「産総研　人工知能セミナー　第４回」で検索
「音声対話システム　POMDP.NET」で検索

End-to-End時代における対話システムの研究動向

奈良先端科学技術大学院大学
助教　吉野 幸一郎

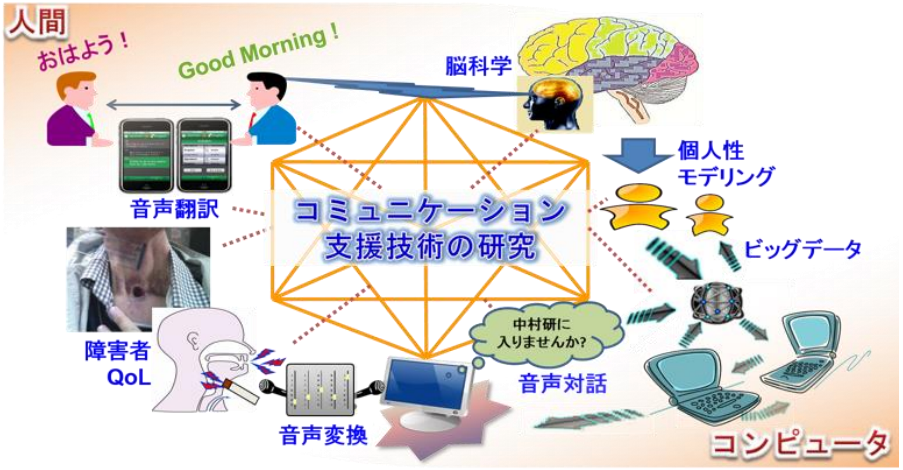
@caesar_wanya

Nakamura SRG
Augmented Human Communication (AHC) Laboratory

Nara Institute of Science and Technology
Augmented Human Communication Laboratory

