実世界における対話システム

理化学研究所ガーディアンロボットプロジェクト
知識獲得・対話研究チームリーダー
吉野　幸一郎

自己紹介

◆吉野 幸一郎 (Koichiro YOSHINO)

- 2005-2009 慶大 環境情報学部 学士（石崎研）
  - 2008 NTT-CS研実習生（東中先生）
- 2009-2015 京大 情報学研究科 修士・博士・PD（河原研）
  - 音声対話システム、音声認識、自然言語処理
  - 第一回 IPSJ-ONE 登壇者
- 2015-2020 NAIST 情報科学研究科 助教（中村研）
  - 音声対話システム、自然言語処理、DS
  - 2016-2020 JSTさきがけ「社会システムデザイン」
- 2020- 理研 GRP チームリーダー
  - 音声対話ロボット、自然言語処理

音声対話・自然言語処理の研究に従事（12年）

12年で何が明らかになったか

◆パターン認識の問題はかなり解けるようになった

- 音声認識
- 物体認識
- 質問応答
- 自然言語処理（形態素解析、構文解析）
- 機械翻訳（敢えて"機械翻訳"と書きますが...）

◆知能情報処理のプロセスは？

- 自由対話
- 翻訳、通訳
- 知識推論

「対話なんて研究しても仕方ない」

◆**2011年頃まで**

- だって何の役に立つの？
- 機械と会話できて何が嬉しいの？

After Siri

これからは対話インタフェースだ！うちでも対話インタフェースを作ろう！

◆**最近**

- だって何の役に立つの？
- 機械と会話できて何が嬉しいの？

なぜこうなってしまったのか？

A new hope:
深層学習の登場

◆音声インタフェースに対する期待の高まり

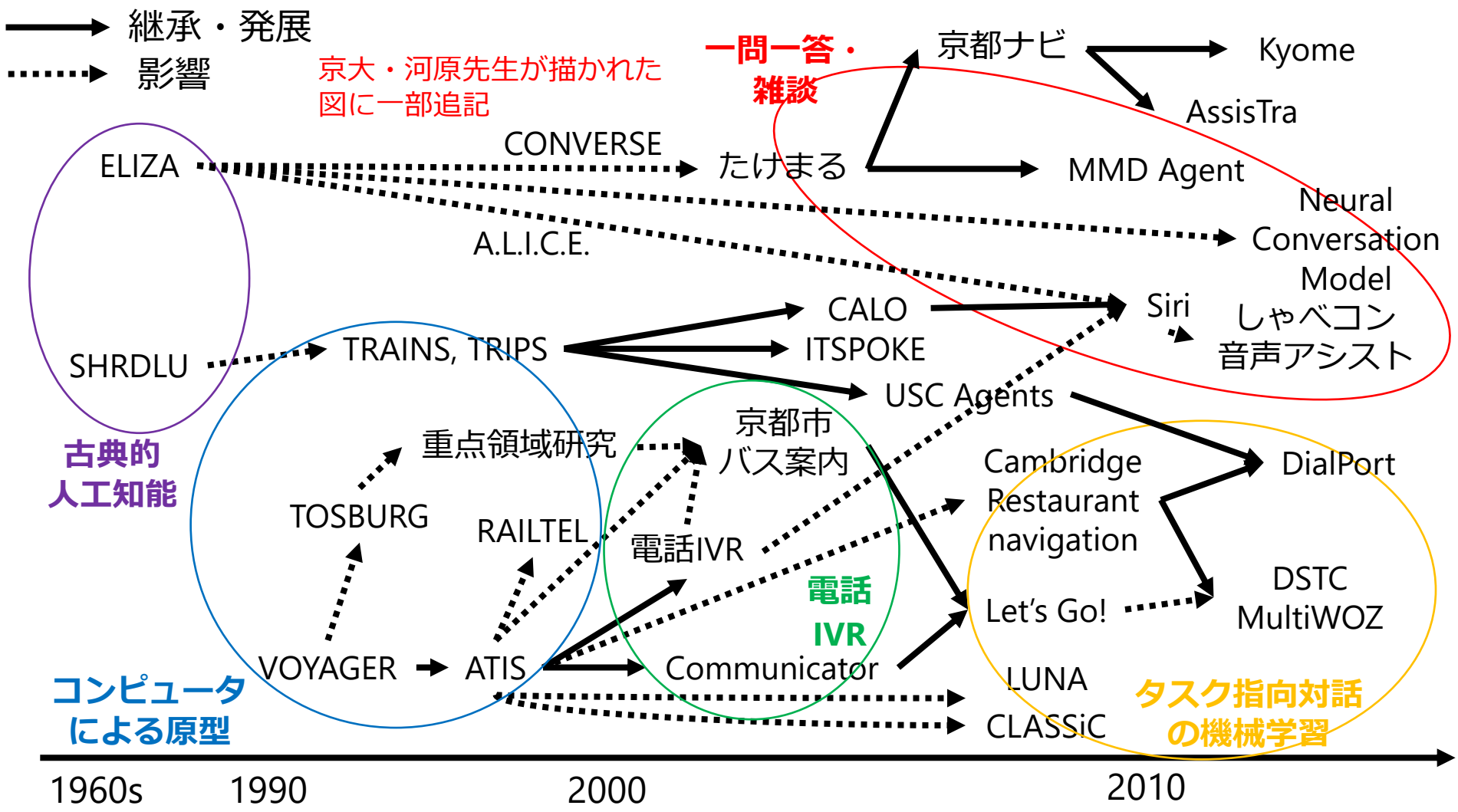
- 音声認識や言語処理などの基盤技術の精度向上
- いくつかのキラーアプリの登場（スマートフォン）

