

その研究 ChatGPT でいいんじゃないですか？ ～LLM時代の対話システム研究～

理化学研究所GRP/奈良先端科学技術大学院大学
吉野 幸一郎



知識獲得・対話研究チーム
Knowledge Acquisition &
Dialogue Research Team



ガーディアンロボット
プロジェクト
Guardian Robot Project

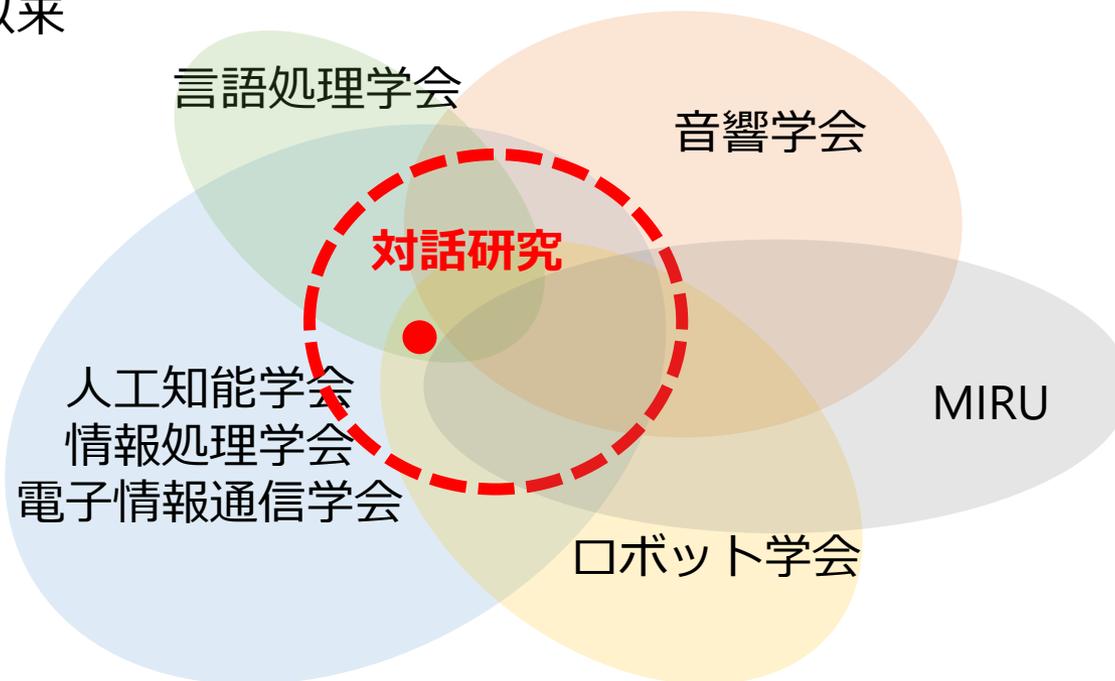


奈良先端大
ロボット対話知能研究室
Intelligent robot dialogue
laboratory, NAIST

吉野 幸一郎



- ◆理化学研究所 ガーディアンロボットプロジェクト チームリーダー
- ◆奈良先端科学技術大学院大学 ロボット対話知能研究室（連携） 教授
- ◆専門は音声対話システム・ロボット対話システム
 - 最近自分でも何の専門家なのかよくわからない...
 - YANSは2015年以来

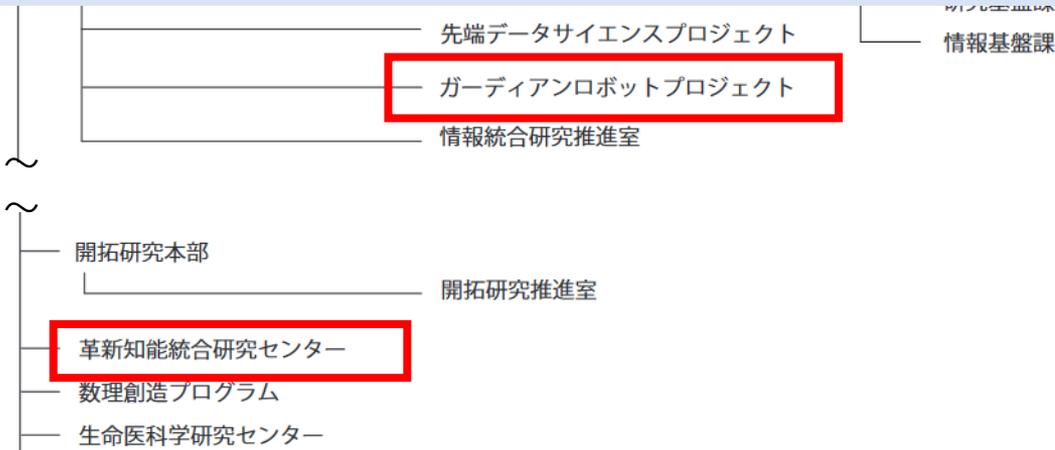


よく勘違いされますが...

◆私は理研@東京ではなく理研@京阪奈にいます

GRPのオフィスはけいはんなATR内

- 奈良先端大、阪大、京大に連携研究室
 - 知能ロボット学研 (中村、港、石井)
 - ロボット対話知能研 (吉野)
 - マルチモーダル環境認識研 (川西)
 - 計算論的認知神経科学研 (佐藤)



2024年2月某日



Transformerを使って〇〇という
対話システムができました！

こうならないうための防衛術が必要

ところで素人質問ですが、
ChatGPTではできないんですか？

うっ...



LLMで本質的に何が解けたか
何が解けていないのか、
一緒によく考えてみましょう

問題

◆ACL2011でタイトルに dialog と入っている論文の件数は？

0件

5件

10件

20件以上

◆ACL2023でタイトルに dialog と入っている論文の件数は？

10件

33件

66件

100件以上

なぜ対話研究は流行らなかったか

◆関連する技術が多過ぎた

- 例えば音声対話システムを動かそうとする場合、モジュールとして「音声認識」「言語理解」「対話制御」「言語生成」「音声合成」を動かさないといけない

◆動かなさ過ぎて何にフォーカスしていいかわからない

- 音声認識だけでも（当時は）エラー率が20%くらいあった
- 対話制御も言語生成もまともに動かない
- 諦めて強化学習のアルゴリズムをイチから書く（！？）

◆一部のモジュールを改善して対話が良くなった気がしない

- モジュールの改善と対話の質的改善が遠い
- 入試で「どこにフォーカスした研究を説明すればいいかわからない」

なぜ対話研究は流行ったか

◆基盤技術が統一化された・パッケージ化された

- ニューラルネットワークで何でもできる
- 研究用ソースコード・プラットフォームの共通化
 - NN framework + Huggingface 🙌

◆とにかく動くようになった

- 音声認識をはじめとするE2Eモデルの隆盛
 - 何でもそこそこの精度でとにかく動く
- Transformerをはじめとする言語モデルの普及
 - 流暢な言語生成ができるようになったのは重要

◆ネットワークの改善=モデルの改善

- 「どういう目的」で「ネットワークのどの部分」を「どのような構造」にするか、「どのような目的関数を与える」か



Siriの登場

A new hope: Transformer 型 LM

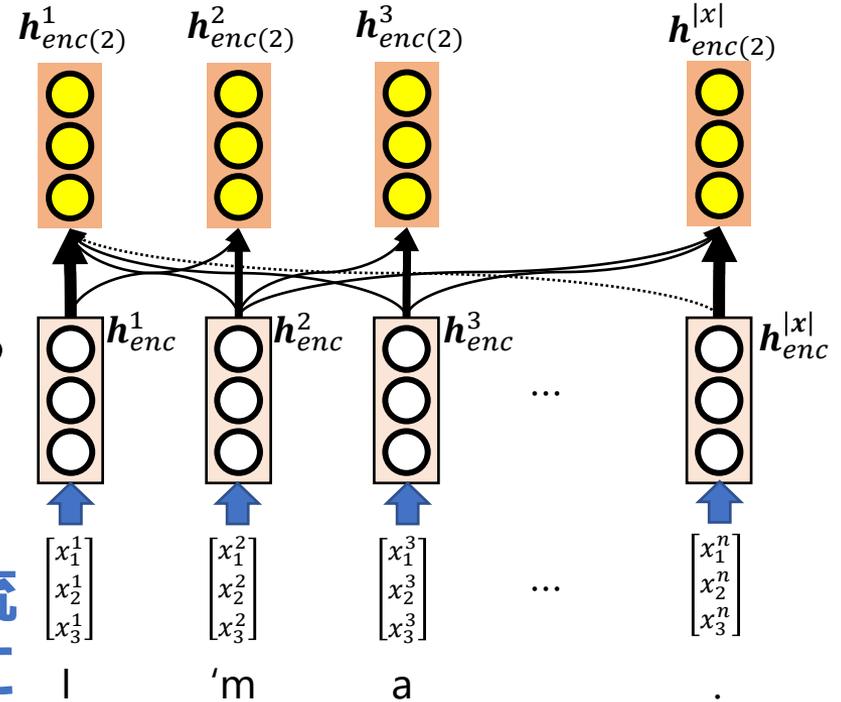
◆Transformer 型ネットワークによる自己注視構造

- 注視 (Attention) によって「入力要素同士の関係」を扱う

◆入力要素同士の相互依存性を陽に扱えるようになった

◆層を積み重ねることで second-order dependencies も扱えるようになった

- 文の構成性を暗に扱っている？



※ 意味とは...？理解とは...？

その研究ChatGPTでいいんじゃないですか？

Neural Networks Strike Back

◆Transformer 型ネットワークはよく出来ている

- 言語生成が非常に流暢にできるようになった

◆Transformer を使った LLM もよく出来ている

- 非常にシンプルな目的関数 (Masked Language Model; MLM) だが scaling law はまだ収束していない
- シンプル故に「データ・パラメータ数の勝負 = お金の勝負」になってしまった

参考: 言語モデルタスク

◆生成AI（大規模言語モデル; LLM）がやっていることは確率的言語モデルそのもの

- 文の生成確率 $P(s) = P(w_1, w_2, \dots, w_n)$
 $= P(w_1)P(w_2|w_1) \dots P(w_n|w_1, w_2, \dots, w_{n-1})$