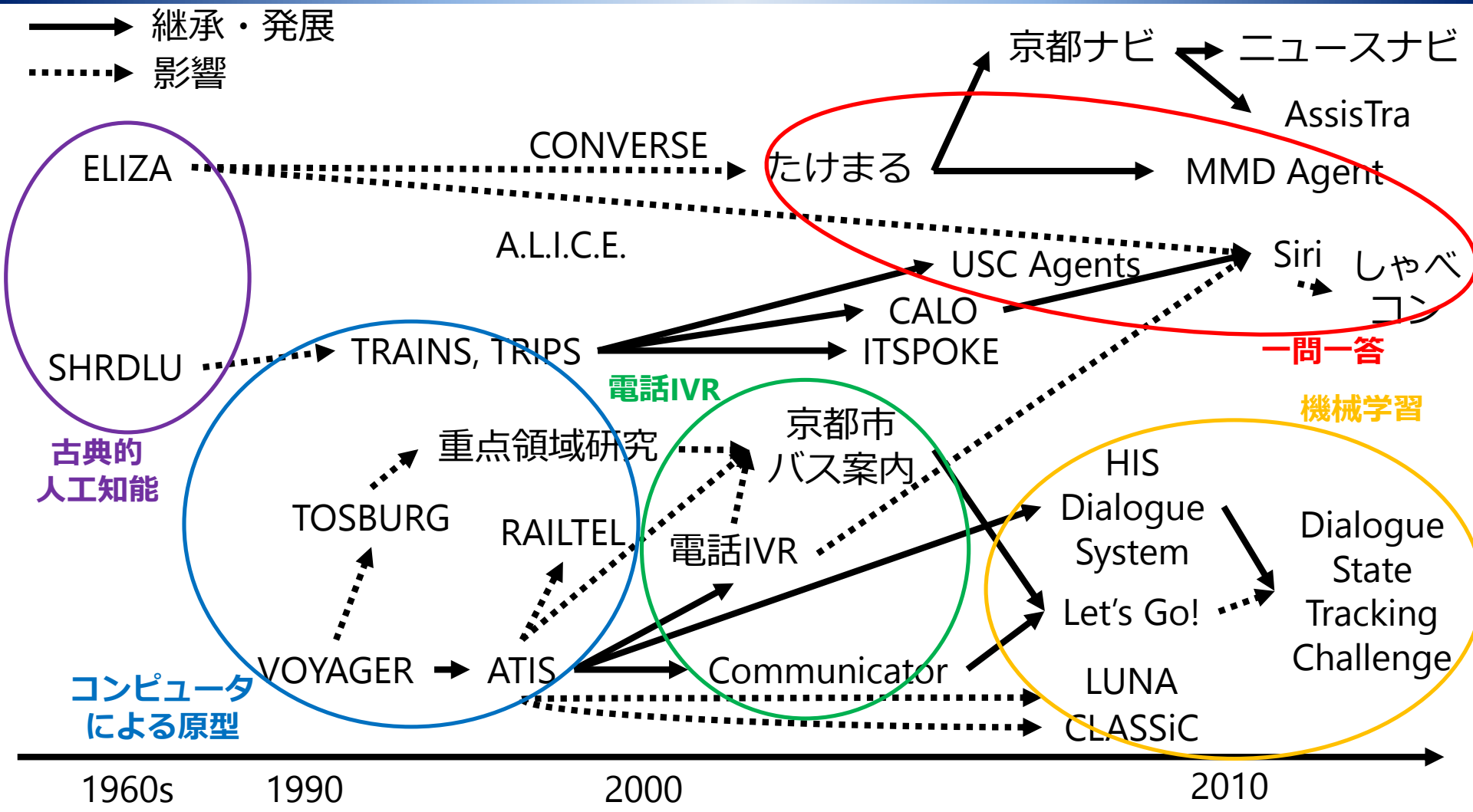
自己紹介

- **吉野 幸一郎**
  - 2005-2009 慶應SFC, 学部（石崎研）
  - 2009-2015 京大情報, 修士博士PD（河原研）
  - 2015- NAIST情報, JK（中村研）

