音声言語処理アプリケーション

奈良先端科学技術大学院大学
吉野 幸一郎

http://www.pomdp.net

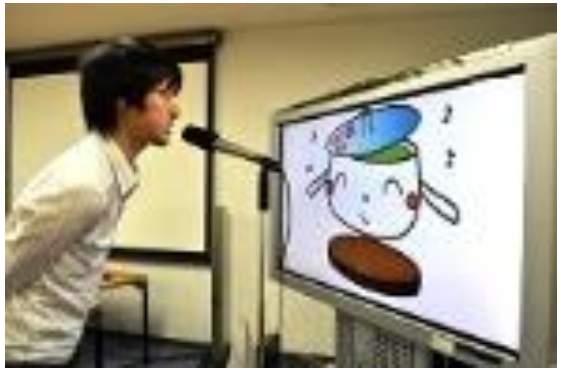
Nara Institute of Science and Technology
Augmented Human Communication Laboratory

自己紹介

- 吉野 幸一郎（よしの こういちろう）

- 2009年 慶大 SFC 石崎研卒業（自然言語処理）
- 2014年 京大 情報 河原研博士修了（音声言語処理）＋PD
- 2015年 NAIST 情報 中村研（音声言語処理、ビッグデータ）

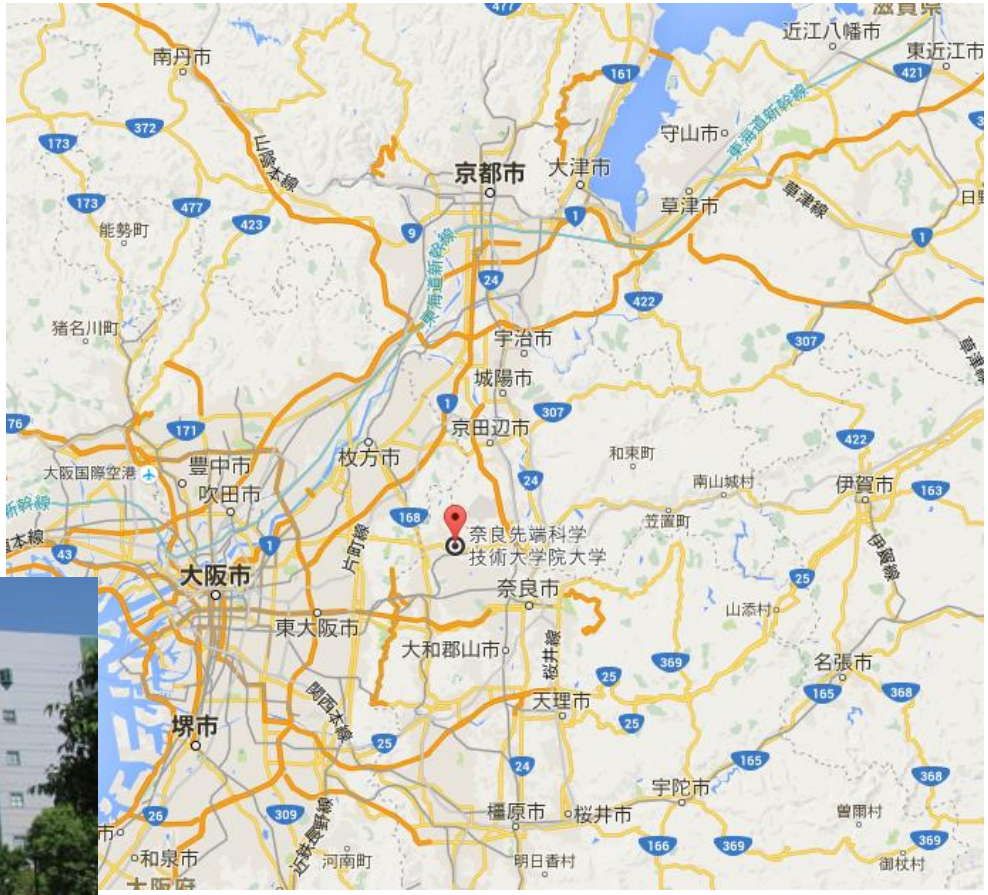
- 研究分野
  - 音声対話システム
  - 音声認識
  - 意味解析
  - ビッグデータ解析

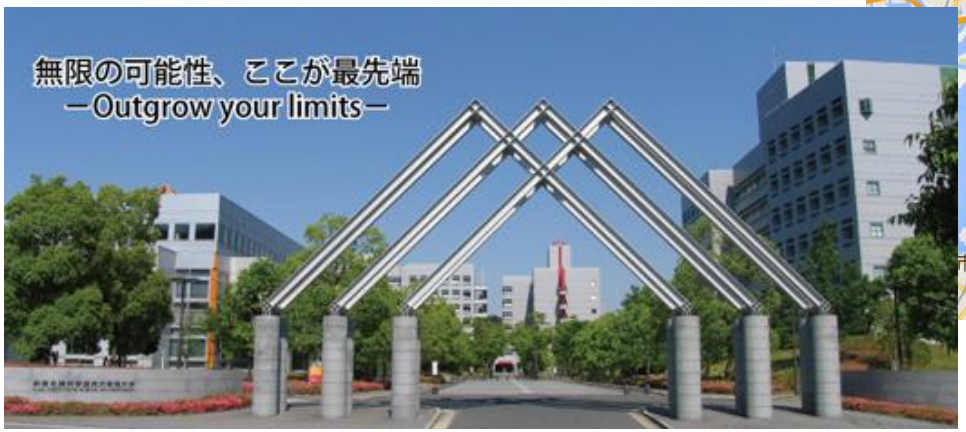


奈良先端科学技術大学院大学（NAIST）

1. 東京から京都へ2.5時間
2. 京都から高の原へ40分
3. 高の原からバスで30分

- 音声・言語処理が盛ん
  （中村研・松本研）



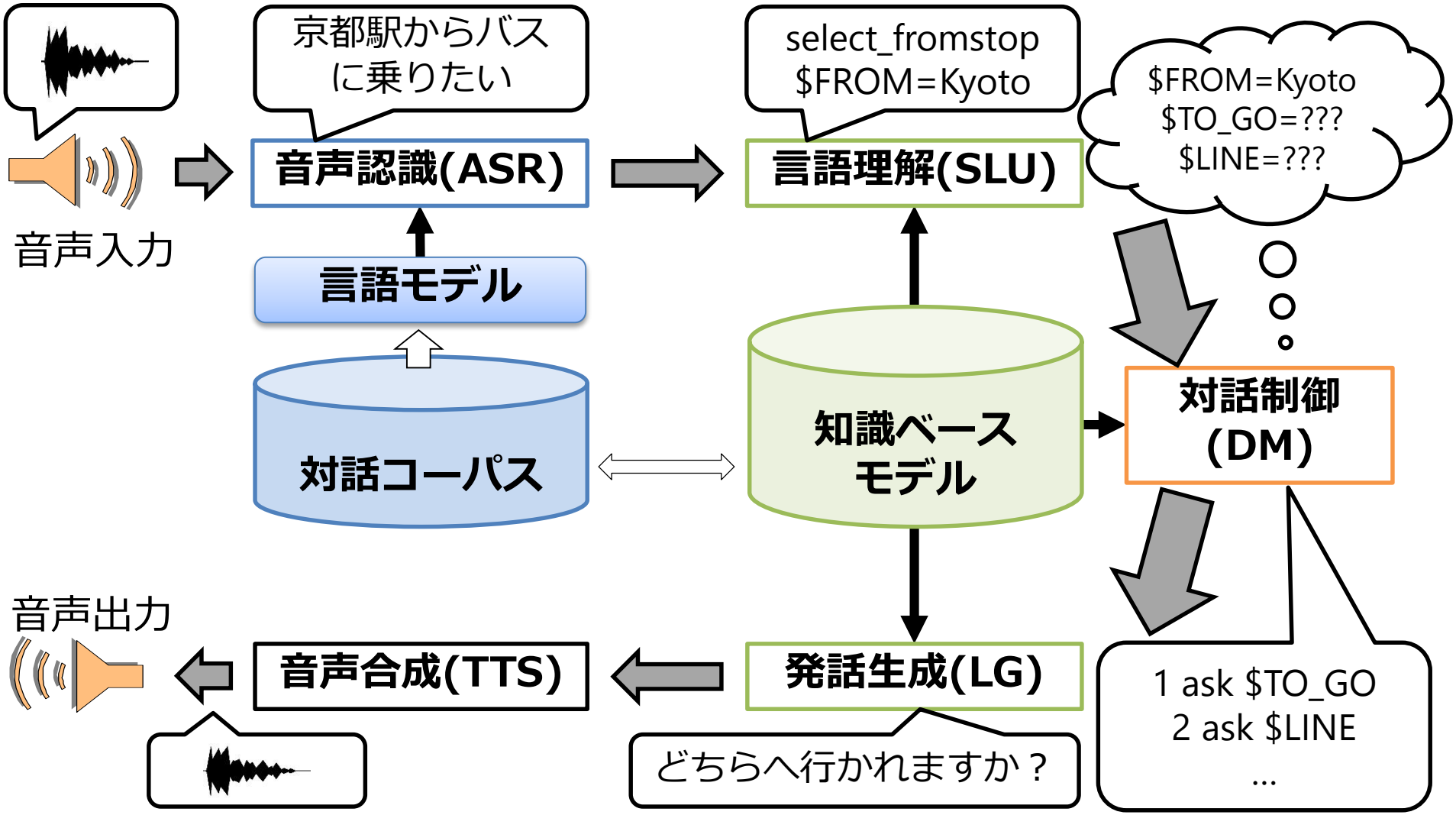


音声言語処理アプリケーション

- **「音声で何かを操作する」ことが普及**
  - 電話を掛ける
  - カーナビを操作する
    - 少なくとも「できる」とは認識されている
    - 使われているかどうかは別

- **現在の音声言語処理は何ができるのか？**
  - 実際に音声認識が使われている場面
  - 音声言語処理の性能を理解したアプリケーションの構築

音声言語処理アプリケーションの構造



音声認識と深層学習

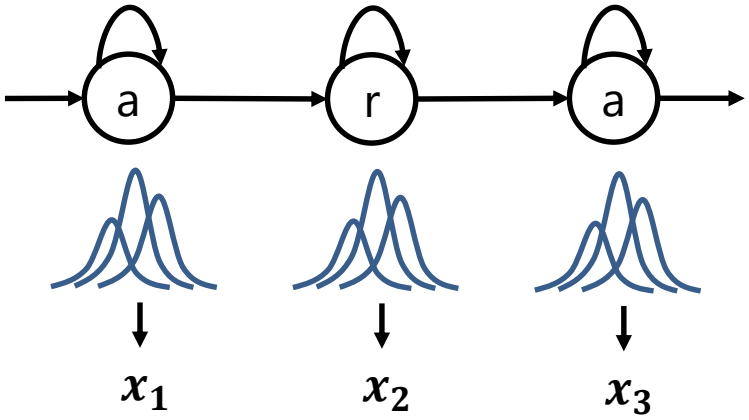
- 音声認識の仕組み

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

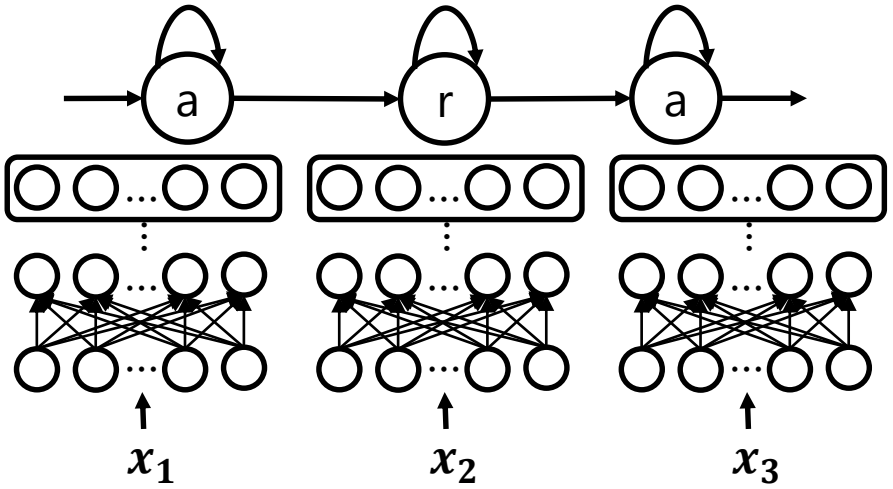
音響モデル　言語モデル

Wは単語列、 Xは音声

GMM-HMM



DNN-HMM



音声認識の現在

- **既にできること**
  - 大規模計算クラスタを用いたクラウドでの大語彙音声認識
    - ドメインに適応できればさらに認識精度は向上
  - 接話マイクでの音声認識
    - 衆議院の議事録作成、スマートフォンの認識アプリ