◆特に音声認識の精度向上は音声インターフェースへの期待を抱かせた

- Microsoft が人間レベルの書き起こし精度を実現 (2016)
- Google 音声認識 (API) の精度・使いやすさ（素晴らしい）

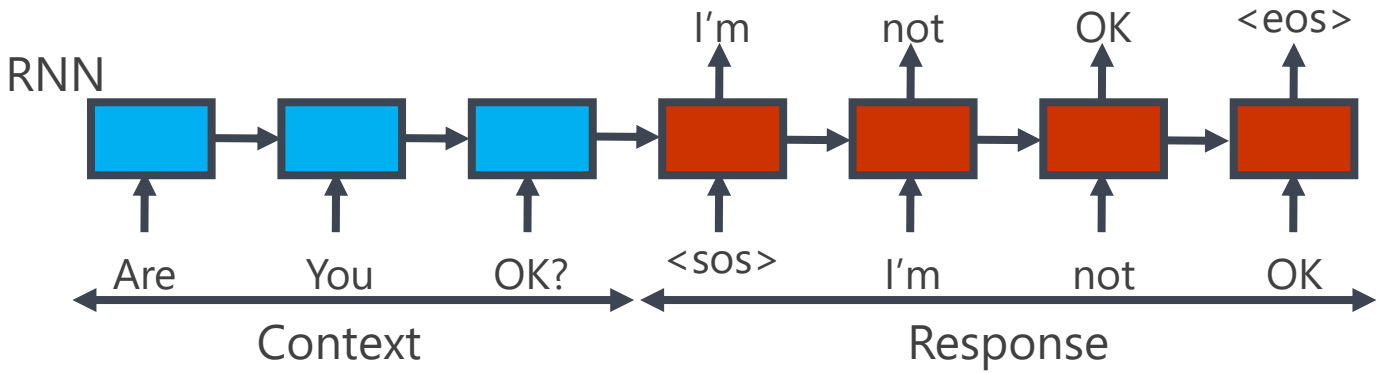
音声認識ができるなら対話できる！
チャットボットを作ろう！

音声対話システム-開発の系譜-



E2Eニューラル対話システム

◆**Neural Conversation Model (NCM) 発話ー応答の対応を学習**

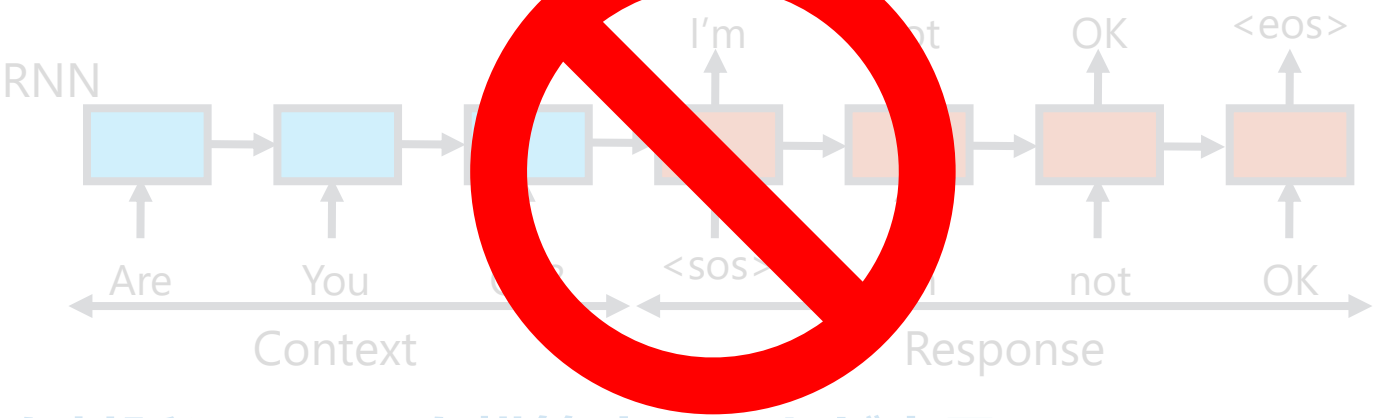


◆**誰でも対話システムを構築することが容易に**

- ドメインの学習データを用意しさえすればチャットボットができる
- 問題定義がシンプルでわかりやすい
- 誤差関数はクロスエントロピー誤差（応答中の単語予測）

The Neural-Network Strikes Back:
E2Eニューラル対話の問題

◆Neural Conversation Model [Vinyals+ 15] 発話ー応答の対応を学習



◆誰でも対話システムを構築することが容易に

◆相手の発話を理解した上で発話を生成しているわけではない

◆入出力の対応を取る非線形関数を学習しているに過ぎない

- 観測事象を接地して推論する過程がない
- **対話モデル学習の目的関数≠言語モデル学習の目的関数**
- 結局対話の目的を決めて対話システムを作らないと使われない

何のための対話研究か

◆**その研究が『何の役に立つ』か、『何を明らかにする』か**

- ●解くべき問題に合わせた最適化を行う
  - ●e.g., クロスエントロピー誤差で対話応答を生成することに（工学的、科学的に）どのような意図があるか
    - **●雑談対話の『雑談』は未定義**

何のための対話か（対話インタフェースか）という問題をよく考える

なぜ対話（言語）研究を行うか？

◆情報システムに人間の意志を伝え、仕事をさせるために、プログラミング言語や種々の約束を持った指令、ハードウェア操作など、いわゆるヒューマンインタフェースが開発されて来ているが、**人間の意志を伝える道具としてもっとも大切なものが自然言語**であることは間違いない。言語はすべての人が特別な訓練を経なくても使え、自分の意志を非常に微妙な所まで伝えることができる道具である。
「**自然言語処理（編: 長尾 真） まえがきより**」