◆文の構成要素の意味は周辺要素から理解可能（分布仮説）

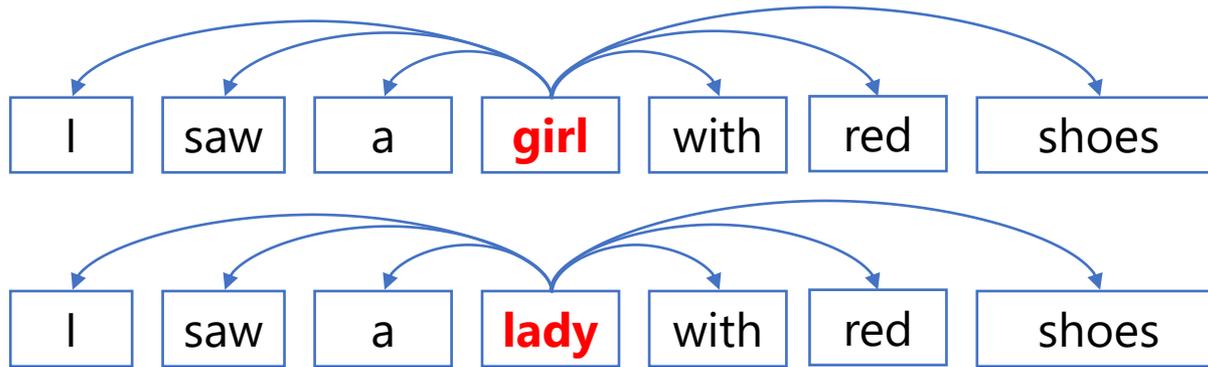
- $P(w_m|w_1, w_2, \dots, w_{m-1}, w_{m+1}, \dots, w_{n-1}, w_n)$

◆文の生成は文脈を与えられた言語モデルで可能（生成AI）

- $P(w_n|w_1, w_2, \dots, w_{n-1})$

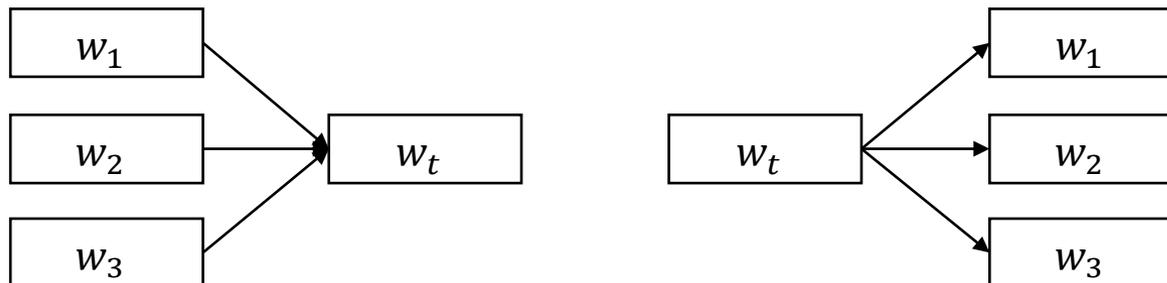
参考: 分布仮説

◆ 同じ文脈で出現する語が同様の意味を持つ傾向があるという仮説



• 今日の分散意味論の基礎

- CBOW, Skip-gram, masked LM などはこの考えに基づく



参考: 分布仮説の研究の歴史



- ◆ **Harris, Z. (1954). Distributional structure. *Word*, 10(23): 146-162.**
- ◆ **Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in Linguistic Analysis*, pp. 1-32. Oxford: Philological Society. Reprinted in F.R. Palmer (ed.), *Selected Papers of J.R. Firth 1952-1959*, London: Longman (1968).**
- ◆ **McDonald, S., and Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. In *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pages 611-616.**
- ◆ **Sahlgren, Magnus (2008) The Distributional Hypothesis. *Rivista di Linguistica (Italian Journal of Linguistics)*, 20 (1). pp. 33-53.**

https://aclweb.org/aclwiki/Distributional_Hypothesis

最初の論文には何が書かれているのか？

参考: 分布仮説を提唱した言語学者

ゼリグ・ハリス

文 17の言語版 ▾

ページ ノート

閲覧 編集 履歴表示 ツール ▾

出典: フリー百科事典『ウィキペディア (Wikipedia) 』

ゼリグ・ハリス（Zellig Sabbetai Harris、1909年10月23日 - 1992年5月22日）は、アメリカ合衆国の言語学者。数理言語学が専門で科学の方法論にも博通する。元はセム語学者で構造主義、談話分析、生成統語論で知られる。

略歴 [編集]

- 1909年、ウクライナ南西のオデッサ州の**バルタ**に生まれる。
- 1913年、アメリカのペンシルベニア州の**フィラデルフィア**に移住する。
- ペンシルベニア大学の東洋学科に入り1930年に学士、1932年に修士、1934年に博士の称号をとる。
- 1931年にペン（Penn）で教務につき、1946年同大学の言語学科に勤める。

弟子 [編集]

- **ノーム・チョムスキー**（Noam Chomsky）

<https://ja.wikipedia.org/wiki/ゼリグ・ハリス>

参考: 分布仮説の基礎

- 1. 言語の各要素は任意の順序・相対位置では発生しない**
 - 単語の生成順序、生成内容には制約がある
- 2. あらゆる要素の生成はその出現位置に制約を持つ**
 - この制約は確率的に議論することができる
- 3. 文中の他の要素との相対状態を見ることで、その要素が文法規則に従うか知ることが出来る**
 - スラングなど従来は文法規範に当てはまらないと考えられてきた現象も他の語との相対位置で説明できる
- 4. 共起する他の要素の分布を見ることで語の説明ができる**
 - もちろん等確率にはならない

※ element を言語の要素、文中の要素と訳しているが、典型的には単語や音素を指す

参考: 分布仮説の基礎

1. 言語の各要素は任意の順序・相対位置では発生しない
→ 言語の順序・相対位置には何らかの意味がある
2. あらゆる要素の生成はその出現位置に制約を持つ
→ 確率的言語モデルそのもの
3. 文中の他の要素との相対状態を見ることで、その要素が文法規則に従うか知ることが出来る
→ 言語モデル尤度に対する考え方
4. 共起する他の要素の分布を見ることで語の説明ができる
→ 今日の分布意味論 (distributional semantics) の基礎

※ element を言語の要素、文中の要素と訳しているが、典型的には単語や音素を指す

参考: 分布で意味を捉える

◆周辺語の分布で形態素の意味（語義）を捉えることは可能

◆ただし**スパーシティの問題**は大きい

- 形態素同士の組み合わせは使われ方に依存
 - neighbor → neighborhood はあらゆる場合に成り立つか？
- 形容詞は名詞を修飾するが、あらゆる形容詞と名詞の組み合わせが存在するわけではない
- 音素の組み合わせに関しても同様
 - 日本語における `q u e` など

これら組み合わせの偏りを意味のクラスとして捉える
e.g., suffix に “hood” を持つ単語群

参考:語の意味から文の意味へ

◆文中の任意の部分に対して分布仮説が成り立つ

- 文の順序それ自体も意味を持つ

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, ArXiv 2018.

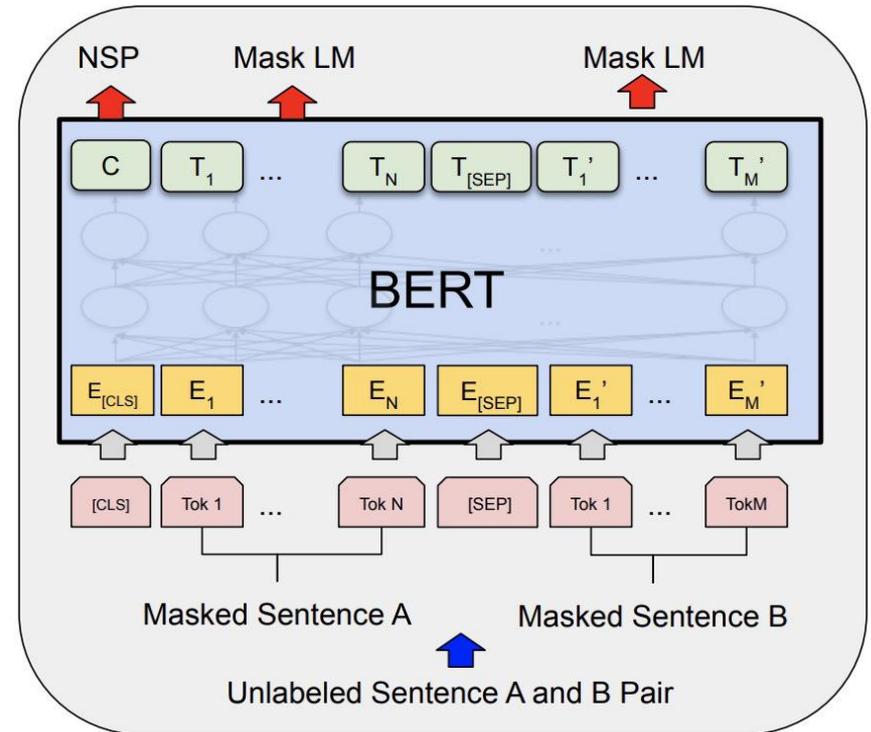
- Masked Language Model (MLM), Next Sentence Prediction (NSP)

◆MLM: 周囲の文脈から単語の穴埋め予測を行う

- 単語予測を行うネットワークが文脈の意味を学習する
- Self supervised learning により学習事例を無限に自動生成可能

◆NSP: 文全体から前後の文を予測する

- 愚直な分布仮説の拡張
- そこまでうまくいっていない
 - データ量の問題? 問題設計が悪い?
 - 文同士で分布仮説成り立つ?



参考: 言語モデルの評価・perplexity



◆任意のテストセットに対して モデルパラメータから対数尤度を計算

- MLM（またはNSP）で訓練したニューラルネットワークのパラメータ θ によってテストセットに出現する単語列の尤度を計算

$$ENT = -\frac{1}{n} \sum_i^n \log P(w_i; \theta)$$
$$PP = 2^{ENT}$$

- テストセットを予測しようとしたときにどれくらい別の候補と迷うか（テストセットを再現するために何ビット必要か）

結局 LLM ができて

◆入力した文の意味※がよく埋め込めるようになった

- 文の構成性も学習できているかもしれない

◆言語生成が非常に流暢にできるようになった

- 特に与えられた情報にある程度一意に対応するような文はかなり流暢に生成できるようになった

◆対話の研究はこれを使えば何でもできる（！？）

- ある意味では真かも

※ 意味とは...？理解とは...？

ある医療情報処理の先生の言葉

医療情報処理の研究目的とは何でしょうか？

医療情報処理は医学の一部です。

医学の研究目的は最後は

「患者が何人良くなったか」

「患者の寿命がどれだけ伸びたか」

しかありません

対話研究の研究目的曖昧がち問題

◆今の対話システムの評価は対話システムによる社会変革に直結しているのか？

- Perplexity が下がることにどういう意味があるのか
- 発話の魅力度が 0.1 上がることに何の意味があるのか

◆そもそも対話研究の目的って？

- 目的が決まらないと評価も決まらない
- 何のために対話システムを作るのか？



機械と会話できて
何が嬉しいの？

その研究ChatGPTでいいんじゃないですか？

主たる研究の目的（スタンス）



工学的立場

- ◆研究によって解決したい何らかの問題が存在する
- ◆研究の主評価は「その問題がどの程度解決されたか」
- ◆いくつかのサブゴールが設計し得るが、サブゴールは研究の主評価に結びつかなければならない