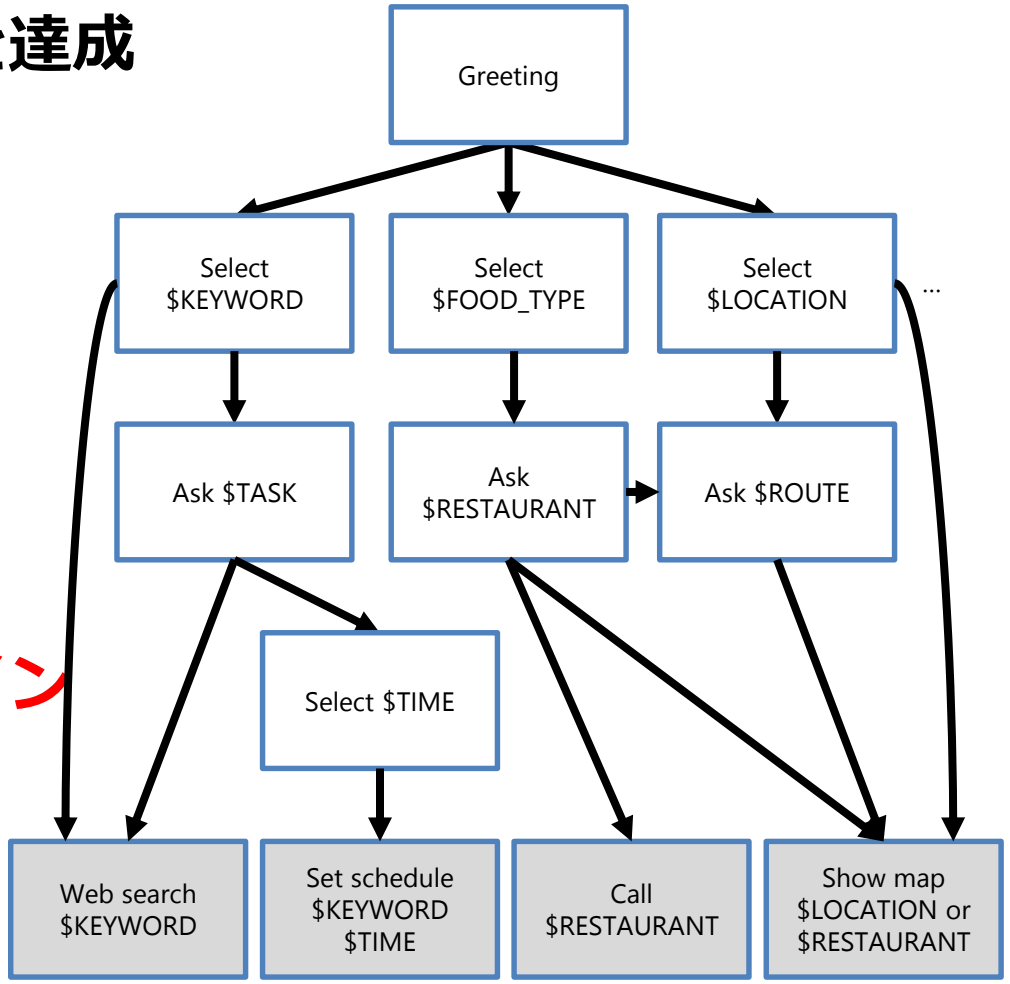
- **これからの課題**
  - モバイル上でのスタンドアローンでの音声認識
    - 現状は認識はクラウド → **リアルタイムではない**
  - 非接話での認識
    - 離れると劇的に認識精度が低下・マイクアレイが必要

実用的な音声言語処理アプリを作るために

- **入力される音声が想定しやすいデザイン**
  - システム側から発話の形を誘導する
  - 対話システムに対する目的を明確化する
- **音声認識率は100%にならない**
  - 必ず認識誤りを想定した処理を行う
  - 100%に近付ける努力（音響・言語モデルの適応）
- **競合する入出力インタフェースとの比較**
  - 音声よりも効率的な入力手段はないか
  - 他のモダリティが使いづらい状況

タスク指向型対話システム

- **ユーザの目的（ゴール）を達成**
  - チケットの予約
  - レストランの案内
- **ゴールに合わせたタスク・ドメイン知識の定義**
  e.g. オートマトン＋RDB
- **定義されたタスク・ドメインでよく動く☺**
- **タスクフロー・ドメイン知識の定義が大変☹**



ゴール・タスク・ドメイン知識

- **ゴール**
  - **対話参与者（ユーザとシステム）で共有される対話目標**
    - バス案内システム: 次の銀閣寺行きのバスの時間　…
    - 質問応答システム: 富士山の高さ、金閣寺の拝観料　…
- **タスク**
  - ゴールに到達するために定義される
    - タスクフロー、質問のパターンなど
- **ドメイン知識**
  - タスクを実現するのに必要な知識
    - バス停の名前　など

タスク指向型対話システムの成功例（京都市バス案内システム）

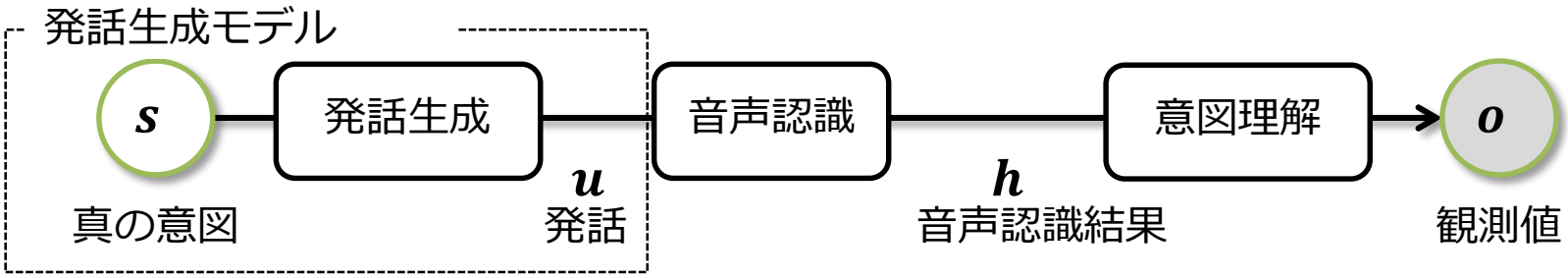
- Flexible guidance generation using user model in spoken dialogue systems. Komatani et al. In Proc. ACL, pp.256—263, 2003.

- 京都市バスのサービスとして実運用
  - サービスの電話番号に電話すると IVR（自動音声応答）

- 乗車場所、降車場所、系統番号を音声で入力
  - 指定したバスがどれくらいで到着するかが得られる

- 制御: VoiceXMLを動的に生成
- 語彙: バス停: 652, 名所・施設: 756

認識誤りを考慮した言語理解



- **ユーザの発話生成（真の意図から発話）**
- **音声認識（声を発話内容へ）**
- **意図理解（音声認識結果を意図理解結果へ）**

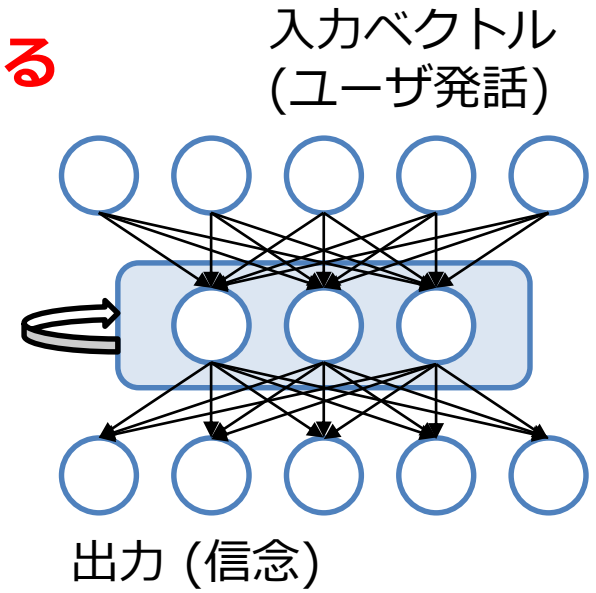
$$P(o|s) = \sum_h P(o,h|s) \approx \sum_h \underbrace{P(o|h)}_{\text{意図理解の識別確率}} \underbrace{P(h|u)}_{\text{音声認識の尤度}}$$

対話における前後の文脈の依存

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

  観測確率　状態遷移確率　現在の信念

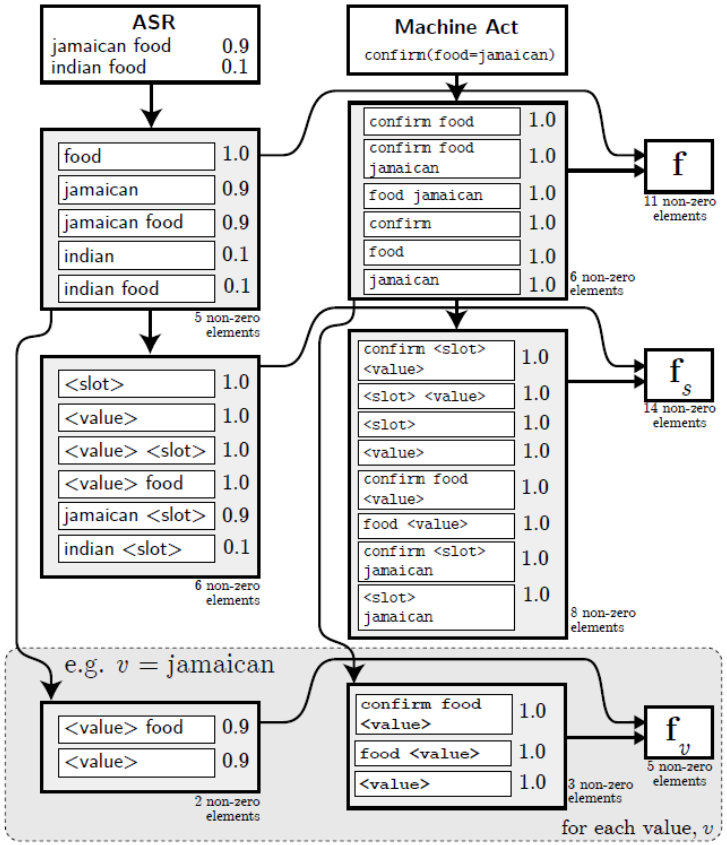
  - $s \in I_s$　ユーザ状態
  - $a \in K$　システムの行動
  - $o \in I_s$　観測状態
  - $b_s = P(s|o^{1:t})$　ユーザ状態が s である信念 (確率変数)