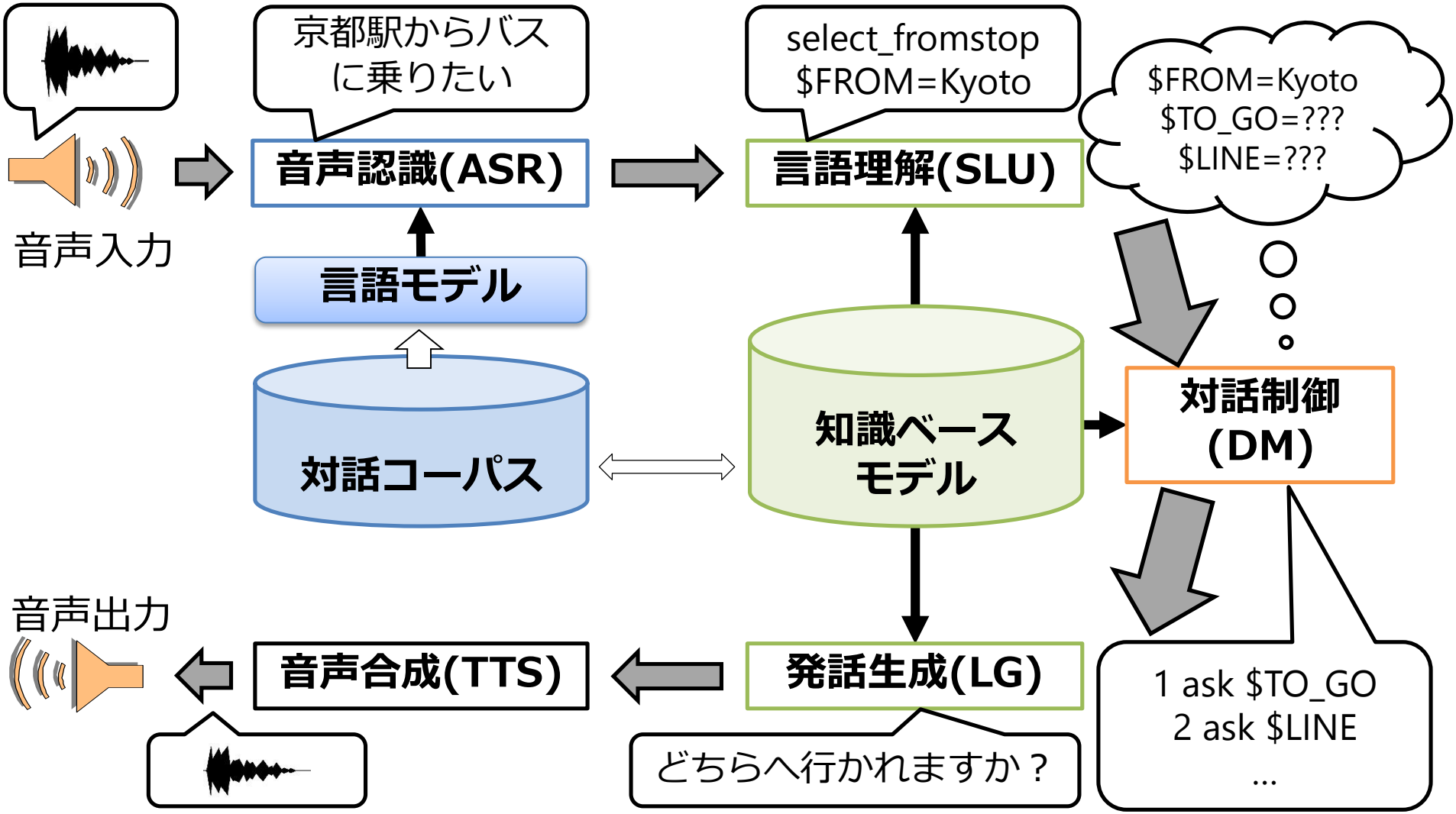


音声認識・対話・翻訳に興味のある方は是非NAIST中村研へ

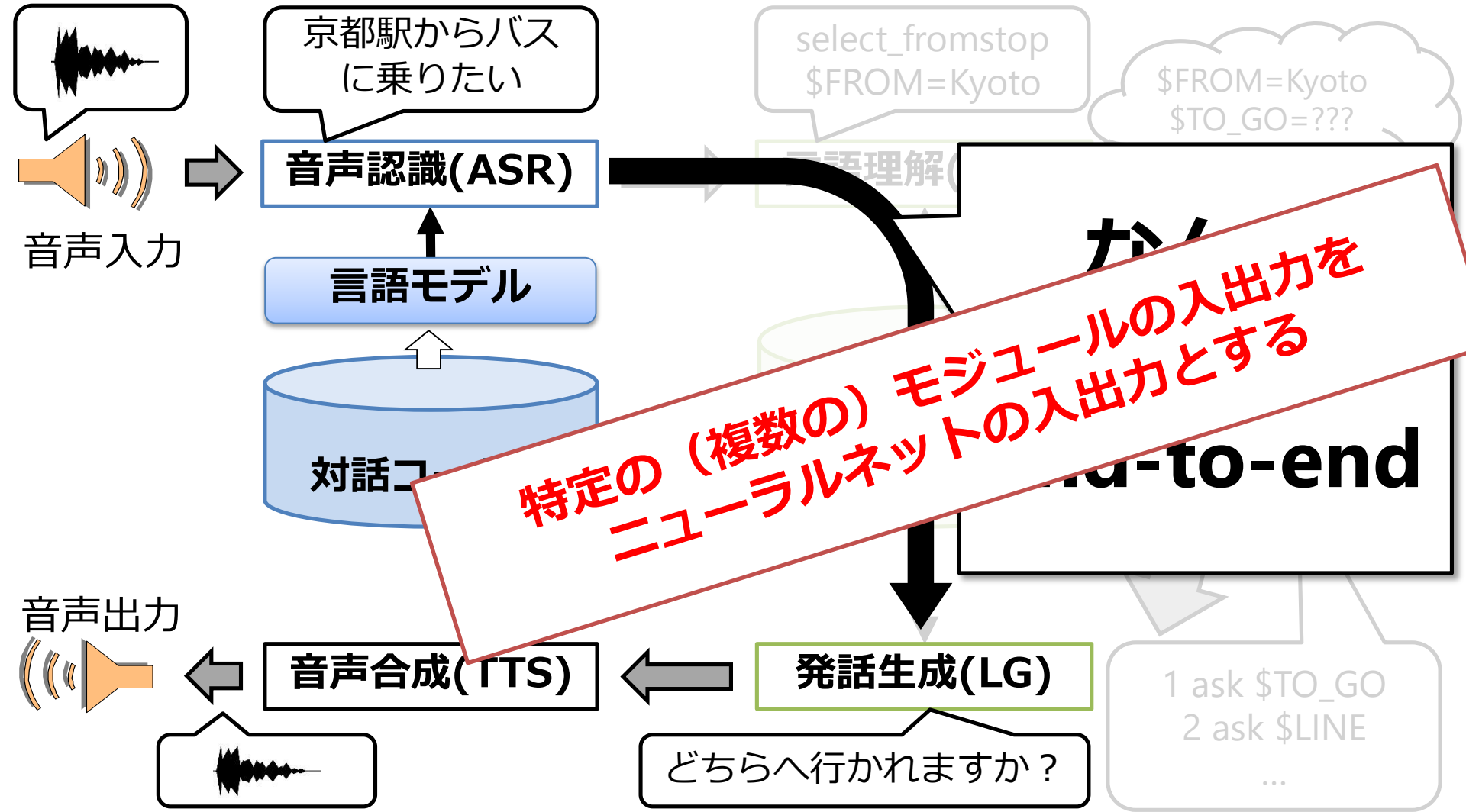
音声対話システムの系譜



音声対話システムの基本的枠組み

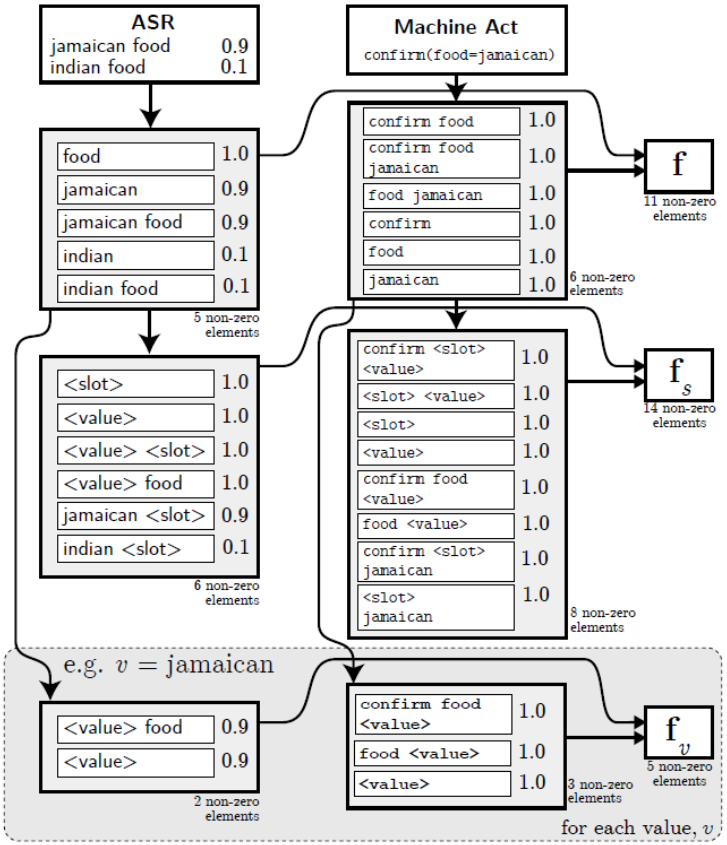


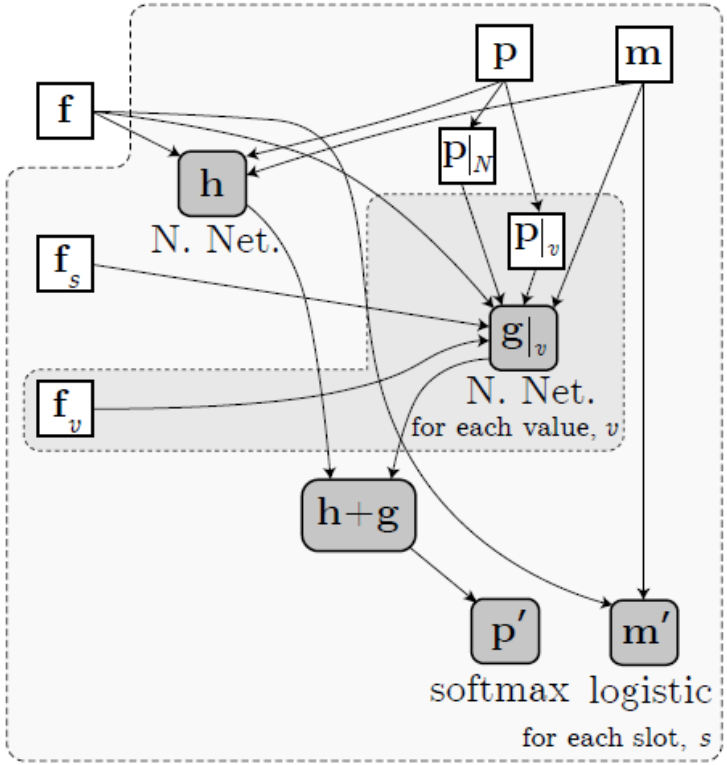