理学的立場

- ◆研究によって明らかにしたい未知の現象が存在する
- ◆研究の主評価は「その現象をどの程度明らかにしたか」
- ◆いくつかのサブゴールが設計し得るが、サブゴールは対象とする現象に結びつかなければならない

対話研究にはどういう目的があるか？

工学的立場

◆対話システムを作ること
で何らかの問題を解決する

◆タスク対話

- タスク達成率がどれだけ
上がったか

◆雑談対話

- 雑談対話のゴール？

理学的立場

◆対話における何らかの現象
を仮説ベースで明らかにする

◆ACTの分析

- 特定の対話における行動レ
ベルでの振る舞いが明らか
になったか

◆雑談の分析

- 雑談における何らかの振る
舞いがモデル化できるよう
になったか

言語モデルで科学をする



◆ AI for Science

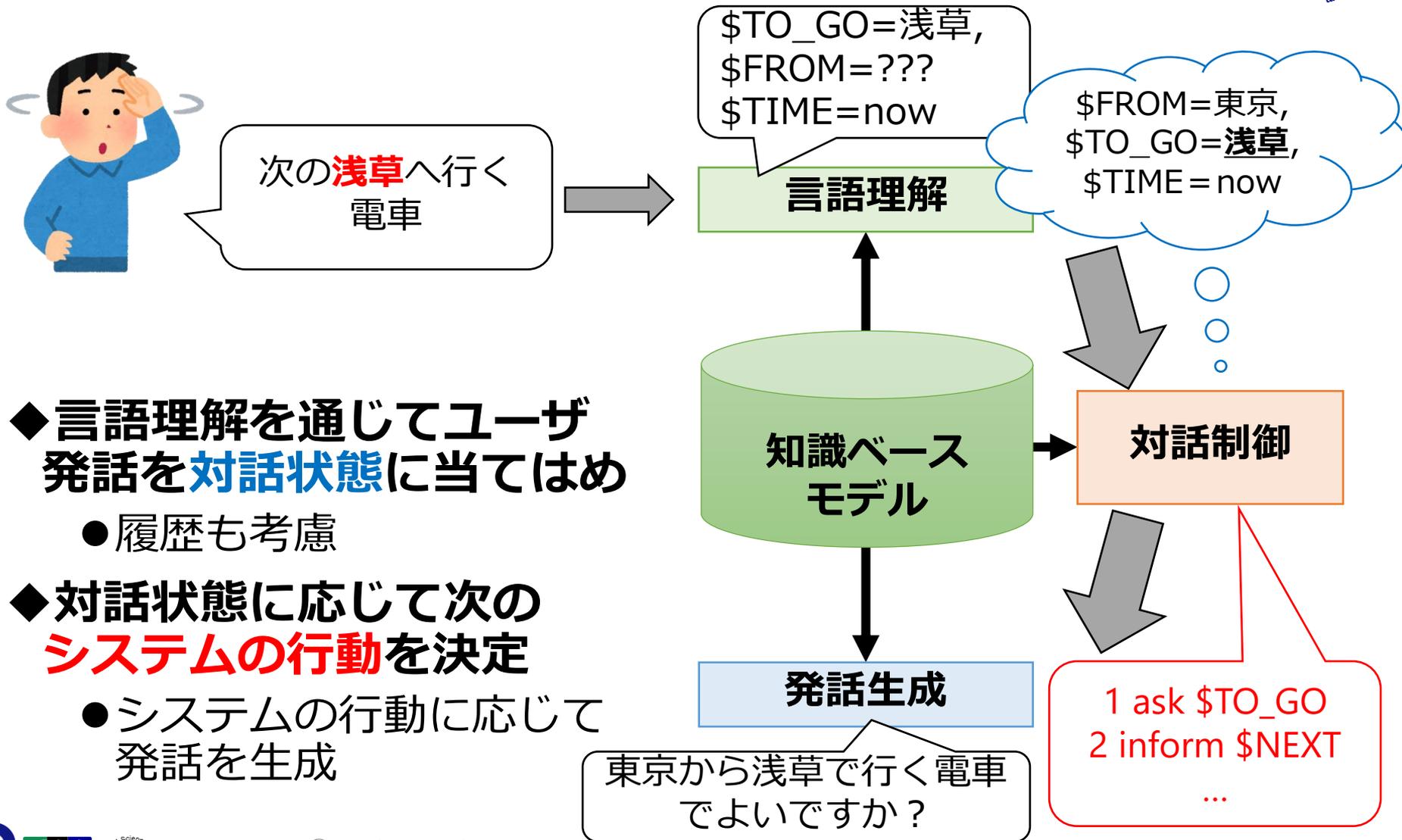
◆ 例えば言語モデルが与える尤度と言語モデルに与える条件からある文のサプライザル（文処理に掛かる負荷）を計測することができるのでは？

- [S5-P28] ニューラル言語モデルのサプライザルに基づいた会話におけるスタイルシフトの評価
- 言語モデルが与える尤度が低=その言語モデルから外れた文

◆ 言語モデルの振る舞いを分析することで「一般的なテキスト生成」の振る舞いを明らかにすることができるのでは？

◆ その他「対話のモデル」の正当化・反証を言語モデルによって行うことができるのは？

Return to the Task-Oriented Dialogue



◆言語理解を通じてユーザ
発話を**対話状態**に当てはめ

- 履歴も考慮

◆対話状態に応じて次の
システムの行動を決定

- システムの行動に応じて
発話を生成

その研究ChatGPTでいいんじゃないですか？

言語理解と対話制御

◆言語理解

- ユーザの発話を機械が理解可能な形に変換



◆対話制御

- 言語理解の結果と過去の履歴から次の行動を決定

言語理解

```
Train_info{$TO_GO=浅草,$FROM=???, $TIME=now}
```

履歴

```
$FROM=東京
```

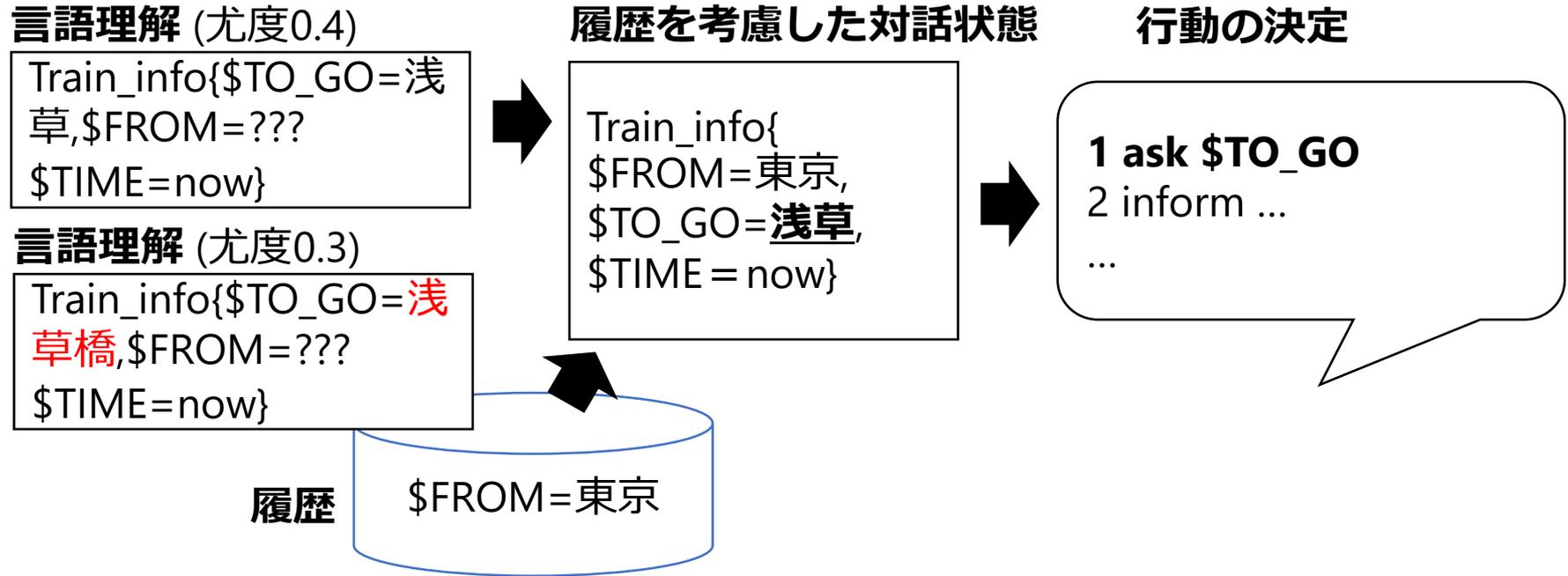
履歴を考慮した対話状態

```
Train_info{
  $FROM=東京,
  $TO_GO=浅草,
  $TIME=now}
```

行動の決定

```
1 ask $TO_GO
2 inform ...
...
```

認識誤りの利用



◆最後のタスク達成率を最大化することが目標

- 音声認識や言語理解が間違っている可能性がある → 聞き返す
- 教師信号が遅れて与えられる → 強化学習

対話状態とシステム行動のデザイン



◆対話状態はフレームとフレームが持つSlot-Valueで管理

- フレームがドメインを表わす
- フレームに記述されたSlotとValueの組により状態を記述

```
フレーム名{  
Slot1=Value1  
Slot2=Value2  
Slot3=Value3  
}
```

```
Train_inform{  
$FROM=東京  
$TO_GO=浅草  
$TIME=???  
}
```

◆システムの行動 = 対話システムの挙動

- 基本的な挙動が定義できれば
あとはフレーム + Slot-Valueの組に対応させる
- 基本的な行動: 実行、確認など

対話状態推定と対話制御の数理モデル

◆問題定義: 言語理解の系列が与えられた場合の 現時刻における最適行動の決定

- 最適行動 = 早く正確に情報を提供する
- $s^t \in I_s$ ターン t のユーザ状態 (ゴールを含む)
- $a^t \in K$ システムの行動
- $o^t \in I_o$ 観測状態
- $b_s^t = P(s|o^{1:t})$ ユーザ状態が s である信念 (確率変数)