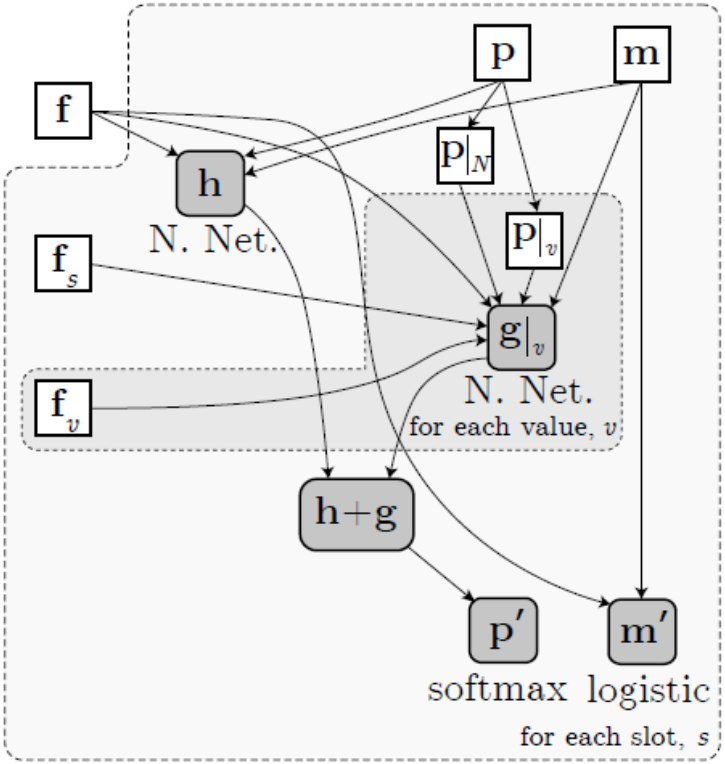


- Recurrent Neural Network と相性がよい！

Recurrent Neural Networkを用いた言語理解

- Word-Based Dialog State Tracking with Recurrent Neural Networks. Henderson et al., In Proc. SIGDIAL, pp, 292-300, 2014.





図は論文より引用

Recurrent Neural Network → Long Short Term Memory Neural Network

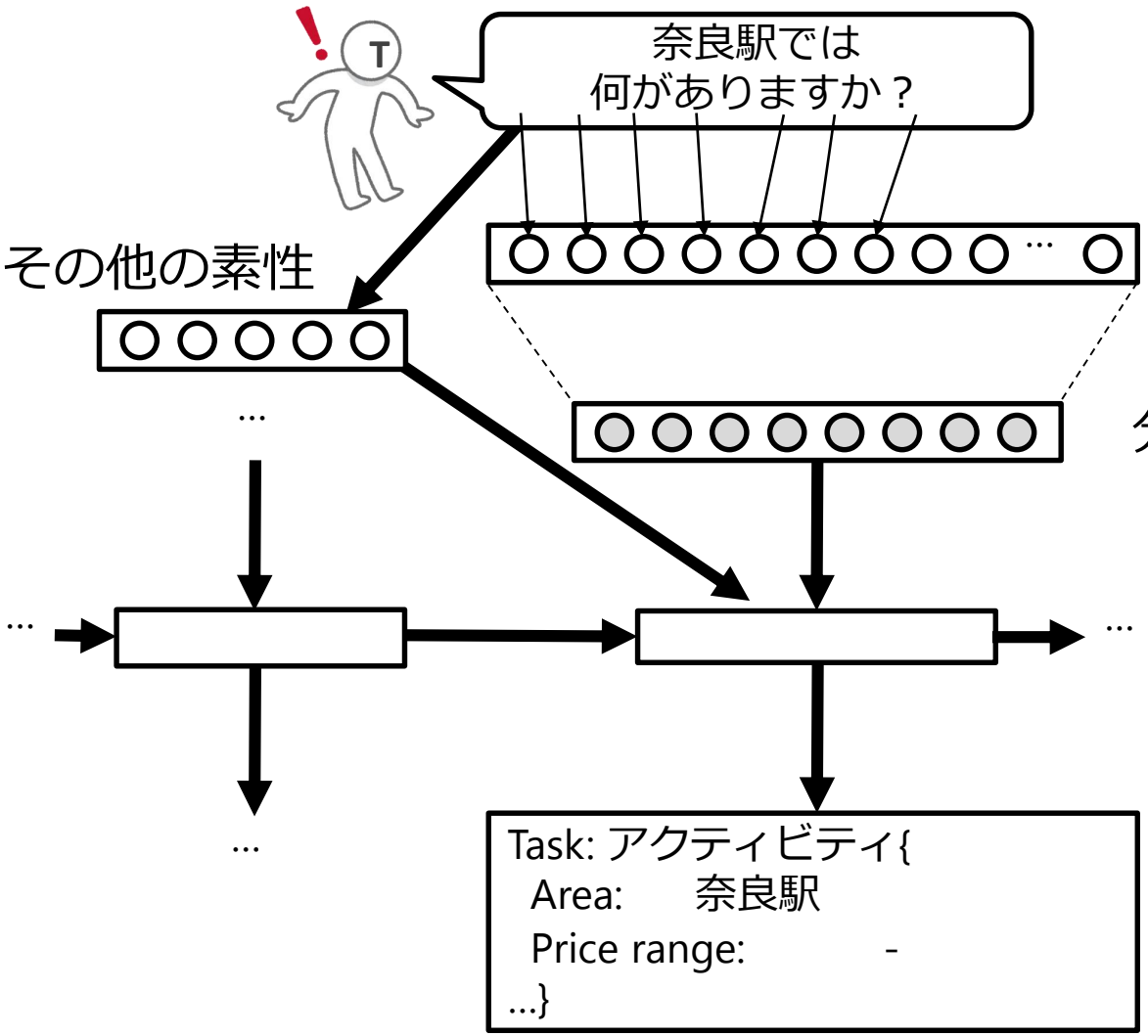
- LSTMは（大まかに言うと）より距離が離れた系列情報を保持可能なRNN

- 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.