音声対話システムにおけるEnd-to-end



Dialogue State Tracking with RNN（入力文→対話状態）

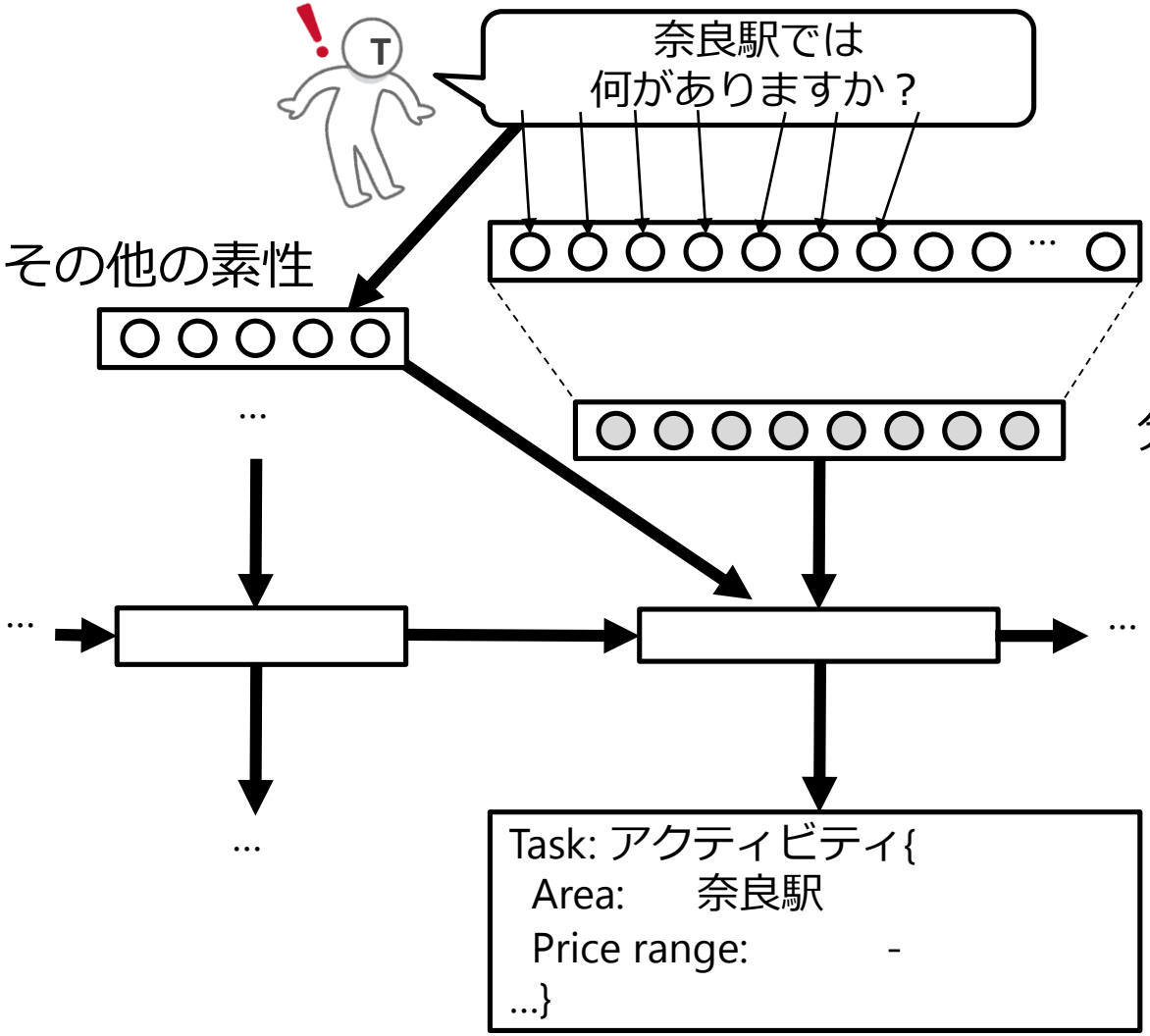
- **Word-Based Dialog State Tracking with Recurrent Neural Networks. Henderson et al., In Proc. SIGDIAL, 2014.**





図は論文より引用

Dialogue State Tracking with LSTM（入力文→対話状態）



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

Dialogue State Tracking using Long Short Term Memory Neural Networks. Yoshino et al., In Proc. IWSDS, 2016.