◆どう b_s^t を更新するか → Belief update

◆どう a^t を見つけるか → 強化学習

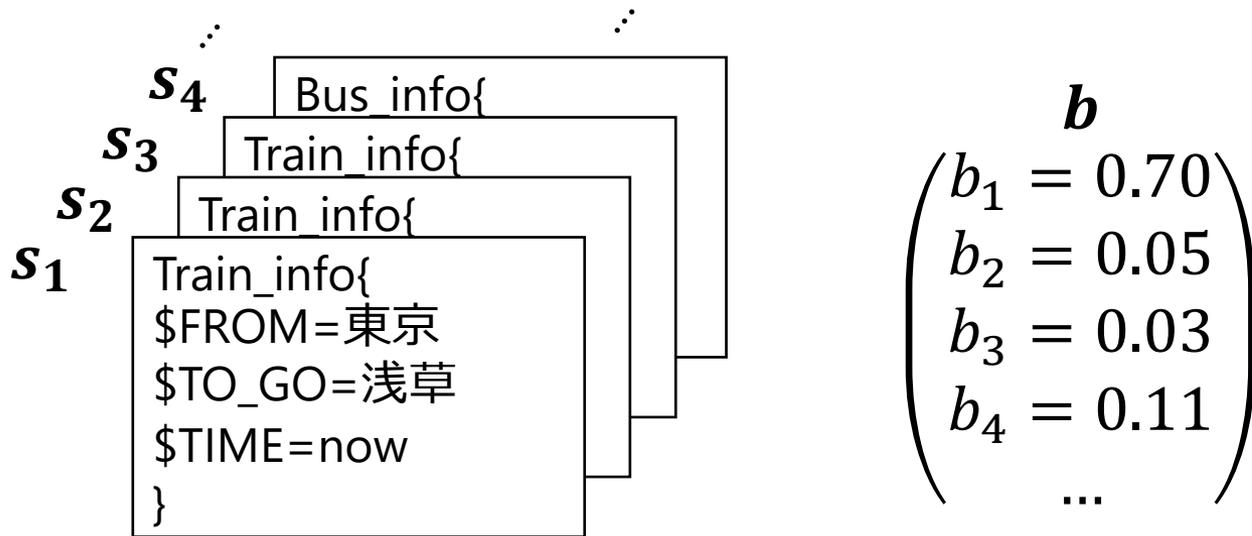
Belief update

◆言語理解の結果は全時刻分与えられる

- 過去の履歴も考慮

$$\begin{aligned}
 b^{t+1} &= P(s^{t+1} | o^{1:t+1}) = \frac{P(s^{t+1}, o^{t+1} | o^{1:t})}{P(o^{t+1} | o^{1:t})} \\
 &\propto \sum_s P(o^{t+1}, s^{t+1} | s) P(s | o^{1:t}) \propto \sum_s P(o^{t+1}, s^{t+1} | s) b^t \\
 &= P(o^{t+1} | s^{t+1}, s^t) \sum_{s_i} P(s'_j | s_i) b^t \\
 &\approx \underbrace{P(o^{t+1} | s^{t+1})}_{\text{観測確率}} \sum_{s_i} \underbrace{P(s'_j | s_i)}_{\text{状態遷移確率}} \underbrace{b^t}_{\text{belief}}
 \end{aligned}$$

対話状態推定: 実際の動作や変数の中身



◆対話状態推定

- そのターンにおける適切な対話状態（スロット値つきフレーム）を
どれだけ尤度高く（または1-bestで）予測できるか
 - 言語理解の出力する事後確率 $P(o_i | s_i)$
 - RNNに現在のユーザ発話を入力した結果の出力

Belief update から RNN へ

◆ Belief update

$$\bullet b^t \approx \underbrace{P(o^t | s^t)}_{\text{観測確率}} \underbrace{\sum_{s_i} P(s_i | s^{t-1})}_{\text{状態遷移確率}} \underbrace{b^{t-1}}_{\text{belief}}$$

◆ RNN: 与えられた発話の系列から対話状態を出力

$$\bullet h^t = \tanh(\underbrace{W_{Xh} X^t}_{\text{観測確率}} + \underbrace{W_{hh} h^{t-1}}_{\text{状態遷移確率}} + \underbrace{b_h}_{\text{belief}})$$

観測確率 状態遷移確率 belief

● X^t は時刻 t における単語系列

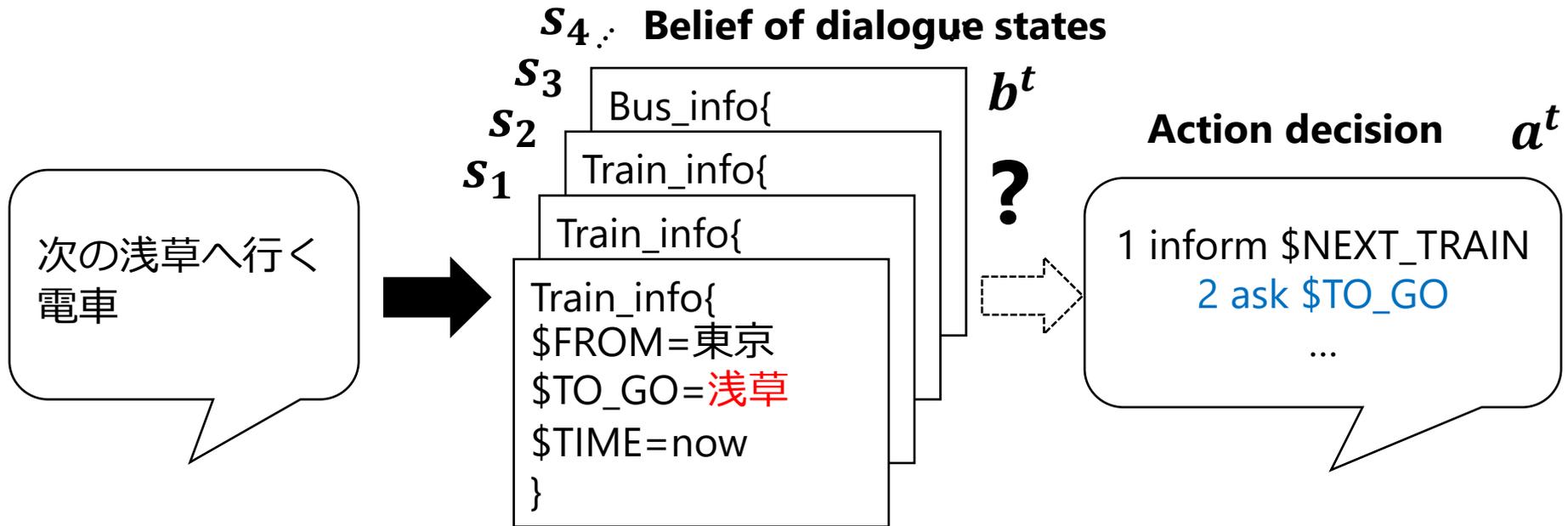
● 隠れ層 h^{t-1} が過去の履歴を集約したもの

$$\bullet y_p^t = \operatorname{argmax} \left(\operatorname{softmax} \left(W_{hy_p} h_t + b_{y_p} \right) \right)$$

● 対話状態 y_p^t を尤度に基づいて出力

部分観測な状況での行動決定

◆信念 b^t が入力された時の行動 a^t を決定



◆強化学習で決定

- 報酬: 対話のゴールに到達する (+), 時間が経つ (-)

強化学習による行動決定

◆ある状態 s のときに行動 a を選ぶ確率 $\pi(s, a)$

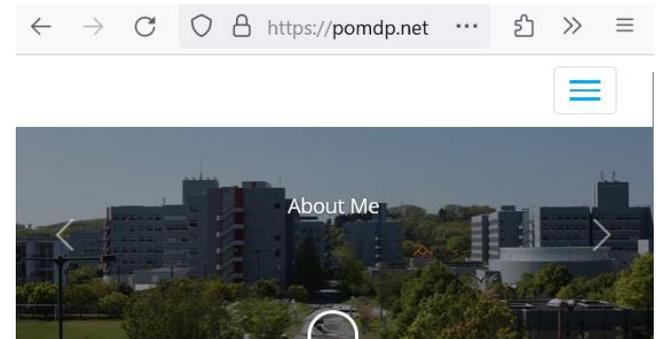
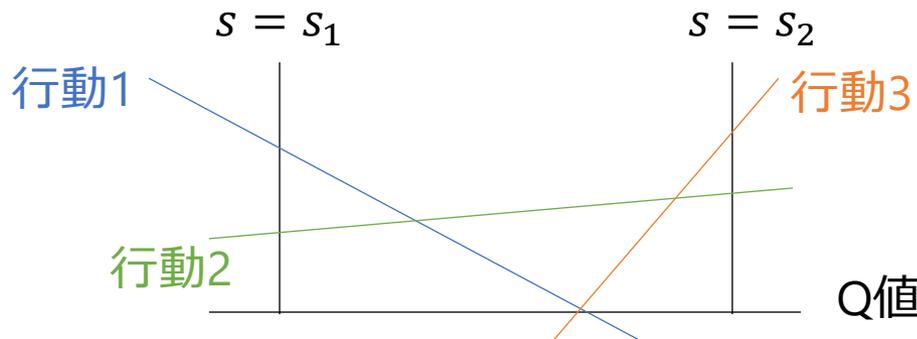
- 強化学習で $E(r) = \sum_{t=0}^{\infty} \gamma^t r_{t+1}$ を最大化する方策 π を見つける
- Q学習やSARSAが用いられてきた (2000年代)
 - $Q(s, a) \leftarrow (1 - \alpha)Q(s, a) + \alpha(r + \gamma \max_{a'} Q(s', a'))$ で高いQ値を出す行動を求める

◆Belief のスパースさの問題

部分観測マルコフ決定過程 (POMDP) の利用!