Long Short Term Memory Neural Networkを用いた言語理解



ユーザ発話

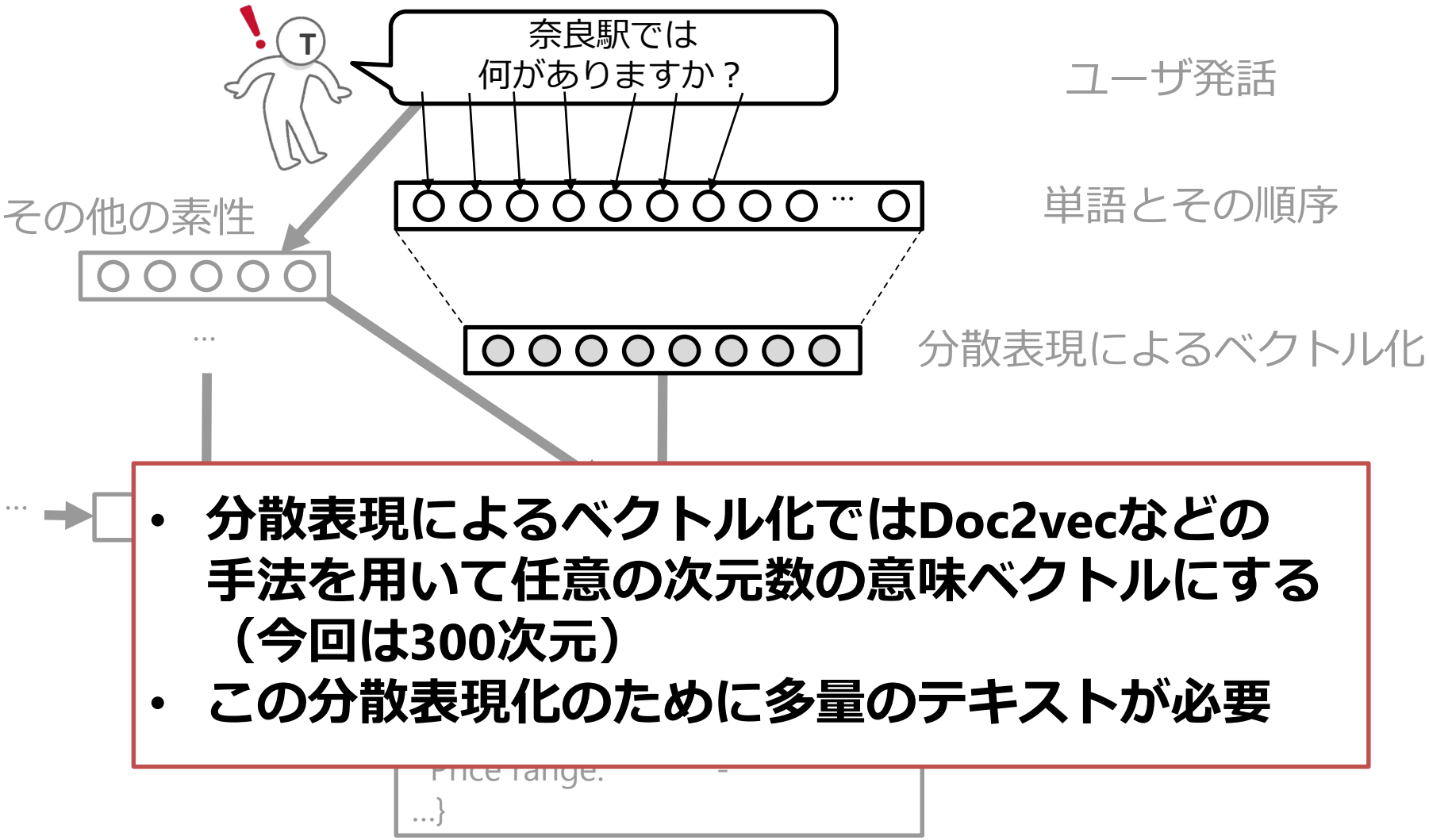
単語とその順序

分散表現によるベクトル化

LSTM

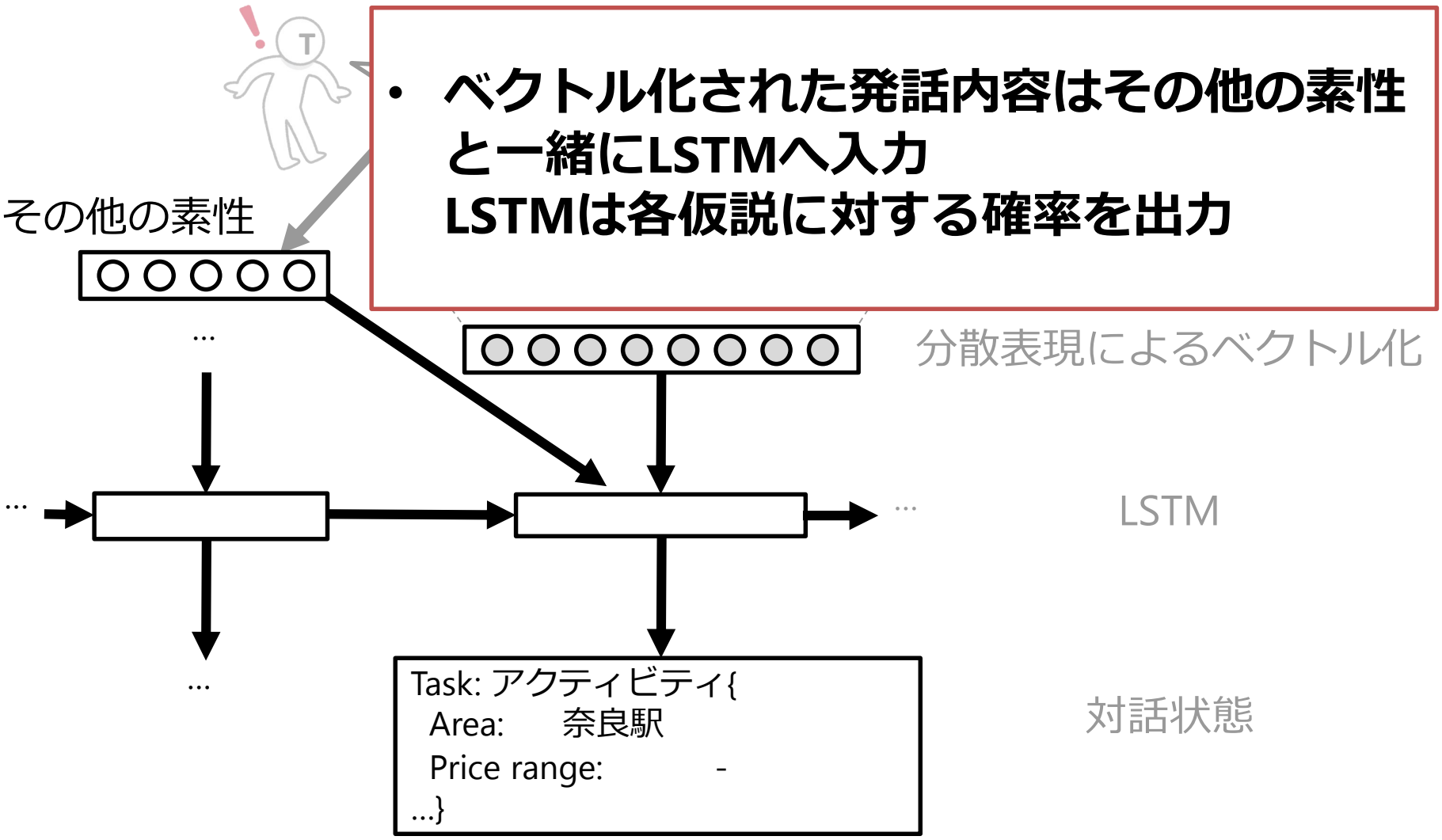
対話状態

Long Short Term Memory Neural Networkを用いた言語理解



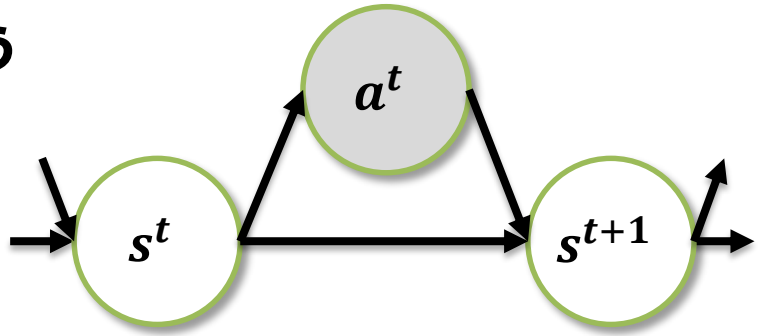
- **分散表現によるベクトル化ではDoc2vecなどの手法を用いて任意の次元数の意味ベクトルにする（今回は300次元）**
- **この分散表現化のために多量のテキストが必要**

Long Short Term Memory Neural Networkを用いた言語理解



言語理解結果に対する行動選択

- **s^t: ターン t のユーザの行動**
  - 具体的な行動:　Select $FROM, Select $TO_GO …
  - 対話の履歴:　$FROM=神保町駅, $LINE=半蔵門線
- **a^t: ターン t のシステムの行動**
  - 次の行動:　Ask $TO_GO, Ask $LINE, Confirm …
- **ユーザの行動は $P(s^{t+1}|s^t, a^t)$ に従う（マルコフ性があると仮定する）**
  - **強化学習で解ける**



強化学習を用いた対話制御

- $s \in I_s$ ユーザ状態
- $a \in K$ システムの行動
- $R(s,a)$ 報酬関数 タスク達成時に報酬を与える
- $\pi(s) = a$ 政策関数 これを効率よく学習したい
- ε 学習率
- γ 忘却率

- 価値関数 $V^{\pi}(s) = \sum_{k=0}^{\infty} \gamma^k R(s^{t+k}, a^{t+k})$ を最大化する政策関数の選択
- Q学習では以下の式で政策関数を学習する
  - $Q(s^t, a^t)$

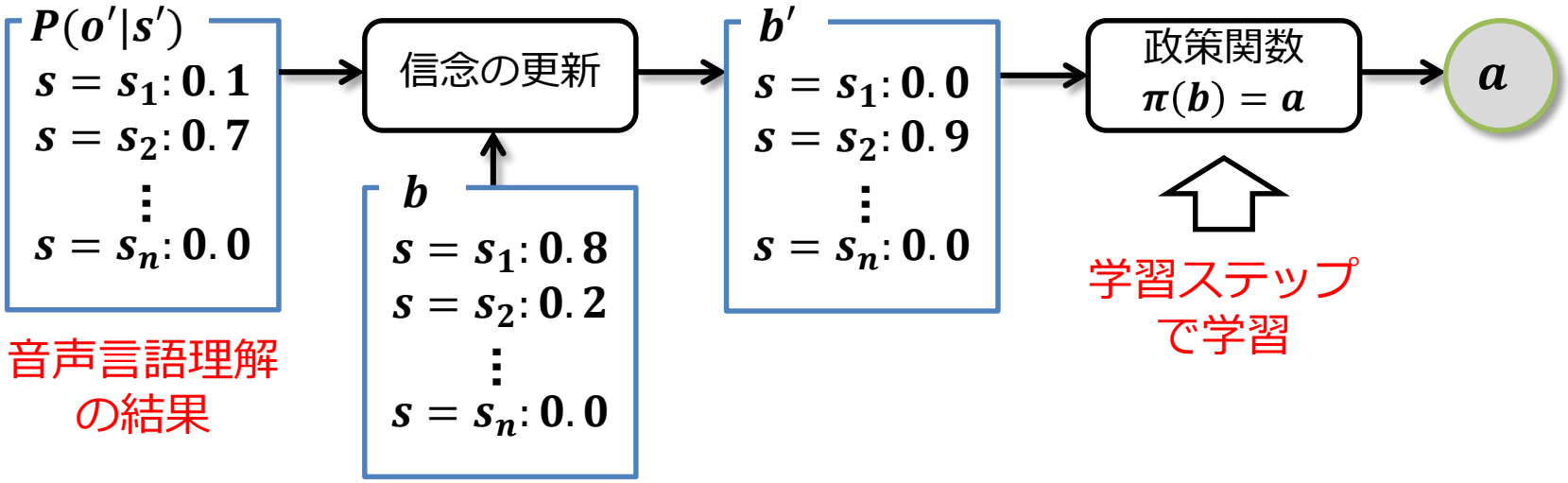
$$\xleftarrow[update]{} (1-\varepsilon)Q(s^t, a^t) + \varepsilon\left(R(s^t, a^t) + \gamma \max_{a^{t+1}} Q(s^{t+1}, a^{t+1})\right)$$

曖昧な言語理解結果に対する行動選択

- **いずれの手法も言語理解結果は確率変数として与えられる**
  - アプリケーションは入力に対して行動選択が必要
  - 与えられるのは $\boldsymbol{s}$ ではなく $\boldsymbol{b_s}$
- Partially Observable Markov Decision Process（部分観測マルコフ決定過程）による行動選択

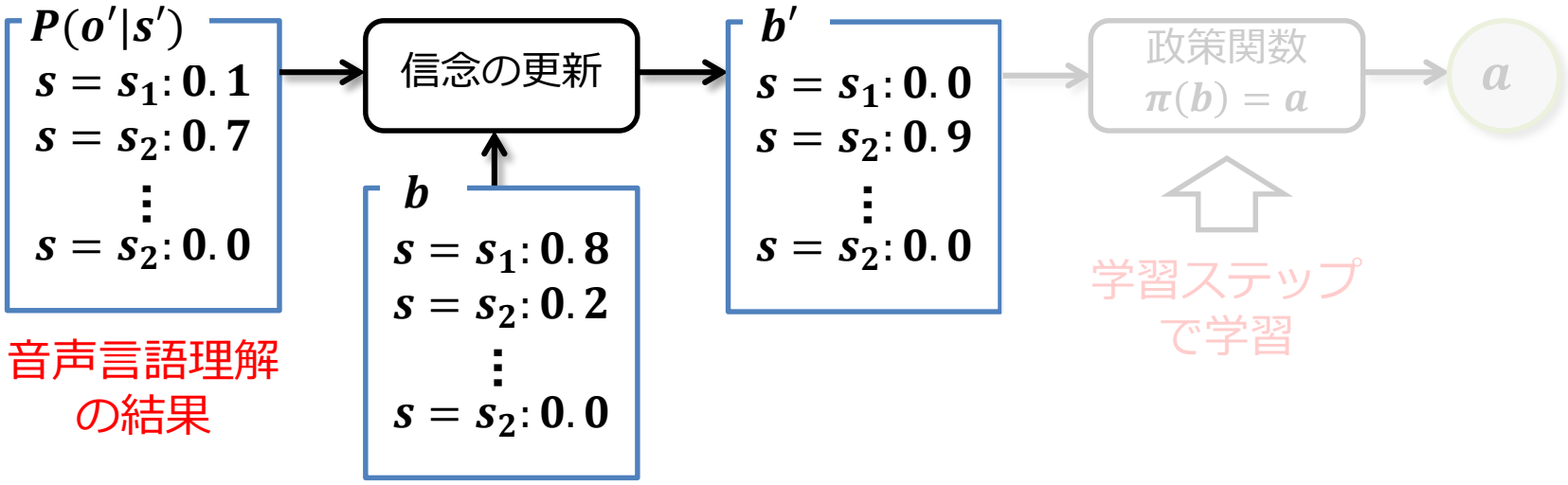
- **部分観測下で最適となる政策 $\boldsymbol{\pi^*(b) = a}$ を学習したい**
  - **対話研究の大きな問題の１つ**
  - **学習に使える対話データの量は限られている**

POMDPによる対話制御



- $s \in I_s$ **ユーザ状態**
- $a \in K$ **システムの行動**
- $o \in I_s$ **観測状態**
- $b_i = P(s_i|o^{1:t})$ $s = s_i$ である信念 (確率変数)
- $R(s, a)$ **報酬関数**
- $\pi(b) = a$ **政策関数**