Dialogue State Tracking with CNN

Chinese word model

Chinese char model

(Translated) English word model

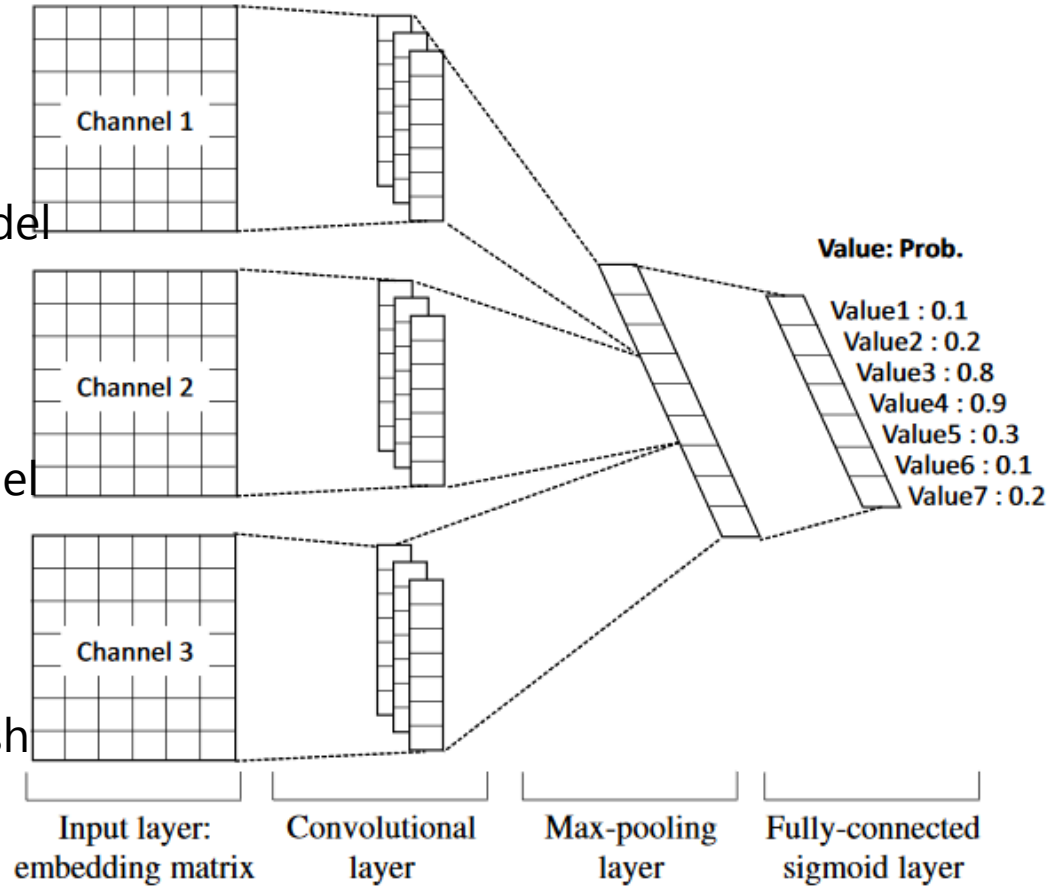


Fig. 2. Multichannel CNN model architecture for three input channels.

A MULTICHANNEL CONVOLUTIONAL NEURAL NETWORK FOR CROSS-LANGUAGE DIALOG STATE TRACKING
Shi et al., In Proc. IEEE-SLT 2016

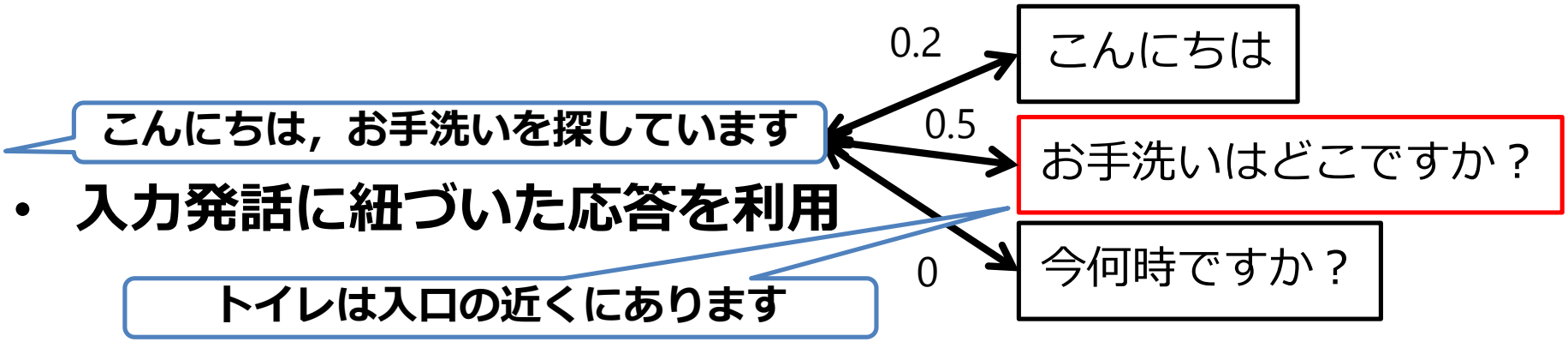
用例対話システム（入力文→出力文）

- **入力発話例/応答文対を用意**

Adaptive selection from multiple response candidates in example-based dialogue. Mizukami et al., In Proc. ASRU, 2015.

