain knowledge

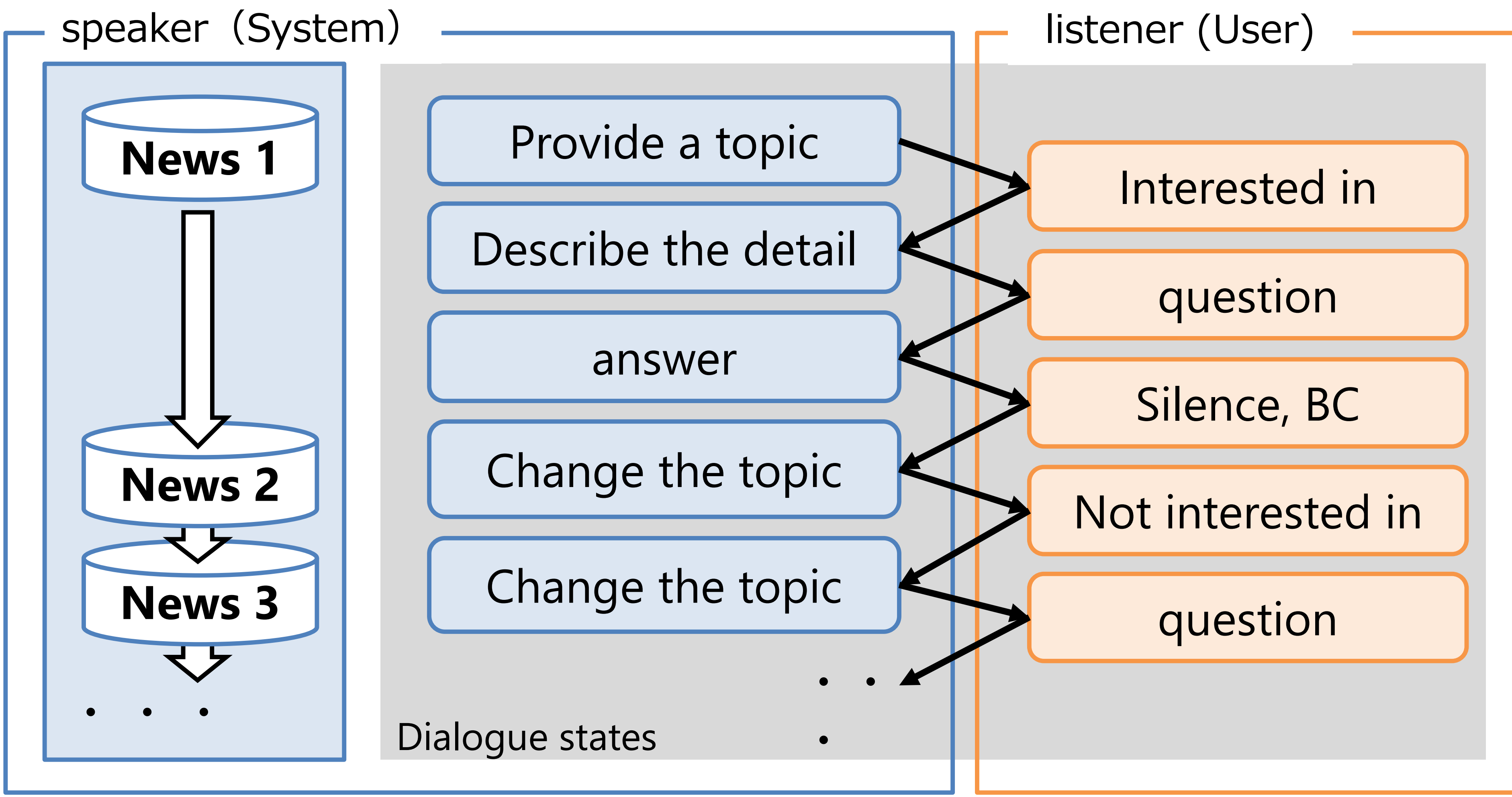
- **The system takes a role of speaker**
  - Clarify what the listener want to know
    - Domain knowledge
    - User intention
    - User focus



Information navigation system



情報案内対話の構造


speaker（System）
News 1
News 2
News 3
・・・
Provide a topic
Describe the detail
answer
Change the topic
Change the topic
・・
・
Dialogue states
listener (User)
Interested in
question
Silence, BC
Not interested in
question