POMDPの更新

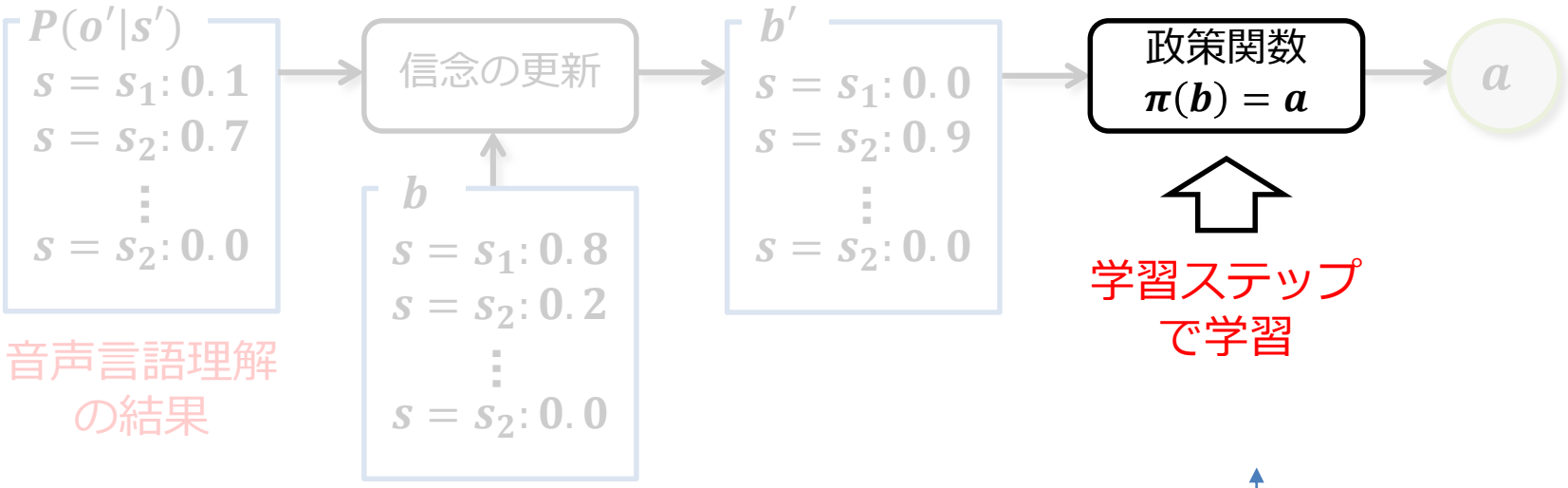


- $b' = P\left(s^{t+1} \middle| o^{1:t+1}\right) \propto P(o'|s'_j) \sum_{s_i} P\left(s'_j \middle| s_i, \widehat{a_k}\right) b^t$

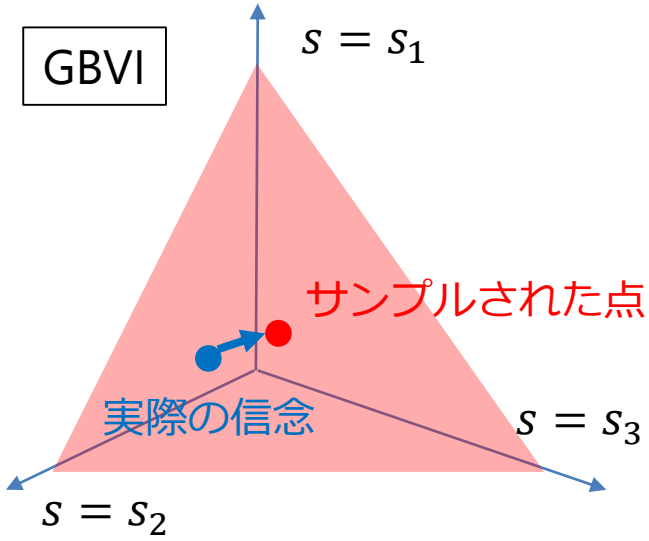
観測確率　状態遷移確率　現在の信念

- 信念を更新
  - 次の行動を出力する政策関数の入力

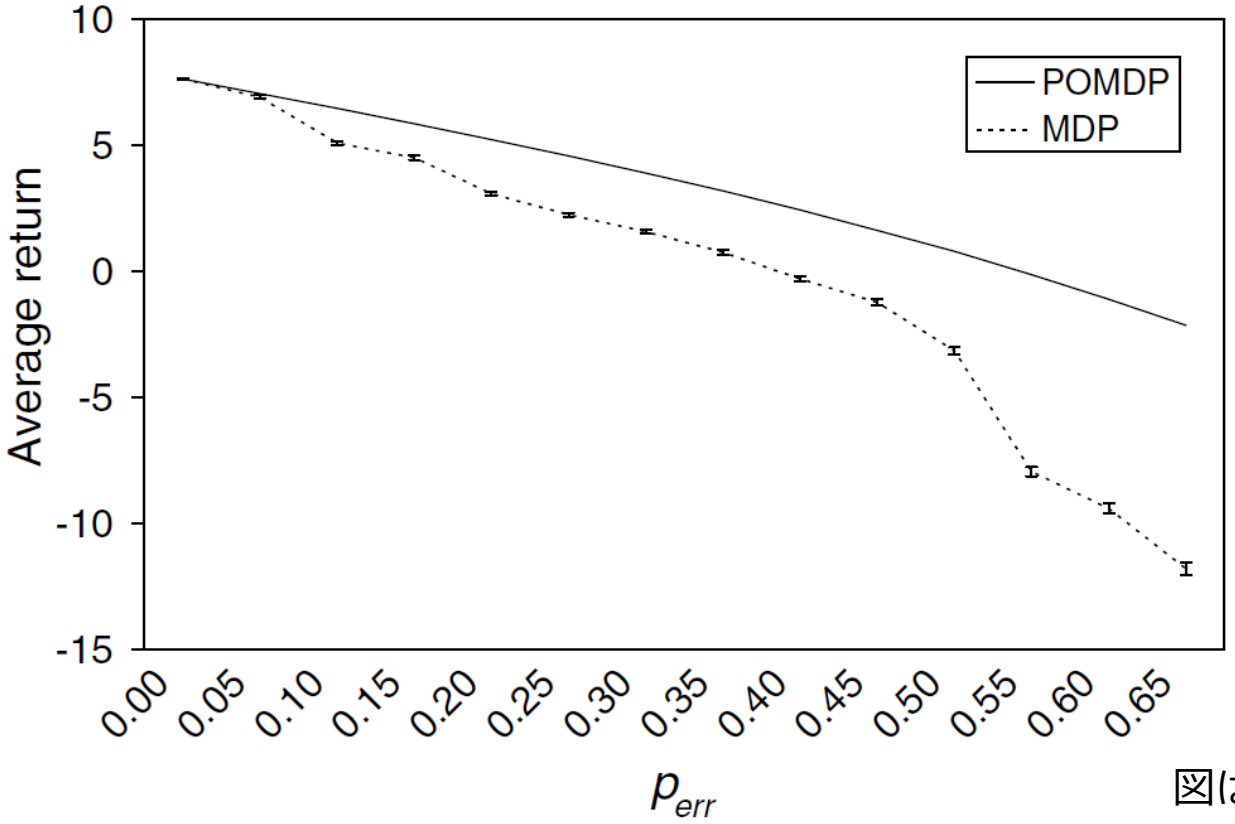
（古典的な） POMDPの学習



- 信念空間上でサンプリングされた任意の点にマッピング
- マッピングされた点においてシミュレータとの学習で得られた政策関数で行動を決定（$Q(b, a)$を最大化）



MDP → POMDPの効果



図は論文より引用

- POMDPの方がエラーが多い場合でも頑健に動作
  - Partially observable Markov decision processes for spoken dialog systems. Williams et al., Computer Speech & Language, 393—422, Vol.22, No.1, 2007.

対話システムにおけるPOMDPの問題

- **最適な $\pi^*()$ を得られるほど学習データがない**
  - 効率的な学習手法が必要

1. **ルールとPOMDPの併用**
2. **効率的なサンプリング**
3. **効率的なQ関数の計算**

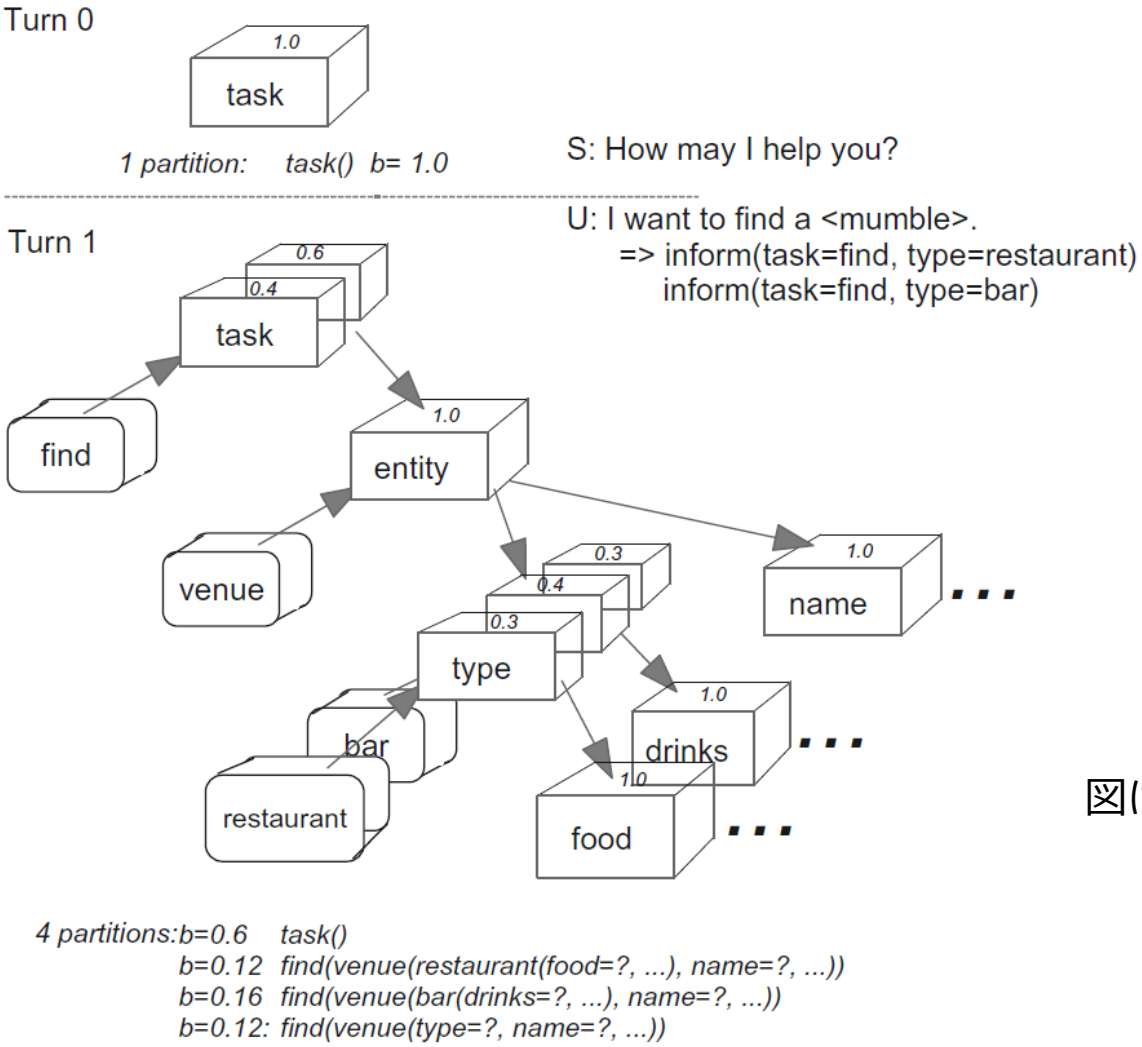
ルールとPOMDPの併用

- 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.