γ は割引率
 α は学習率

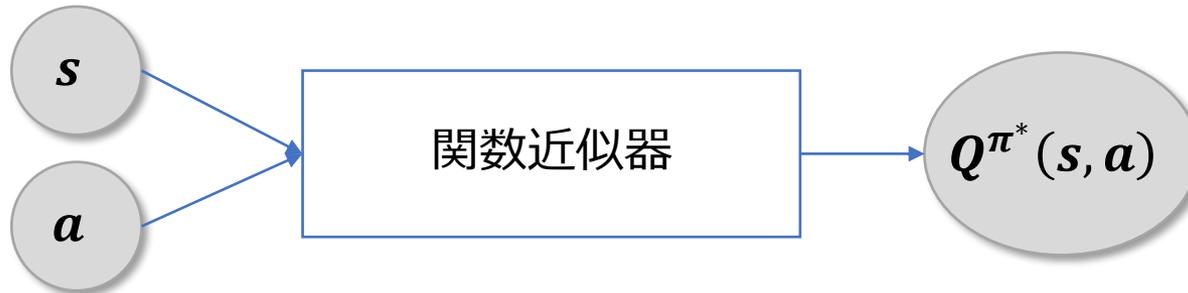
- 与えられるのは状態 s ではなく確率変数 b
- b の組み合わせは無限にあり求めるのは難しい



深層強化学習の登場

◆関数近似と深層強化学習

- Q値を求めるニューラルネットワークを学習する (Q-network)
- 高いQ値を出すかどうかNNに聞いてみよう！

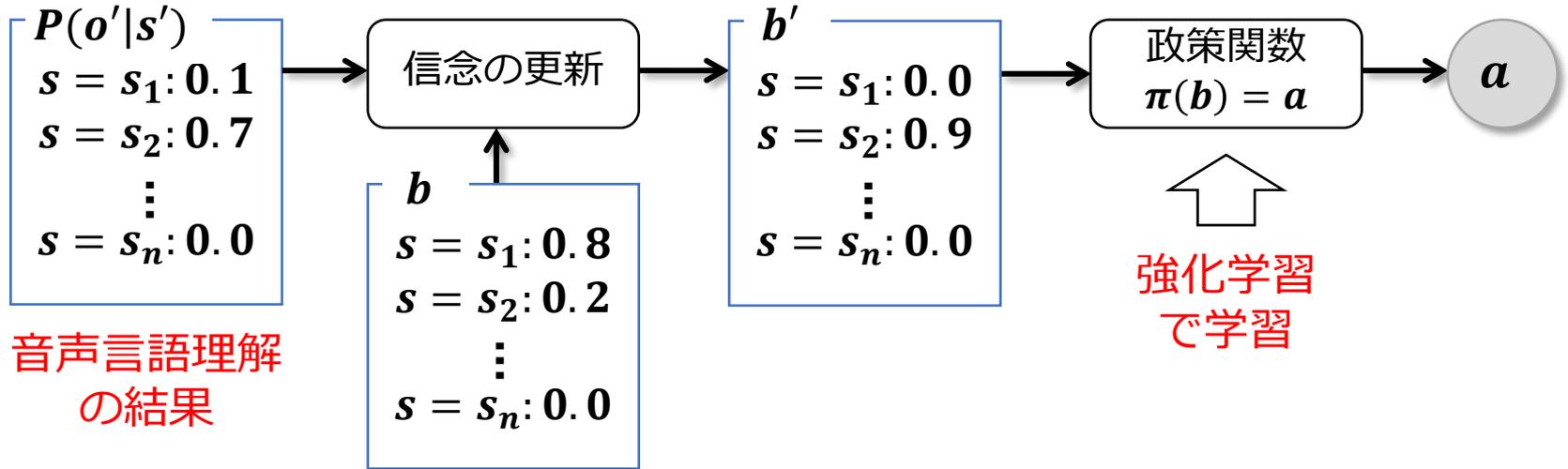


- NNは入出力の対応が与えられれば学習できる
- 状態などを入力して値を出力する回帰モデル

◆この回帰に深層学習を使うのが Deep Q-network (DQN)

- $\mathcal{L}(\theta_i) = E \left[(y - Q(s, a))^2 \right]$
- $y = R(s, a, s') + \gamma \max_{a'} Q(s, a')$

対話制御の全体像 (旧)



◆ $s \in I_s$

◆ $a \in K$

◆ $o \in I_s$

◆ $b_i = P(s_i | o^{1:t})$

◆ $\pi(b, a)$

ユーザ状態

システムの行動

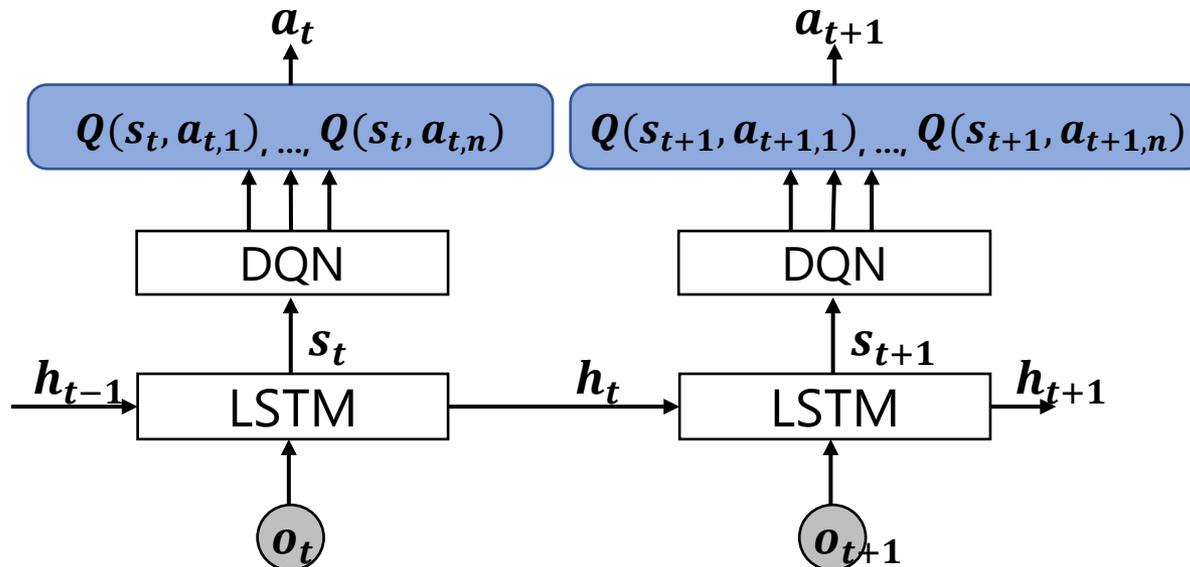
観測状態

$s = s_i$ である信念 (確率変数)

政策関数

対話制御の全体像 (新)

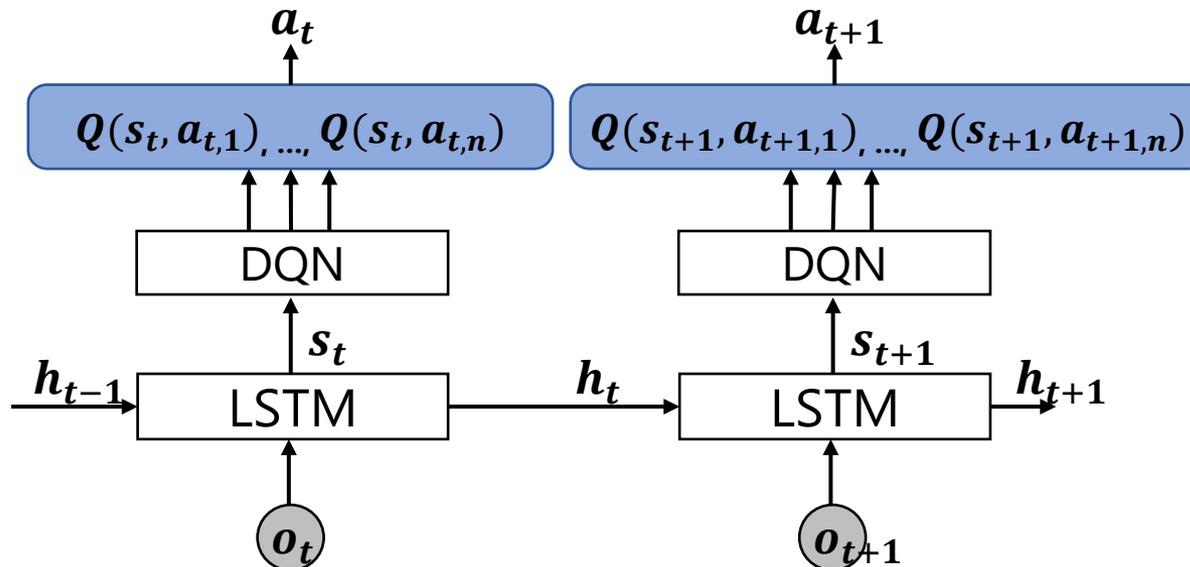
- ◆ Belief update を RNN (LSTM) で置き換え
- ◆ 行動決定部分を深層強化学習 (DQN) で置き換え
 - LSTM と DQN を直結すれば belief も考慮可能
 - s を h で置き換えてDQNに直接入力



対話状態はいらないのでは？

◆ Modular-based → E2E

- 全体の fine-tuning は DQN の目的関数で最適化
- 入口と出口が決まっていればそれでよい
 - タスク対話で E2E の論文も出るようになってきている
- エンコーダは Transformer や事前学習モデルの出力でもよい



入力モダリティの拡張

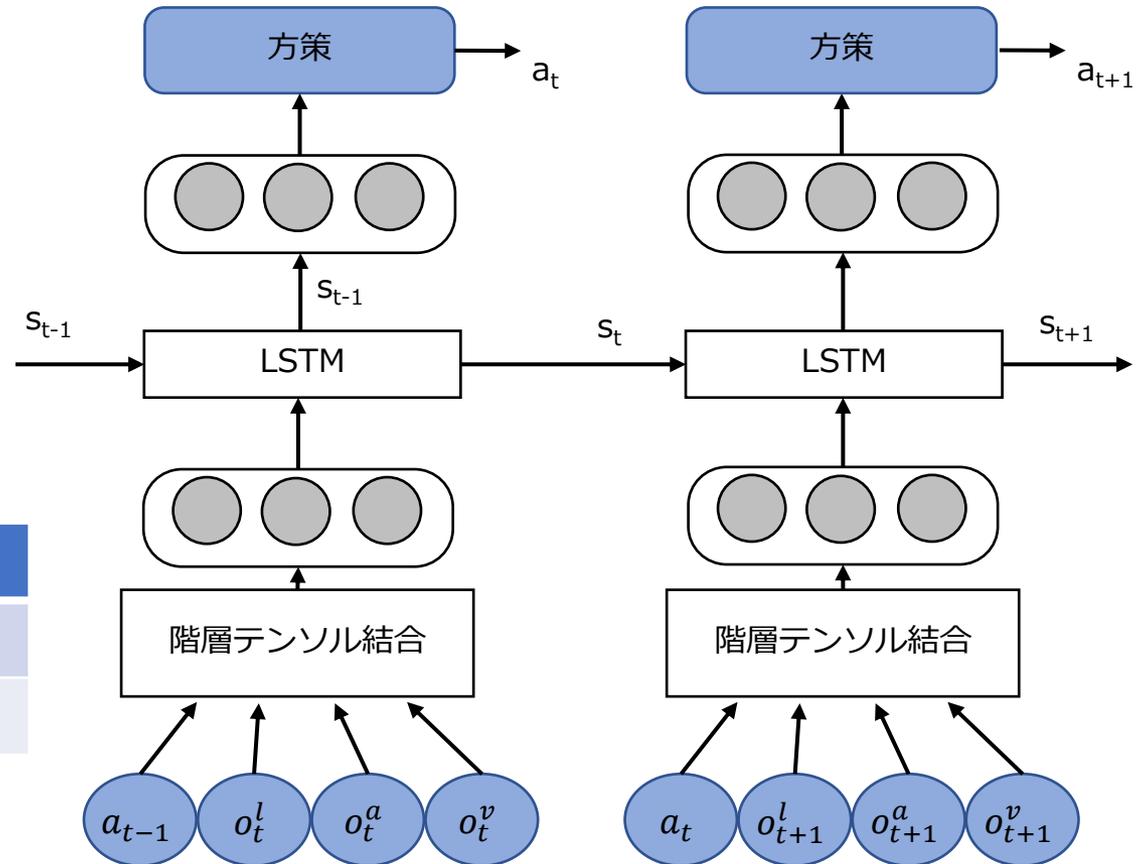
◆あらゆるモダリティの特徴量を入力

- o^l : 言語特徴
- o^a : 音響特徴
- o^v : 画像特徴

◆中間の対話状態をアノテーションしなくても学習可能

The Tung Nguyen, et al., Policy Reuse for Dialog Management Using Action-Relation Probability
IEEE Access, 2020.