入力発話例	応答文
こんにちは	こんにちは
お手洗いはどこですか？	トイレは入口の近くにあります
今何時ですか？	今は<Hour>時<Minute>分です

- **入力発話例と実際の入力との類似度を計算**

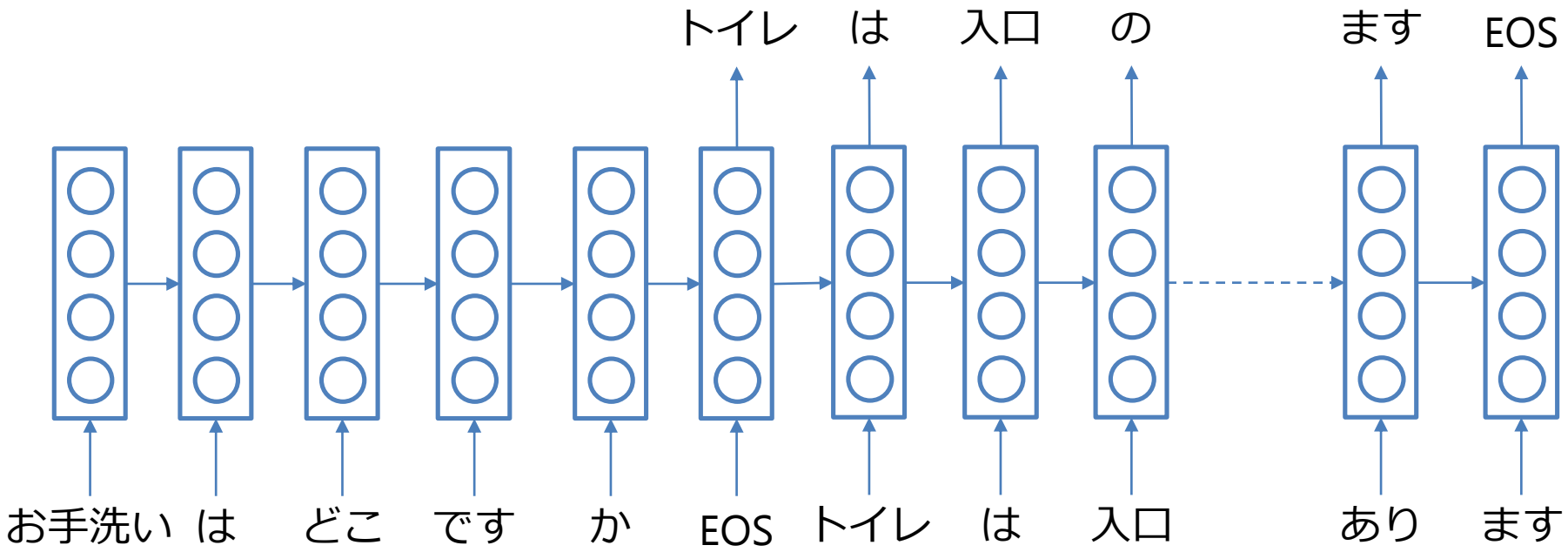


- **入力発話に紐づいた応答を利用**

トイレは入口の近くにあります

Seq2seq
（入力文→入力文）

- **Recurrent Neural Network (RNN)を用いたエンコーダ・デコーダモデルによる発話生成**



Deep Reinforcement Learning（強化学習に対するDNNの適用）

- **POMDPの問題は任意の b, a に対する $Q(b, a)$ の計算**
  - Q値を最大化するペアの探索

- **学習データに存在する $Q(b_i, a_i)$ から未知の $Q(b_k, a_k)$ を求める → 教師あり学習**

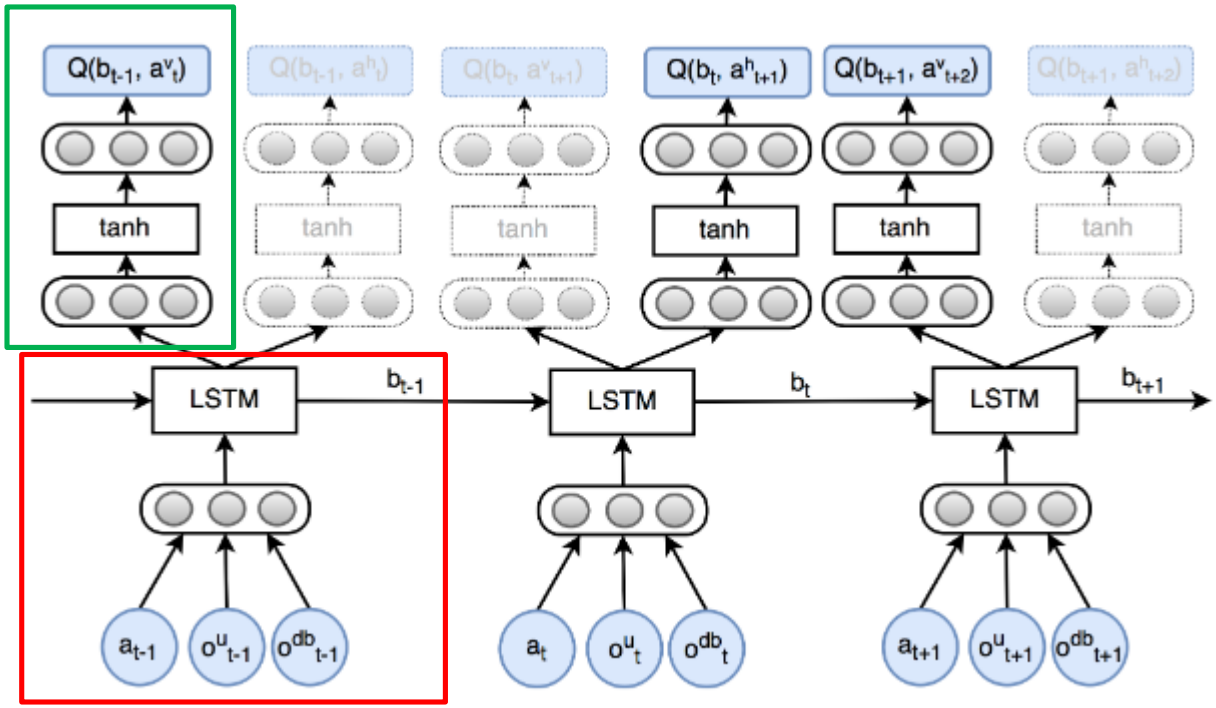
$$\mathcal{L}(\theta_i) = E_{(s,a,r,s')}[(y^{DQN} - Q(s, a; \theta_i))^2] \quad (1)$$

$$y^{DQN} = r + \gamma \max_{a'} Q(s', a'; \theta_i^-) \quad (2)$$

Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. Zhao et al., In Proc. SIGDIAL, 2016