States and actions in information navigation

- **Dialogue management activates the appropriate dialogue module**

User state s:	**system action a:**
– ***TP***: request to present a topic	**TP**: topic presentation
– ***ST***: request to describe detail	**ST**: story telling
– ***QA***: question	**QA**: answer
– ***GR***: greeting	**GR**: greeting
– ***II***: meaningless input due to ASR errors	**KS**: don't do anything **CO:** confirmation
– ***NR***: silence	**PP**: proactive presentation

Introducing a user focus

Example 1

…
Usr: Where was **Tanaka** training?
Sys: Tanaka practiced at the bullpen in the New York Yankees' camp on Feb. 20th.
Usr: (silence)
Sys: By the way, Tanaka also practiced a drop two-seam on Feb. 18.
…

Example 2

…
Usr: What happened?
Sys: The godzilla cannon came back to the Giants‘ spring camp in Miyazaki.
Usr: (silence)
Sys: To be a ace pitcher, has Fujinami improved from the rookie year?
…

- **System continue the dialogue if the user has a focus (=attentional point) (Example 1)**
- **System change a topic if the user does not have any focus (Example 2)**

Extension of dialogue management

- **Introduce a Boolean $f = 0\ or\ 1$ (user have a focus or not)**
  - $b' \propto P(o'_{s'}, o'_{f'} | s'_j, f'_m) \sum_i \sum_l P(s'_j, f'_m | s_i, f_l, \widehat{a_k}) b^t_{s_i, f_l}$
- **observation**
  - $P\left(o'_{s'}, o'_{f'} \middle| s'_j, f'_m\right) \approx P\left(o_s^{t+1} \middle| s'_j\right) P\left(o_f^{t+1} \middle| f'_m\right)$
- **transition**
  - $P\left(s'_j, f'_m \middle| s_i, f_l, \widehat{a_k}\right) = P(f'_m | f_l, s_i, \widehat{a_k}) P\left(s'_j \middle| f'_m, f_l, s_i, \widehat{a_k}\right)$
- **policy**
  - $\hat{a} = \pi^*\left(b_{s,f}\right)$

Evaluation of information navigation

- **Criteria**
  - DST (accuracy of dialogue state tracking)
  - ACT (accuracy of action selection)

- **Evaluation data**
  - 12 users, 24 dialogues, 626 utterances with real user
  - Annotate $\boldsymbol{s}$ for each utterance and the appropriate action $\boldsymbol{a}$ for the utterance
  - Annotator agreements
    - $\boldsymbol{s}$: 0.958 (kappa=0.938)
    - $\boldsymbol{a}$: 0.944 (kappa=0.915)

Evaluation of information navigation

	Rule	POMDP w.o. focus	POMDP proposed
DST	0.812 (=508/626)	0.853 (=534/626)	**0.867** (=543/626)
ACT	0.788 (=539/684)	0.751 (=514/684)	**0.854** (=584/684)

- **DST is improved by introducing user focus**
- **ACT is improved by introducing user focus (significant)**
- **Proposed method proactively present a related information according to the user's interest**
  - 35 proactive presentations evoked 17 more questions of users

Construction of non-task oriented systems

- **Clearly define the situation of dialogue**

  →to know information to be observed, and system action

- **Classification of user intention into good granularity**
  - It depends on which method will you use
  - Rule is still efficient in some cases
    - Statistical method is not necessary

Generation

- **Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. Wen et al., In Proc. EMNLP, 2015.**

- **Existing approaches of generation in dialogue systems**
  - **Use rules, templates**
    - It is hard to generate variational responses
    - Difficult to extend other domains

  - **Statistical methods**
    - Solve problems of rules and templates
    - Often generate ungrammatical, meaningless sentences