| DM model | DA selection |
|----------|--------------|
| Modular | 71.2% |
| E2E | 69.95% |



その研究ChatGPTでいいんじゃないですか？

システムの『意図』の存在意義

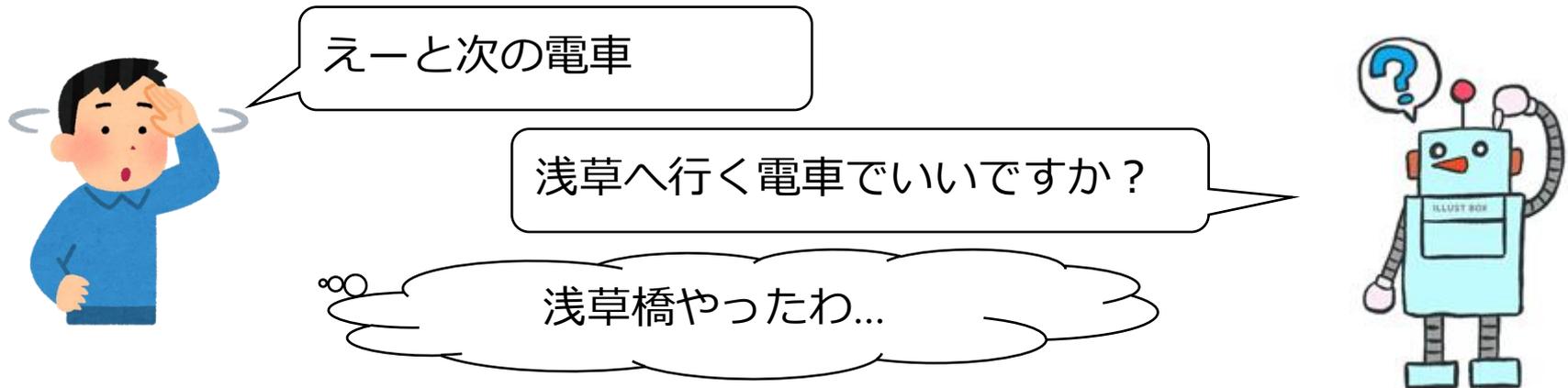
◆E2Eで対話状態などの中間状態は意味がなくなったのか？

- アノテーションと精度向上のコストの見合い
- 真に必要なアノテーションを定義する必要

◆システムがどの行動をするのか、なぜその行動をするのか

- E2Eのシステムでは意図は隠蔽化

◆説明可能であるためには意図は重要



◆何のための『意図（状態）』かの再考

- そのタスクにその状態が真に必要なかどうか

LLM とタスク対話

◆E2E（ユーザ発話を入力してシステム発話を出力）でもタスク指向対話はそこそこ解ける

- End-to-end task-oriented dialogue

| Model | MWOZ | | SMD | | CamRest | |
|--|--------------------|--------------------|--------------------|--------------------|--------------------|--------------------|
| | BLEU | Entity F1 | BLEU | Entity F1 | BLEU | Entity F1 |
| DSR (Wen et al., 2018) | 9.10 [‡] | 30.00 [‡] | 12.70 [‡] | 51.90 [‡] | 18.30 [‡] | 53.60 [‡] |
| KB-Retriever (Qin et al., 2019) | - | - | 13.90 | 53.70 | 18.50 | 58.60 |
| GLMP (Wu et al., 2019) | 6.90 [‡] | 32.40 [‡] | 13.90 [‡] | 60.70 [‡] | 15.10 [§] | 58.90 [§] |
| DF-Net (Qin et al., 2020) | 9.40 | 35.10 | 14.40 | 62.70 | - | - |
| GPT-2+KE (Madotto et al., 2020) | 15.05 | 39.58 | 17.35 | 59.78 | 18.00 | 54.85 |
| EER (He et al., 2020b) | 13.60 [§] | 35.60 [§] | 17.20 [§] | 59.00 [§] | 19.20 [§] | 65.70 [§] |
| FG2Seq (He et al., 2020a) | 14.60 [§] | 36.50 [§] | 16.80 [§] | 61.10 [§] | 20.20 [§] | 66.40 [§] |
| CDNET (Raghu et al., 2021) | 11.90 | 38.70 | 17.80 | 62.90 | 21.80 | 68.60 |
| GraphMemDialog (Wu et al., 2022) | 14.90 | 40.20 | 18.80 | 64.50 | 22.30 | 64.40 |
| ECO (Huang et al., 2022) | 12.61 | 40.87 | - | - | 18.42 | 71.56 |
| DialoKG (Rony et al., 2022) | 12.60 | 43.50 | 20.00 | 65.90 | 23.40 | 75.60 |
| UnifiedSKG (T5-Base) (Xie et al., 2022) | - | - | 17.41 | 66.45 | - | - |
| UnifiedSKG (T5-Large) (Xie et al., 2022) | 13.69* | 46.04* | 17.27 | 65.85 | 20.31* | 71.03* |
| Q-TOD (T5-Base) (Tian et al., 2022) | - | - | 20.14 | 68.22 | - | - |
| Q-TOD (T5-Large) (Tian et al., 2022) | <u>17.62</u> | 50.61 | 21.33 | <u>71.11</u> | 23.75 | 74.22 |
| Ours (T5-Base) | 17.23 | <u>53.68</u> | <u>24.79</u> | 69.79 | <u>25.04</u> | 73.09 |
| Ours (T5-Large) | 18.77 | 54.72 | 25.91 | 71.30 | 25.53 | 74.36 |

Wan, Fanqi, et al. "Multi-Grained Knowledge Retrieval for End-to-End Task-Oriented Dialog." Proc. ACL(2023).

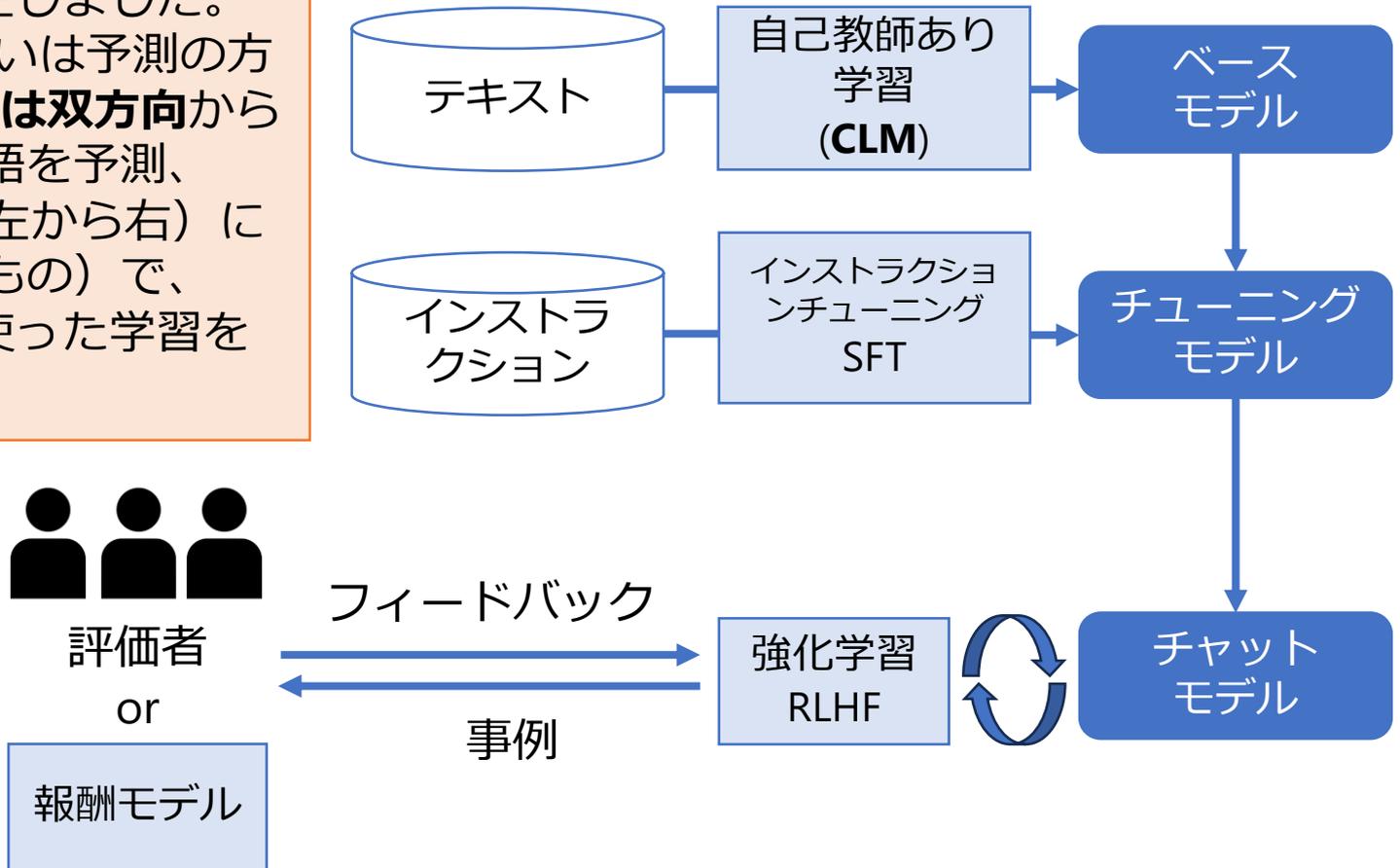
Task-orientedなのに
BLEUで評価するとは
... ???

システムの意図は
... ???