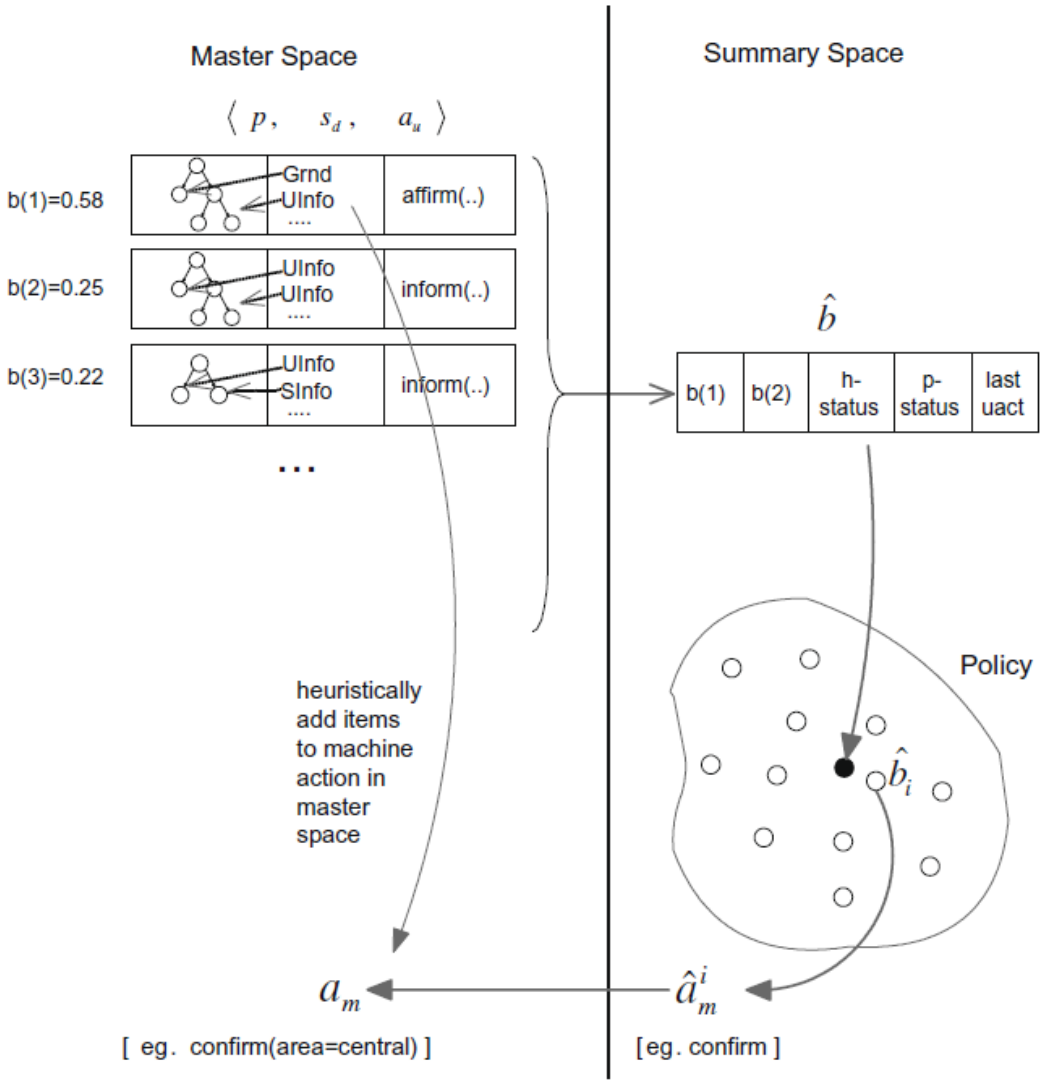
- **人手で与えたルールを探索空間の制約とする**

Hidden Information State Model



図は論文より引用

Hidden Information State Model



図は論文より引用

Intention Dependency Graph

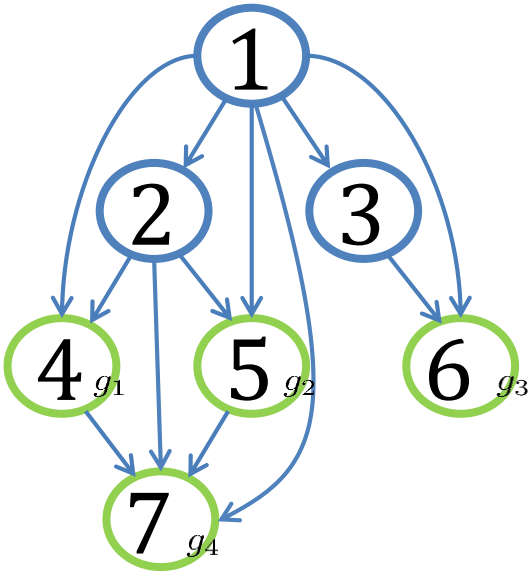
- **あらかじめ定義されたタスク構造間の遷移確率を定義**

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]

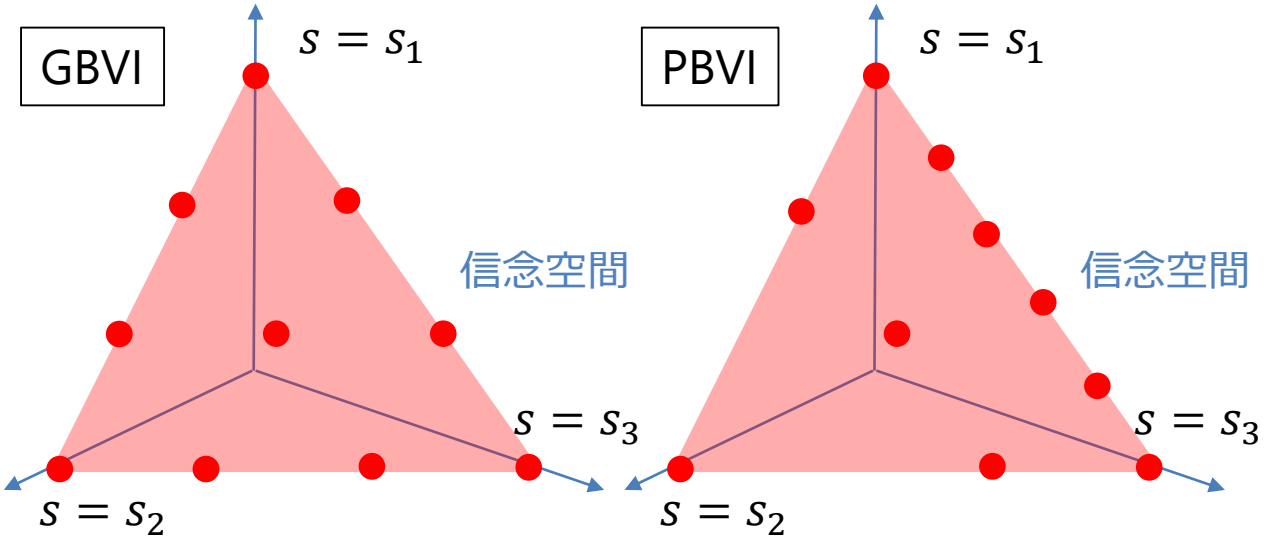


ルールやタスク構造を併用するメリット

- **新ドメイン・システムのローンチ**
  - まずはルールベースでデータを集める
  - スムーズに統計ベースにシフトできる
- **未観測の状態・系列に対して重み付け可能**
  - 全ての状況をカバーする対話データを学習用に用意することは困難
- **未観測の状態・新しいドメインへの適応は大きな課題**

効率的なサンプリング

- 均等に信念空間をサンプルするのは非効率



- GBVI: 均等なグリッドに沿って Belief point を選択
- PBVI: 実際の分布にあわせて Belief point に偏りを持たせる
  - 例では s_1 と s_3 がよく間違われやすい状態

効率的なQ関数の計算

- POMDPの学習は $\boldsymbol{Q}(\boldsymbol{b}, \boldsymbol{a})$ の最大化問題
  - あらゆる $\boldsymbol{b}, \boldsymbol{a}$ に対しQ値が計算できればよい

- $\boldsymbol{Q}(\boldsymbol{b_i}, \boldsymbol{a_i})$ が既知の場合、 $\boldsymbol{Q}(\boldsymbol{b_k}, \boldsymbol{a_k})$ を求めるためには $(\boldsymbol{b_i}, \boldsymbol{a_i})$ と $(\boldsymbol{b_k}, \boldsymbol{a_k})$ の類似度を用いればよい
  - この類似度を求めるカーネルを学習する

- 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.

一問一答の質問応答

- 2000年頃から盛んに行われていた研究タスク
  - NTCIRなど
- **IBM Watson の登場**
  - 質の良い質問・応答ペアを集めるというアプローチ
  - 近年では推論の研究が盛ん

- **これまでの多くの音声対話システムは**
  - **特定のタスク達成対話システム**
  - **一問一答の質問応答**
  - **Webなどを用いた検索**

  などの組み合わせで構成されている

タスク指向対話システムの構築に重要な点

- **ユーザが明確な意図発信をできるタスクデザイン**
  - ユーザが何を言っていいかわからない状態を防ぐ
    - ファーストフード店の優秀な店員を目指す
- **必要十分なタスク構造**
  - 大ざっぱすぎると何も出来ない
  - 細かすぎると制御の学習がうまくいかない
- **対話が失敗したときのフォールバック**
  - Webで調べるなどで何もできない印象を軽減

非タスク指向におけるシステム構築

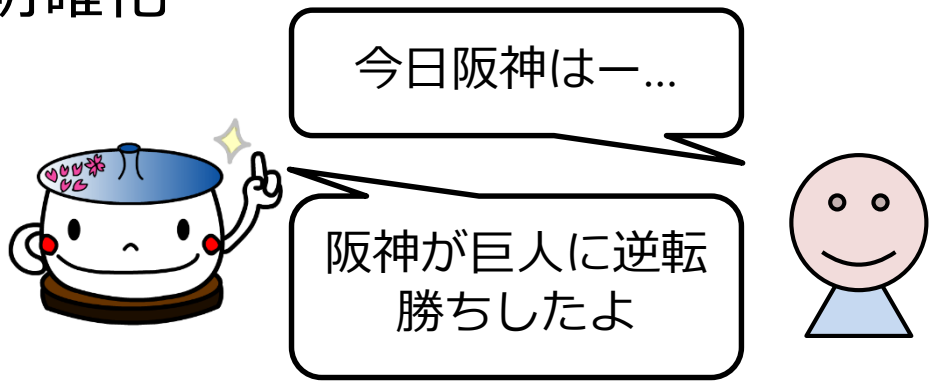
- Conversational System for Information Navigation based on POMDP with User Focus Tracking. Yoshino et al., Computer Speech & Language, Vol.34, Issue.1, pp.275--291, 2015.