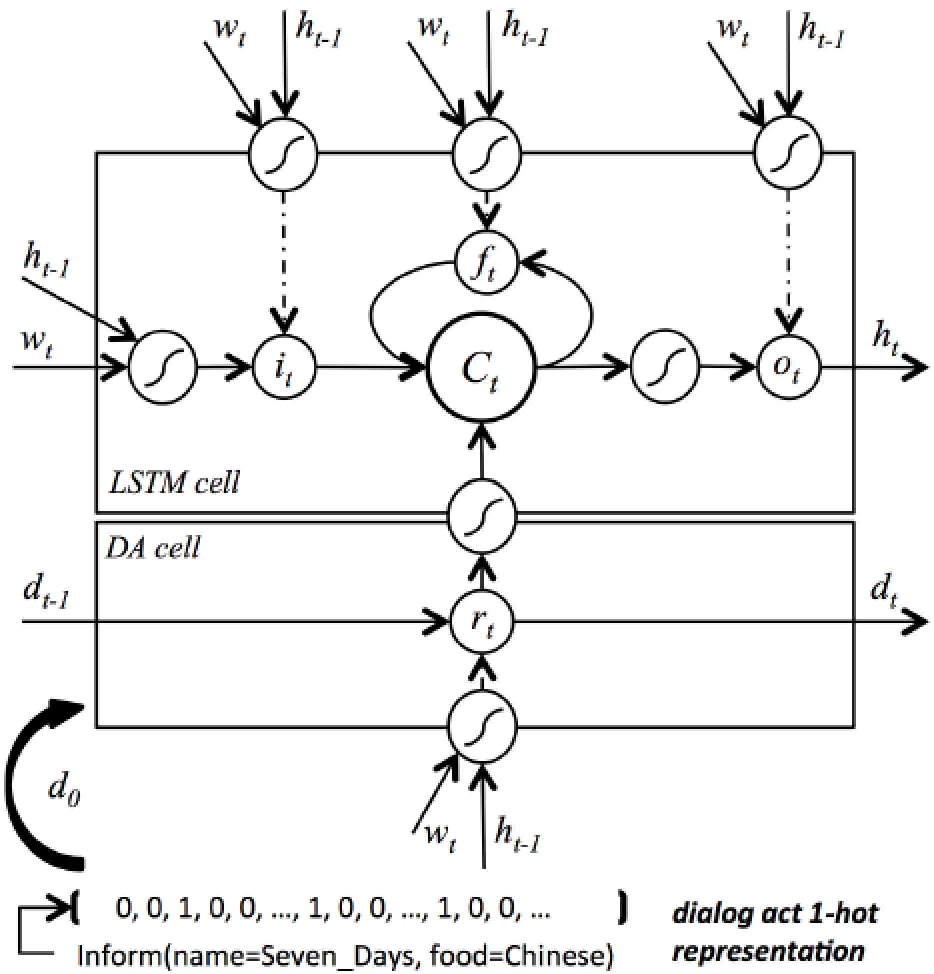
  - **Appropriateness, naturalness, understandability, variation**
    - It is difficult to fulfill all of them

Generation using LSTM

recurrent hidden layer

embedding of a word

1-hot dialog act and slot values



Upper LSTM sell is a language model to control “how to say”

Lower RNN cell is a semantic condition to control “what to say”

The figure is cited from the paper

End-to-end SLP

- **Skip SLU and DM**
  - Generate response given an input utterance
- **Example-based dialogue system**
  - Prepare large example base consist of pairs of a user query and a system response
  - Calculate the similarities between the utterance and queries in DB
- **Adaptive selection from multiple response candidates in example-based dialogue. Mizukami et al., In Proc. ASRU, 2015.**
  - Quantify the value of example with user satisfaction
  - Relevance feedback to adapt to the user's preference
- **End-to-end memory networks. Sukhbaatar et al., In Proc. NIPS, 2015.**
  - Directly generate response with Neural Network (LSTM) given a query

Risks to use statistical method on frontend

- **Statistical methods are hard to control (e.g. NN)**
  - Banned words, expressions
    - Microsoft AI praises Hitler
  - Grammatically correct, but semantically incorrect
- **Some filtering methods are proposed**
  - Filtering of training data
  - Evaluation metrics for semantic correctness
- **Understanding for applied method is required**
  - Some methods are used as black-boxs

Open domain system

- **Open-domain chat-oriented system**
  - Extend the example base to cover a variety of domains

- **Open-domain task-oriented system**
  - Policy committee for adaptation in multi-domain spoken dialogue systems. Gasic et al., In Proc. ASRU, 2015.
  - The system consists of several expert systems on different domains, and committee decides that which system is the most appropriate to speak in the current situation

Other applications in SLP

- **More natural, and adaptive TTS**
  - TTS for dictation already works better than non-native
- **Information security (for handicap people)**
  - SIG-AAC（IPSJ-SIG of accessibility)
- **CALL system for second language learner**
  - Assist listening by using ASR results as a caption
  - Assist speaking by using non-native ASR system
  - Government of China decided to develop a English CALL system as a national project
- **Communication skill training with SLP**
  - Developmental disorders
- **Supportive system for elderly people**

Future directions

- **Multi-modal system**
  - Gaze, gesture, etc…
- **Continuous dialogue**
  - Not limited by voice activity detection (VAD)
- **Incremental processing**
  - Real-time communication
- **Semantics**
  - How do we use meaning?