でもLLMがすごいのは
Zero-shotでそこそこできること

LLMの学習

柴田さんの指摘を受けてMLMをCLM (causal language modeling) に修正しました。MLMとCLMの違いは予測の方向の違い (**MLMは双方向**からマスクされた単語を予測、**CLMは単方向** (左から右) に単語を予測するもの) で、GPT系はCLMを使った学習をしています。



目的関数を理解する

◆モデルは教師信号を与えられているタスクに強い

- 当たり前だが忘れがち
 - LLM がそれっぽくやってしまうので...
- MLM/CLM 単体で学習したモデルより「後で当該タスクの目的関数が入っているモデル」の方が強い

◆MLM/CLM のタスクで解かれる問題は何か考える

- 言語モデリングのタスクと類似するタスクはシンプルに言語モデルを使う方が強い
 - E.g., 翻訳、要約、儀礼的な対話、音声認識...
 - **パターン認識で入出力いずれかがテキストになっている問題**
- 言語モデリングのタスクから遠いタスクは（LLMのバックアップを得られるとはいえ）FTした方がよい
 - E.g., 知識推論、（一般的内容でない）QA

目的関数にないことはFTには勝てない
テキストに書かれていないことはできない

これはなんでしょう？

「ご視聴ありがとうございました！」

「チャンネル登録よろしくね！」

「最後まで視聴してくださって本当に
ありがとうございます。」

```
[asr_node-1] [INFO] [1693384595.195385386] [asr_node]: {"num_frame":  
-1, "duration": 1540.0, "asr_model": "whisper_large_v2", "words": "視  
聴してくださって 本当にありがとうございます。", "scores": "0.81482410  
4309082", "recog_time": "1693384595.1947806", "sep_port": "5562"}  
[asr_node-2] [INFO] [1693386684.885362154] [asr_node]: {"num_frame":  
-1, "duration": 2240.0, "asr_model": "whisper_large_v2", "words": "最  
後まで視聴してくださって 本当にありがとうございます。", "scores": "0.  
7629714012145996", "recog_time": "1693386684.8828504", "sep_port": "5  
563"}
```

The Embodied Dialogue Awakens



◆言語の origin は実世界で使われていたところから

- もちろん「言語だけの世界」があることは否定はしない
- 言語は実世界におけるコミュニケーション・知識構築のツール

◆本来実世界性（身体性）と言語は強く結びつくもの

- 例えば動作動詞は人間が可能な動作に強く紐づく
 - 文化的背景で考慮される動作主にも紐づく（例: gallop）

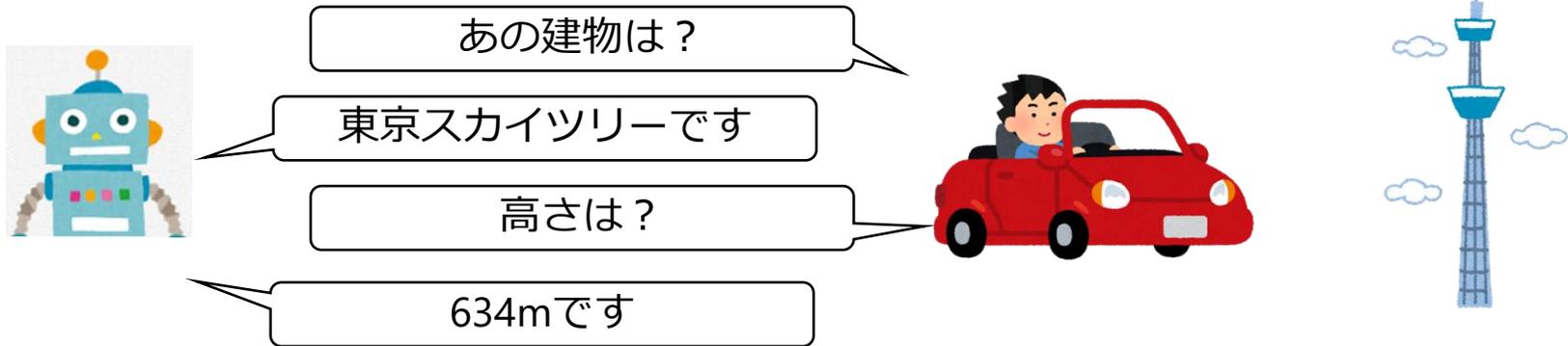
◆実世界と言語の関係について考える時

- GPTをはじめとする言語処理ツールの進展
- 特に動作+言語の研究はこれまであまりうまくいってない

ユーザが置かれた状況

- ◆人間との対話でのコンテキストを理解しなければならない
- ◆対話コンテキストとは「何を話したか」と「どういう状況にあるか」の両方を指す

- 対話コンテキストと実世界の事物との「接地」が必要
- 観測したものの何について話しているか「選択」が必要



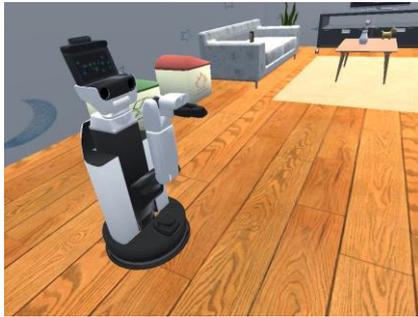
言語+マルチモーダル情報を使った状況の理解

実世界の観測と記憶



Planning における課題

動作（アクチュエーション）



言語（動作指示・動作説明）

キッチンのテーブルからウサギの置物を取ってきて



Yoshino et al., "Caption Generation of Robot Behaviors based on Unsupervised Learning of Action Segments." Proc. IWSDS, 2020.

◆自然言語による動作指示から動作を生成

- 曖昧性を含む言語指示を曖昧性のないアクチュエーション系列に落とし込むことができるか
- そのアクチュエーション系列は実施可能か
- 実世界における事物とのインタラクションを織り込んだ動作の設計 (planning) ができるか

SayCan で解けたんじゃないの？

Saycan by Google



TidyBot by Princeton & Stanford



Place clothes into the laundry basket

◆ SayCan などは動作と一意に変換できる動作プリミティブをあらかじめ用意し、言語 \leftrightarrow 動作プリミティブの問題を解いている

- 思い出して下さい、LLMは言語が入出力にある系列変換が得意

LLM がどのように動作の問題を解くか

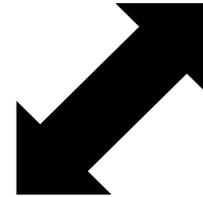
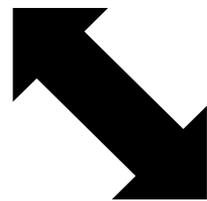
動作 (アクション)



言語 (動作指示・動作説明)

キッチンのテーブルからウサギの置物を取ってきて

ここは写像を仮定



基本行動単位 (細かい動作単位の系列)

MOVE from START to KITCHEN → PICK RABBIT ORNAMENT
→ HOLD RABBIT ORNAMENT → MOVE from KITCHEN to START

確実に動作可能な基本行動の系列予測問題に変換

LLM による行動プランニング

KO

You are a robot, and you have four actions: picking up something, holding the picked thing, putting a thing that you are holding, and moving from place A to place B. Currently, you can see that there is a box, and a cup of coffee is on the box. If someone said you "bring a box," what kind of action will you do?



As a robot with the given actions, if someone instructs you to "bring a box" and you can see a box in your environment, you would perform the following actions:

1. Pick up the box: Use your picking up action to grab and hold the box.
2. Hold the box: Utilize your holding action to maintain a firm grip on the box.
3. Move from place A to place B: Move towards the desired location, carrying the box with you.
4. Put the box down: Once you reach the designated place, use your putting action to set the box down in the desired position.

By following these steps, you would successfully bring the box as instructed.

実世界における動作と常識

◆常識 (commonsense) は当たり前過ぎると言語化されない

- 「コーヒーが上にのった物体を持ち上げると
コーヒーがこぼれる」という当たり前のことを
誰も (テキストとして) 書かない

◆しかしこれまでロボットの課題で詰まってきたのは こうした実世界における**当たり前の知識**

◆LLM により確実に出来ることは増加

- ただラストワンマイルが遠いのも事実 (**ラスト5cm問題**)
- 知識 \leftrightarrow 事物の接地も課題

LLM と推論

推論

ごちそうさまでした

ごちそうさまでと言う
↓s after
ご飯を食べ終える
↓s before
食器を片付ける → ケチャップを片付ける

行動候補
食器を持って来る
ゴミを捨てる
...
ケチャップを片付ける

机の上のケチャップを片付けましょうか？

◆この推論をすることは LLM に可能か？

◆LLM は仮説推論みたいなことはできるのか？

- [S5-P04] 家庭内ロボットの気の利いた行動の実現に向けて

Dialog System Technology Challenge (DSTC)



◆2013年から開催される対話システム関連シェアードタスク



General Chair:
Koichiro Yoshino
RIKEN, Japan



Workshop Co-Chair:
Yun-Nung (Vivian) Chen
National Taiwan
University, Taiwan



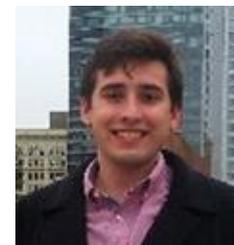
Workshop Co-Chair:
Paul A. Crook
Meta, USA



Publication Co-Chair:
Satwik Kottur
Meta, USA



Publication Co-Chair:
Jinchao Li
Microsoft, USA



Publicity Co-Chair:
Behnam Hedayatnia
Amazon, USA



Publicity Co-Chair:
**Seungwhan (Shane)
Moon**
Meta, USA

Yoshino, Koichiro, et al. "Overview of the Tenth Dialog System Technology Challenge: DSTC10." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).