- **ユーザの意図が曖昧でも動作するデザイン**
  - システムからの能動的な働きかけ
  - 対話が失敗したときのフォールバック
- **ユーザの意図を大まかに抽象化**
  - 対話のシチュエーションを限定
- **システムが行いたい行動に合わせた観測状態の導入**
  - 情報案内システムでは話題の注目状態に相当する焦点を導入

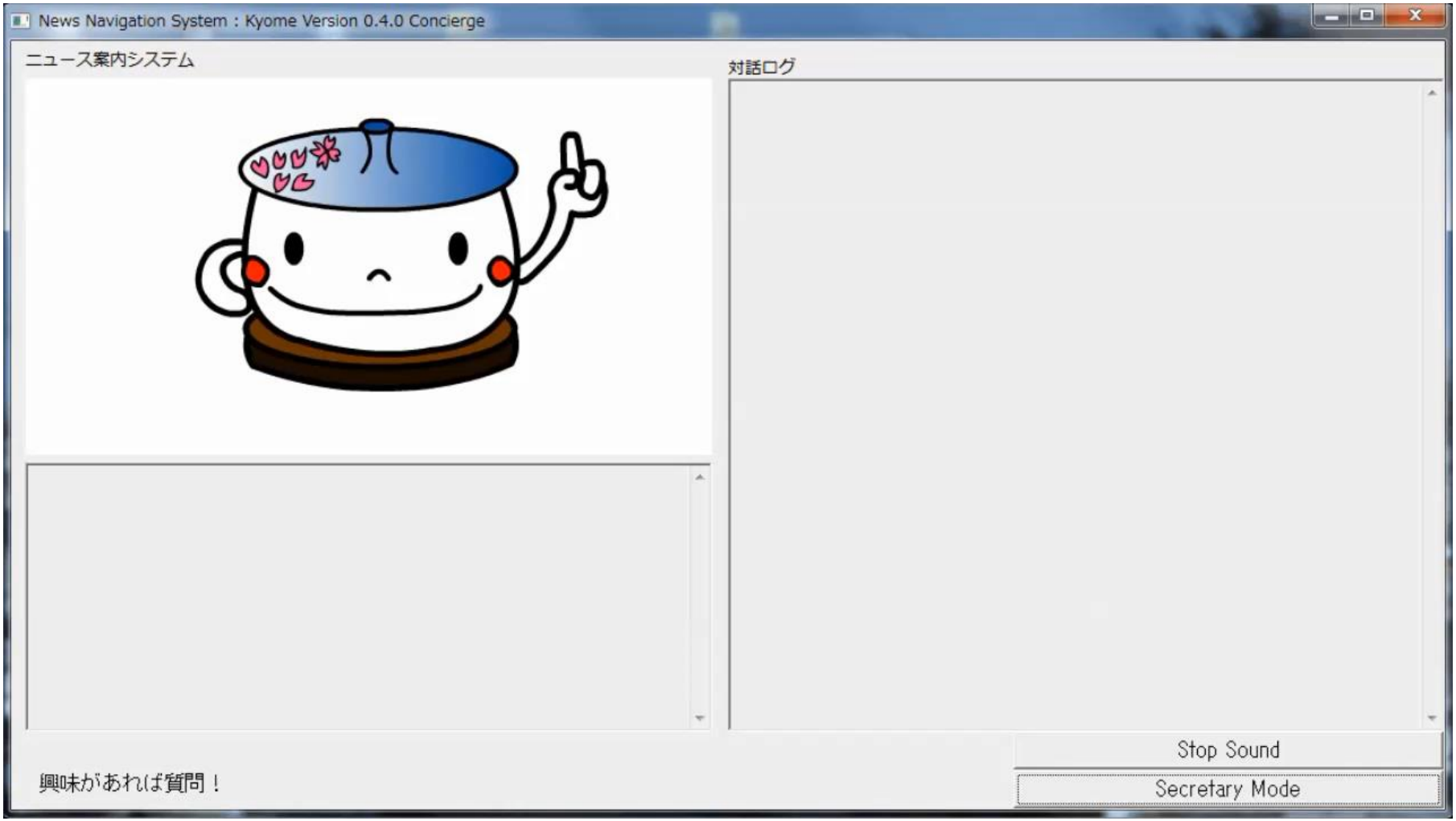
情報案内対話システム

- **知識ベース（文書）に記述された内容を案内するタスク**
  - 日々動的に更新されるニュース記事
  - ドメインを規定
    - 自動抽出したドメイン知識の利用

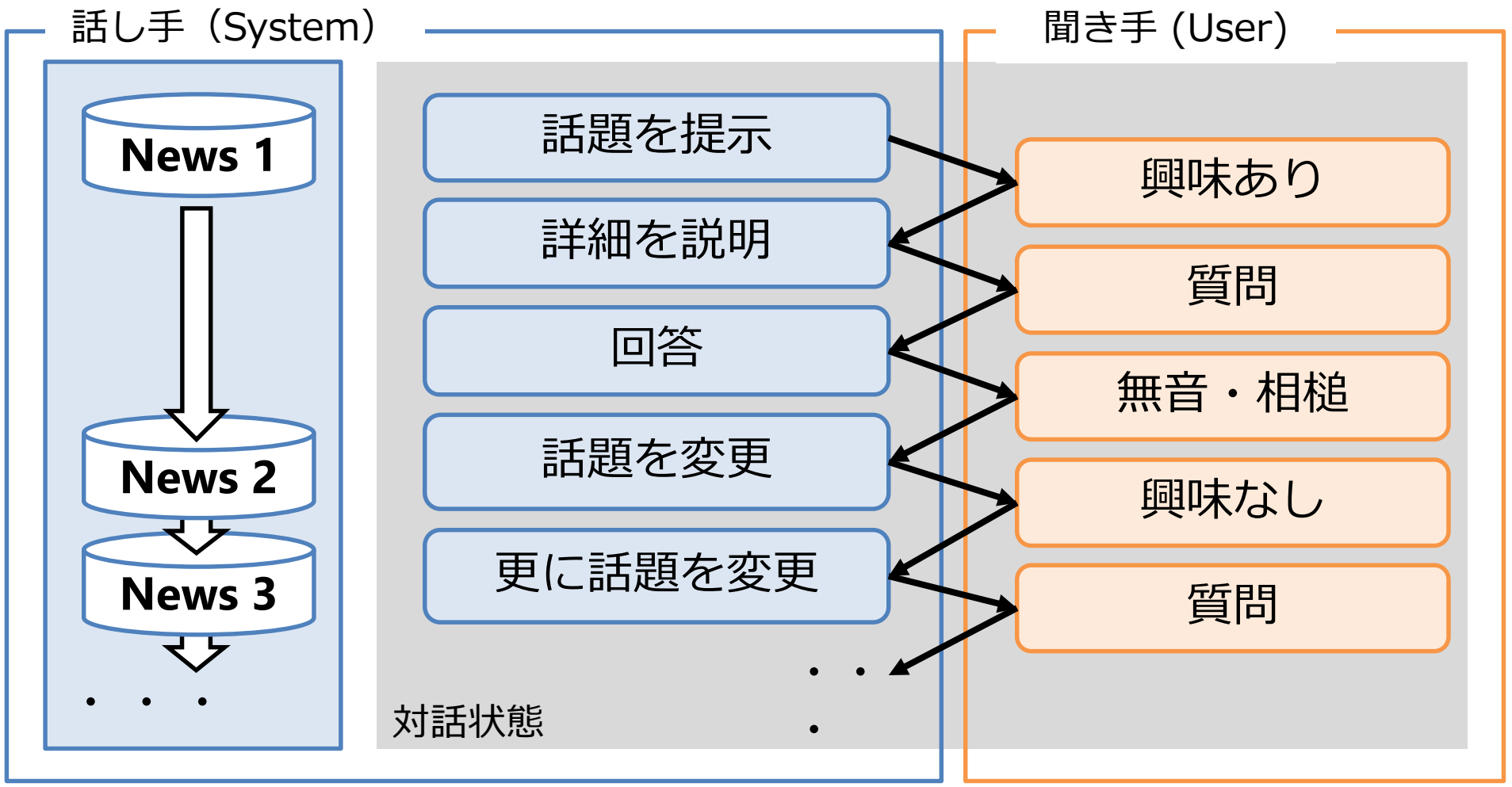
- **話し手（システム）が順番にトピックを提示**
  - 聞き手の聞きたいことを明確化
    - ドメイン知識
    - ユーザの意図
    - ユーザの焦点



情報案内システム


News Navigation System : Kyome Version 0.4.0 Concierge
ニュース案内システム
対話ログ
興味があれば質問！
Stop Sound
Secretary Mode

情報案内対話の構造


話し手（System）
聞き手 (User)
News 1
News 2
News 3
話題を提示
詳細を説明
回答
話題を変更
更に話題を変更
興味あり
質問
無音・相槌
興味なし
質問
対話状態

情報案内対話における対話の機能

- **各ターンのユーザ意図に対する最適なモジュールの選択**

- ユーザ意図 $\boldsymbol{s}$:
  - ***TP***: トピックの紹介要求
  - ***ST***: 詳細の説明要求
  - ***QA***: 質問
  - ***GR***: 挨拶
  - ***II***: 音声認識誤りに起因する無効入力
  - ***NR***: 一定時間の無音

システムの行動 $\boldsymbol{a}$:

TP: トピックの紹介

ST: トピックの詳細説明

QA: 質問に対する回答

GR: 挨拶

KS: 無音（反応なし）

CO: 意図の確認

PP: プロアクティブな情報推薦

対話における焦点

Example 1

…
Usr: **田中は** どこで練習したの？
Sys: 田中は20日、ヤンキースのキャンプでブルペンに入り投球練習をを行ったよ。
Usr: (無音)
Sys: ところで、田中は2月18日にも落ちるツーシームを練習したよ。
…

Example 2

…
Usr: 何かあったんですか？
Sys: 宮崎でゴジラ弾が復活したよ。
Usr: (無音)
Sys: スーパーエースへ。藤浪晋太郎はルーキーイヤーから進化しているのか
…

- 例1: ユーザが焦点を持っているのでシステムは話題を継続
- 例2: ユーザに焦点がないのでシステムは次の話題を提示

適応的な対話制御に発話内での焦点の有無の導入が効果的

焦点は「ユーザへの情報案内に不可欠な対象」として定義

音声言語理解の評価

焦点解析

問題	精度
文節ごとの解析精度	78.5%
発話中に焦点があるか(0 or 1)	99.9%

意図理解

素性	タグ	再現率
	TP	98.7%
	ST	81.1%
	QA	95.1%
	GR	97.7%
	II	31.3%
	All	**93.6%**

- 事前に収集した18話者918発話を書き起こし・アノテーション
  - **ユーザの意図**: どのモジュールがユーザへの応答に最適か
  - **焦点**: 発話で最も情報案内に不可欠な対象 (最大1個)
  - ５分割交差検定
- $\boldsymbol{P(o|h)}$（前後のターンを考慮しない場合）

焦点を用いた対話制御の拡張

- 発話に焦点が存在するかのブール値 $f = 0\ or\ 1$ を導入
  - $b' \propto P(o'_{s'}, o'_{f'} | s'_j, f'_m) \sum_i \sum_l P(s'_j, f'_m | s_i, f_l, \widehat{a_k}) b^t_{s_i, f_l}$
- 観測モデル（独立を仮定）
  - $P\left(o'_{s'}, o'_{f'} \middle| s'_j, f'_m\right) \approx P\left(o^{t+1}_s \middle| s'_j\right) P\left(o^{t+1}_f \middle| f'_m\right)$
- 遷移モデル（焦点が定まってからユーザ状態が定まる）
  - $P\left(s'_j, f'_m \middle| s_i, f_l, \widehat{a_k}\right) = P\left(f'_m | f_l, s_i, \widehat{a_k}\right) P\left(s'_j \middle| f'_m, f_l, s_i, \widehat{a_k}\right)$
- 政策関数
  - $\widehat{a} = \pi^*\left(b_{s,f}\right)$