LSTMを用いた言語理解と DQNによる対話制御の接続（入力文→システム行動）

- **LSTMによる言語理解をDQNの入力として利用**
- **DQNは任意の b, a のQ値を計算**



LSTM: 観測結果から $\boldsymbol{b_t}$ を計算

DQN: 与えられた $\boldsymbol{b_t}$ に対して $\boldsymbol{Q(b_t, a_{t+1}; \theta)}$ を計算

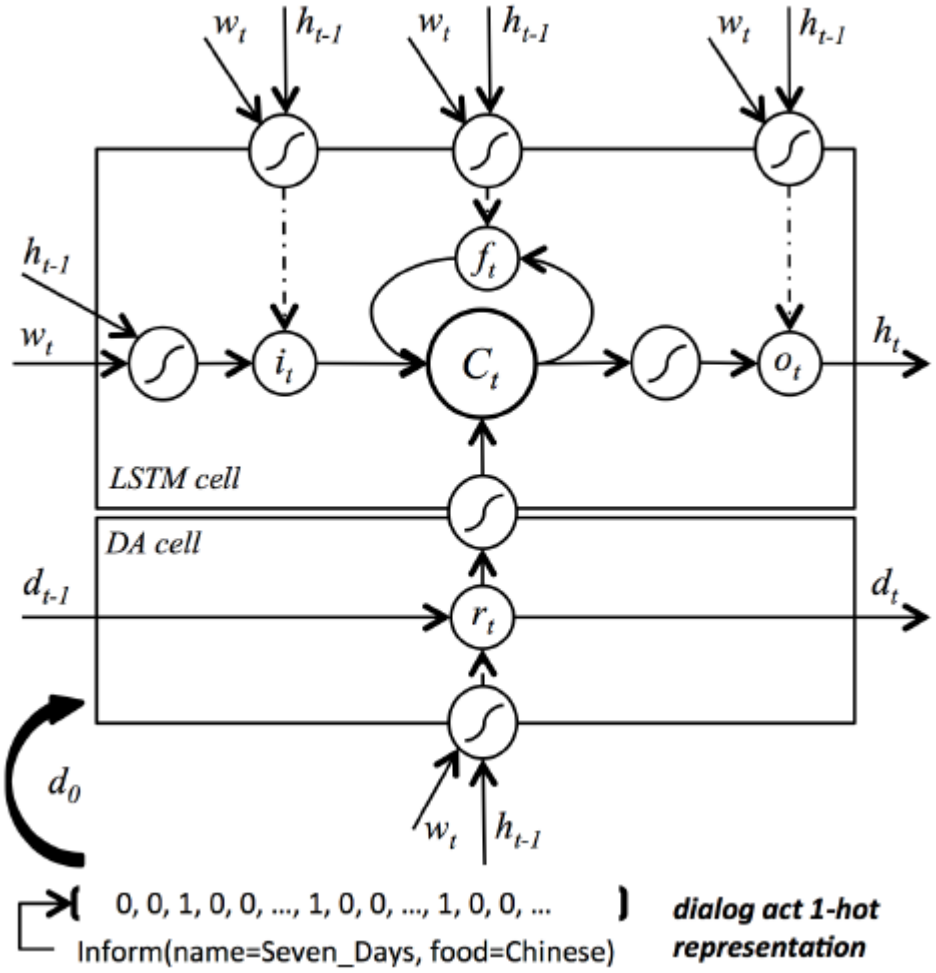
最後に全体のニューラルネットをファインチューニング

LSTM発話生成
（システム行動→出力文）

recurrent hidden layer

embedding of a word

1-hot dialog act and slot values



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

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

なんかEnd-to-end (seq2seq) うまくいかないんだけど？

- **利点**
  - 言語理解・対話制御を設計不要
    - 入力発話から直接出力発話を推定する
    - データさえあればシステムが動く
- **欠点**
  - **応答が一意でない場合学習がうまくいかない**
  - 制御が難しい
    - リスク管理のフィルタは結構大変

現状のNN技術に関する重要な金言

ニューラルネットワークやディープラーニングは、基本的に写像を認識する問題であって、写像を定義するものではありません。

【人工知能はいま　専門家に学ぶ】（１１）
音声認識研究の第一人者、河原達也氏が見るＡＩの世界,
Sankei-biz, 2016.12.19

End-to-endで解けるもの解けないもの

- **「言い換えれば、写像問題に置き換えられて、かつ正解がきちんと与えられるものについてはディープラーニングでほとんど解けるでしょうが、それ以外は難しいということです」（河原達也）**

- **自分が解こうとしてる問題が写像として定義できているのか考えよう**
  - Seq2seqも定型応答は写像として解けるはず
- **新しい組み合わせを写像として定義できればワンチャン**

- **本当に面白い研究は写像を定義できるシステムを作ろうとする挑戦にある…はず**

→そう思う人は中村研へ