◆世界における問題意識とは？

- MOD: Internet Meme Incorporated Open-domain Dialog
- Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations
- SIMMC 2.0: Situated Interactive Multimodal Conversational AI
- Reasoning for Audio Visual Scene-Aware Dialog
- Automatic Evaluation and Moderation of Open-domain Dialogue Systems

Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations

◆非構造化データへの対話的アクセス

- 案内情報を記述した文書を利用した対話
- タスク対話的な評価

◆音声対話の利用

- 認識誤りへの対応 (n-best)
- Joint goal は最高で 46%

Written Conversation

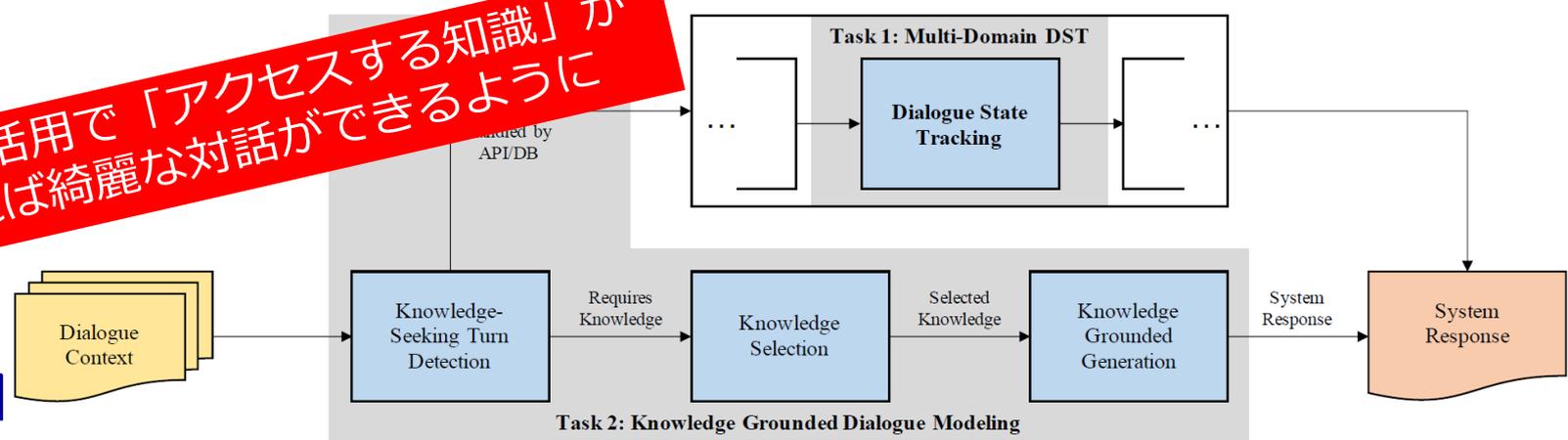
User I need a hotel in Fisherman's Wharf
Agent Is there a particular price range you are looking for?
User I'm looking in the expensive price range
Agent The Suite at Fisherman's Wharf may work for you
User Do you know how much the parking is?
Agent It would cost 25 dollars per day.

Spoken Conversation

User hi ummm i'm looking for a place at uhhh to stay at fisherman's wharf at a hotel in the expensive pressure engine
Agent sure let me see ok so there is one called the suites at fisherman's wharf is that something that would be interesting to you
User can you tell me how much parking coast
Agent sure okay this hotel charges twenty five dollars per day

Figure 1: Examples of written and spoken conversations

LLMの活用で「アクセスする知識」が定めれば綺麗な対話ができるように



SIMMC 2.0: Situated Interactive Multimodal Conversational AI

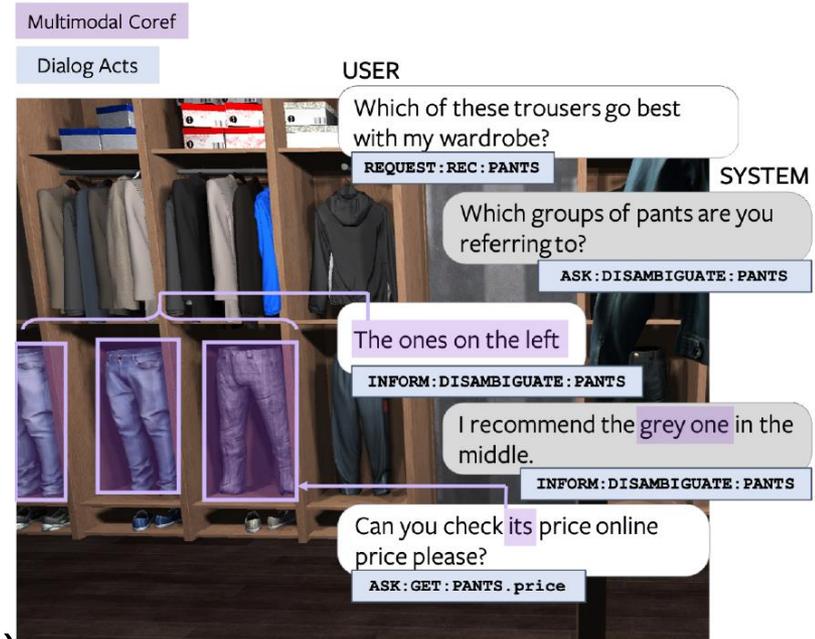
◆VR空間でのお買い物対話

◆マルチモーダル情報の利用

- 「それ」が何を指すか
- 実空間に接地された対話という文脈で非常に興味深い

◆いくつかの部分タスク

- マルチモーダル曖昧性解消 (95%)
- マルチモーダル共参照解析 (76%)
- マルチモーダル対話状態推定 (96%)
- 応答選択・生成 (R@5 で 95%程度)



世界への紐づけはいかにLLMへの入力にするかが課題

Reasoning for Audio Visual Scene-Aware Dialog



◆ビデオからの状況理解を行う対話

- Visual question answering のビデオ+対話コンテキスト
- ビデオと対話コンテキスト双方を正しく理解しないと答えられない対話

LLMが何に注視するかは
人が与えないといけない

- Q1: 女性はビデオ開始時に映っていますか？
- A1: はい、女性はシーン開始時に映っています。
- ...
- Q6: 女性は壁にあるものをずっと触っていましたか？
- A6: ずっとではなく10秒くらい、頭の上にあるものを直していました
- Q7: 彼女はビデオの中で何か言いましたか？
- A7: 全体を通して会話はありませんでした。

◆BLEU-4 0.385, LOUGE_L 0.539

タスクに合わせた注視

◆人間に求められるタスク=人間の注意・注視の考慮が必要

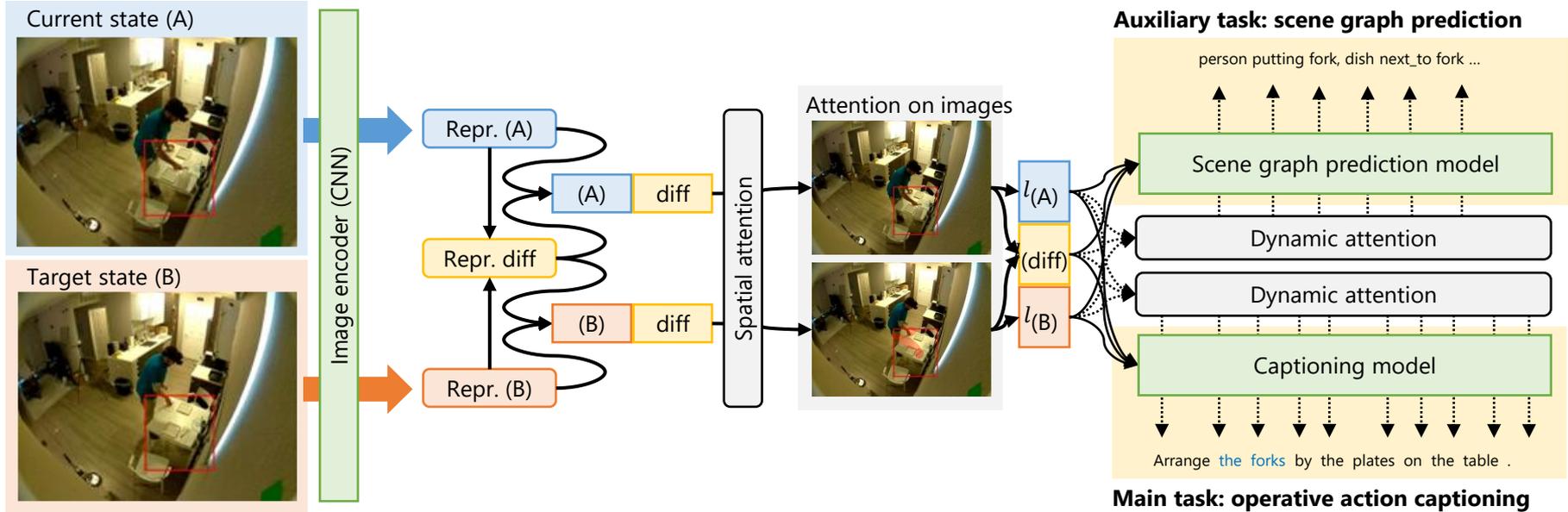
- 「何の情報に着目すればよいか」は自明ではない
- まだ **LLM にも出来ていない**ところ (のはず)

◆マルチモーダル事前学習モデルはまだ**これという事前学習の目的関数がきまっていない**ように見える

- CLIPのような対照学習では「テキストに書かれていない部分」に注視ができない
- EfficientNetのような物体名を扱うものでも画像中の物体間関係を扱うのは結構難しいのでは？
- お前の情報古い、これが定番、とかあったら教えて下さい

言語は MLM/CLM の目的関数がすごい
MLM/CLM で何ができないか考えるのがいいのでは？

言語生成の観点を与える



支援動作前の状態 (A)



支援動作後の状態 (B)



ベースライン
IH機を洗っている
 シーングラフ (全体)
 食洗機にコップを入れた
 シーングラフ (差分)
 食洗機のかごを引き出している

The Last Problem: 雑談対話

◆雑談の研究は本当に難しくてわかりません...

- 雑談の目的、評価指標、とにかくわからない...
- まずは評価指標を作らないとどうにもならないと思っている
- 単一リファレンスのBLEUはダメ**

◆重要だと思っているのは記憶について

- 前話したことを覚えている、前話した内容から深掘りできるというのは人間の重要な機能では
 - 重要な機能の割に今まで手薄だったのでは
 - 最近 prompt tuning の流れで記憶も注目されるようになってきていると思う
- [S4-P08] 経験に基づく知識の想起と深化を伴う対話システムの構築に向けて

◆対話研究が ChatGPT でいいかどうか

- ある部分はYes: 与えられた情報に対応したテキスト生成は非常に流暢なので活用すればよい
- ある部分はNo: LLM で解けない問題はいくらでもある

◆（当たり前だが） LLM にも得手不得手はある

- LLM が得意なタスクは適切に LLM に任せる
 - 専門家は「LLMがどういう仕組み・どういう背景理論を持っているか」をきちんと学んで活用すべき
 - 解くべき問題にフォーカスするため LLM を使う

◆ LLM の振る舞い自体にも謎は多い

- ただしかなり色々解明されつつある

研究で大事なこと

◆研究の目的をよく考えよう

- 自分の研究は工学なのか理学なのか
 - 最終的なゴールは何なのか
 - それが社会にどのような変革をもたらすのか

◆先人の知恵を借りよう

- 温故知新
- 大事なことは昔の論文に書かれている



べいえりあ @mr_bay_area · 8月25日

昔某所でインターンしてた時、当時NLP業界に革命を起こし続けてた某氏がチームの人々に「お前ら昔の論文読まなすぎ。昔の論文はアイデアの宝庫やぞ?」という話をされていて、なるほどなと思ったことがある。



16

117

9,158



◆病みの魔術に対する防衛術

- 研究活動は心に負荷がかかる
- 過度に負荷がかからないように
- ビックテックの研究にあまり振り回されても仕方ない
 - 今後10年で自分が解くべき問題をきちんと考える

最後に



- ◆ LLM で広がった対話研究の世界を存分に楽しみましょう
- ◆ May language model be with you