情報案内システムの評価

- **評価指標**
  - DST（ユーザの意図の追従精度）
  - ACT（システムの行動の選択精度）

- **評価データ**
  - 12ユーザ24対話626発話の実ユーザとの対話を収集
  - アノテータ２名によって各発話のユーザの意図($\boldsymbol{s}$)
    ・対応するシステムの行動($\boldsymbol{a}$)をアノテーション
  - アノテーション一致率
    - $\boldsymbol{s}$: 0.958 (kappa=0.938)
    - $\boldsymbol{a}$: 0.944 (kappa=0.915)

情報案内システムの評価

	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)

- **焦点の導入によりユーザ意図の追従精度が向上**
- **焦点の導入により行動選択の精度が向上**
- **提案法はユーザ焦点に応じた情報の推薦が可能**
  - 35回推薦を行い、17回でユーザのさらなる興味を誘発

非タスク指向型システムの構築

- **ユーザがシステムと対話する場面の想定**
  - 場面を明確化
    →システムがとる行動・観測するユーザ情報を明確化

- **必要な会話の粒度に合わせた抽象化・階層化**
  - 使う手法に応じた抽象化が必要
  - 無理に統計的手法を使う必要はない

発話生成（文生成）

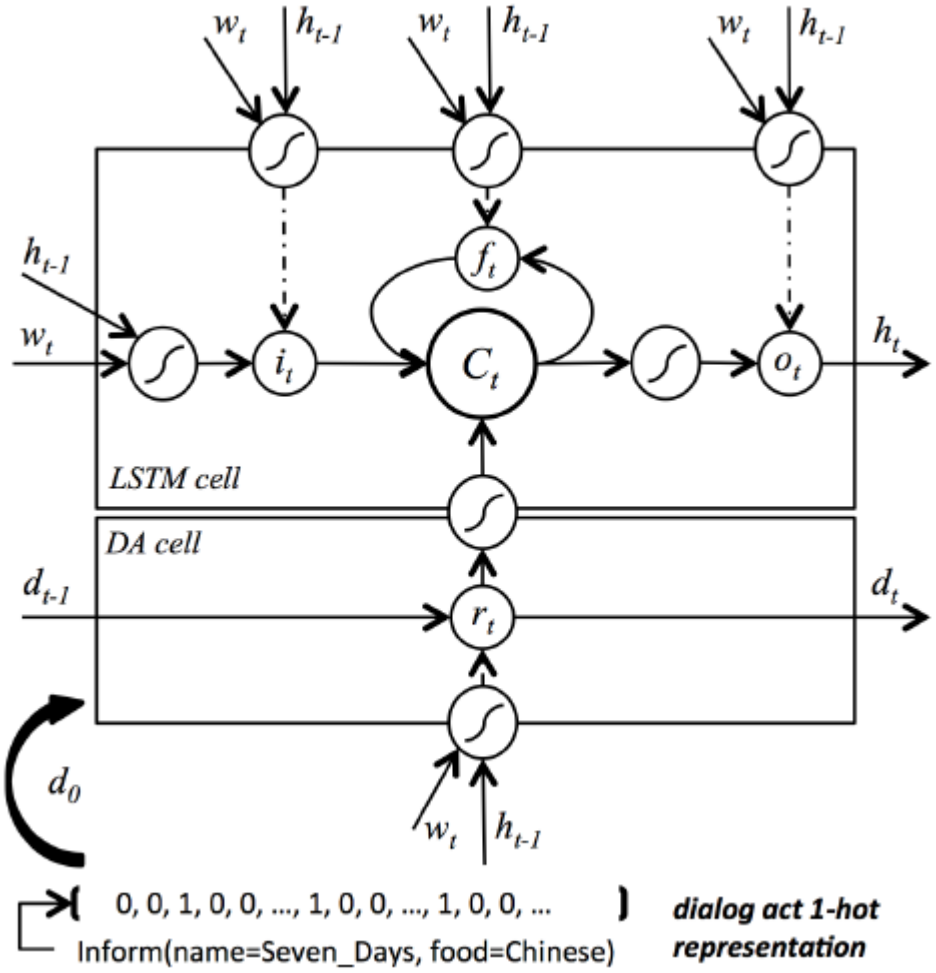
- Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. Wen et al., In Proc. EMNLP, 2015.

- **これまでの対話システムにおける文生成の問題**
  - **ルールやテンプレートを利用**
    - 表現のバリエーションを生み出すことが難しい
    - 異なるドメインに移行するのが大変
  - **統計ベースの制御の難しさ**
    - 上記の問題を解決するが適切な文を生成しないことがある
  - **適切さ、自然さ、理解しやすさ、バリエーションが重要**
    - 同時に満たすのが難しい

LSTMを用いた発話生成

recurrent hidden layer

embedding of a word



1-hot dialog act and slot values

上のセルは言語モデルに相当

下のセルは「言うべきこと」を満たしているかに対応

図は論文から引用

End-to-endの音声言語処理

- **言語理解・対話制御を行わない**
  - 入力発話から直接出力発話を推定する
- **用例対話システム**
  - ユーザ発話とシステムの応答の対を大量に用意
  - 用意された用例のどれに一番一致するか
- Adaptive selection from multiple response candidates in example-based dialogue. Mizukami et al., In Proc. ASRU, 2015.
  - 用例の良さをユーザ満足度で定量化・好みを学習
- End-to-end memory networks. Sukhbaatar et al., In Proc. NIPS, 2015.
  - Neural Network (LSTM)を用いて入力から直接出力発話を推定

フロントエンド出力層に統計的手法を配置するリスク

- **Neural Networkなどの統計的手法は制御が難しい**
  - 出力してはいけない文
  - 文法的には正しいが意味的に正しくない文
- **これらの問題をフィルタする機構が必要**
  - 用例作成の段階でのフィルタ
  - 意味的な正しさを向上させる機構
- **構築する際に手法に対する理解が必要**
  - ブラックボックスのままでは難しい

オープンドメインシステムの構築

- **チャット型対話システムのオープンドメイン化**
  - 用例ベースなどのシステムは拡張が容易
  - 多様なドメインに対応可能

- **タスク対話システムもドメイン適応の研究は盛ん**
  - Policy committee for adaptation in multi-domain spoken dialogue systems. Gasic et al., In Proc. ASRU, 2015.
  - 複数のドメインシステムを構築し「どのドメインについて話すのが適切か判定する」
  - 異なるドメインの対話データも学習に利用可能

その他の音声言語処理の可能性

- **より自然な音声合成**
  - 現在の音声合成は既に非ネイティブ以上
- **講演に対する字幕付与（情報保障）**
  - SIG-AAC（情報処理学会アクセシビリティ研究会）
- **音声言語処理技術を用いた語学教育**
  - 音声認識結果と字幕を使ったリスニング教育
  - 非英語母語話者の音声認識を用いたスピーキング教育
  - 既に中国では国家プロジェクトとしてシステム開発を開始
- **音声言語処理技術を用いた教育**
  - 発達障害に対するコミュニケーショントレーニング
- **高齢者社会へ向けた見守りシステムの実現**

音声言語処理のこれから

- **様々なモダリティの併用**
  - 人間は音声以外にも視線・ジェスチャーなどを利用
- **より粒度の細かいターンテイキング**
  - 従来は「音声区間の検出＝相手のターン」
- **漸進的な処理**
  - リアルタイムコミュニケーション
- **意味の扱い**
  - 文意をどう扱っていくか

より自然な音声言語インタラクションを目指して

- **ERATO 石黒共生ヒューマンロボットインタラクションプロジェクト**

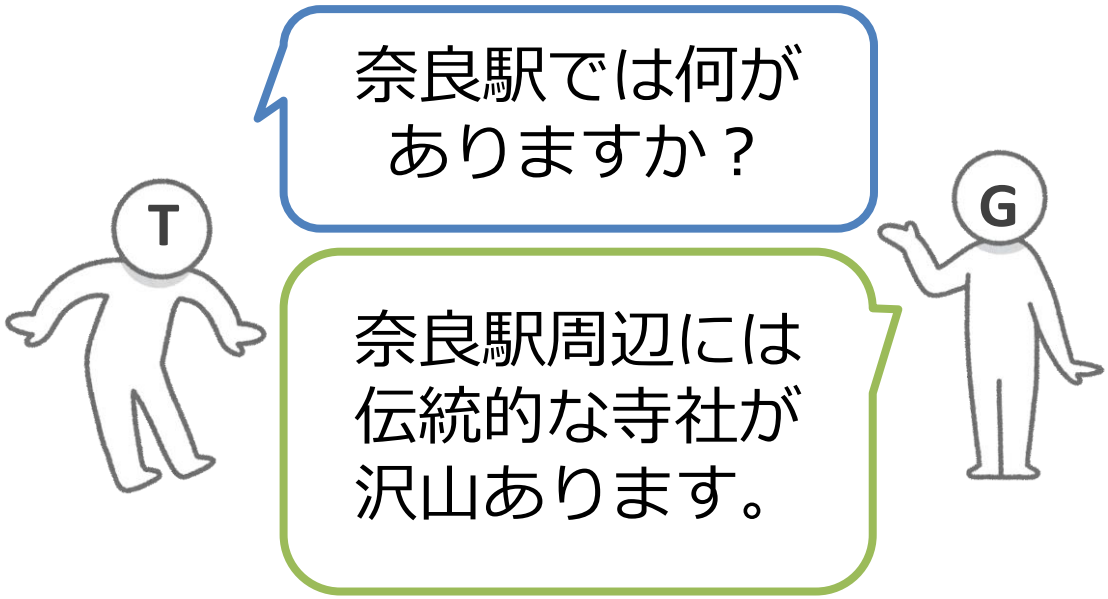
- **人間らしいコミュニケーション技術**
  - 自然な応答タイミング・ジェスチャー・相槌 etc…
- **遠隔音声認識**
  - 非接話・雑音環境下での音声認識
- **適切な意図理解**
  - タスク指向・非タスク指向

音声言語処理分野における国際コンペ

- 音声認識: CHiME challenge
  - 実環境下における音声認識精度の向上
- 対話状態推定: Dialogue State Tracking Challenge
  - タスク対話におけるユーザ発話意図の推定
- 言語処理: CoNLL Shared Task
  - 係り受け、項構造など言語処理の分野における重要なタスク
- その他機械翻訳・音声翻訳・音声認識など多数のコンペ
  - 配布されるデータに対する精度を競う

Dialogue State Tracking Challenge

- 発話の意図理解



意図

```
"frame_label": {
  “情報”:
    [”アクティビティ”],
  ”NEIGHBOURHOOD”:
    [“奈良駅”]
}
```

- 多様な研究機関が参加する国際コンペ
  - 過去に NAIST, Panasonic, MIT, XEROX, IIR, Microsoft, Cambridge などが参加

Take home messages

- 音声言語処理は実用のフェーズに入っている
- ただし現状は目的・手法をよく理解した専門家が必要
- これからを考える上で他分野との連携が重要
- Shared Task などを通じた資源の整備も重要