◆音声対話によるコミュニケーションは、文字が発明されるよりはるか昔の人類から（あるいは動物でも）行われており、生き延びるのに不可欠なものであった。～**人間にとって音声対話は原始的・根源的なもの**である。これに対して、**現在のコンピュータは、～音声対話の能力に関しては人間と比べてきわめて低い。**
「**音声対話システム (著: 河原 達也, 荒木 雅弘)　まえがきより**」

対話研究の目的

◆役に立つ対話インタフェースを作る
- ●何ができれば役に立つのか？
- ●対話ゴールの定義が必要
  - ●レストラン案内、ドキュメント検索
- **●Return of the Task-Oriented Dialogue**
  - ●知識ベース（オントロジー）への接地と理解

◆知能情報処理のプロセスを明らかにする
- ●対話における情報授受のプロセス
  - ●対話における理解・基盤化
- ●基盤化空間での推論
- **●The Embodied Dialogue Awakens**
  - ●実空間の事象に紐づいた理解と推論

タスク対話

◆タスクゴールを設定し対話状態に従い対話する

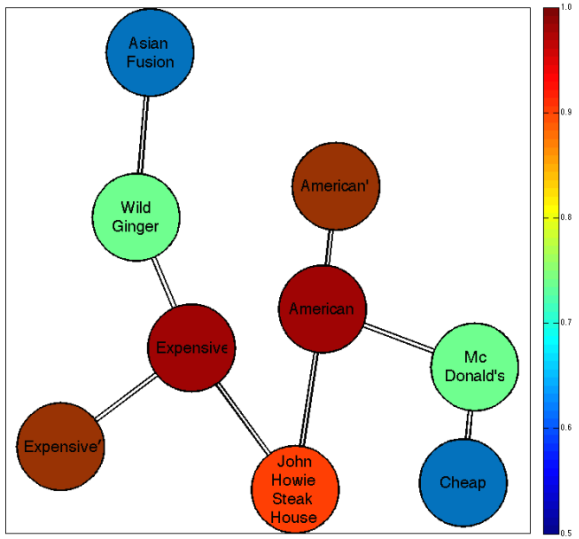
- ●タスクと関わる対話行為、スロットなどのアノテーション
- ●対話状態中で対話のゴールが明示的に定義される

Speaker	Utterance	Dialogue Frame
System	Hello, welcome to the Cambridge restaurant system. You can ask for restaurants by area, price range or food type. How may I help you?	Welcomemsg
User	The cheap restaurant and it should be in the west part of cow	Inform(pricerange=cheap,type=restaurant, area=west)
System	There are 12 restaurants in the west of town. What type of food do you want?	Request(slot=food) Inform(count=12) Impl-conf(area=west)
User	Any	Inform(food=don'tcare)
…	…	…

オントロジーから他の知識源へ

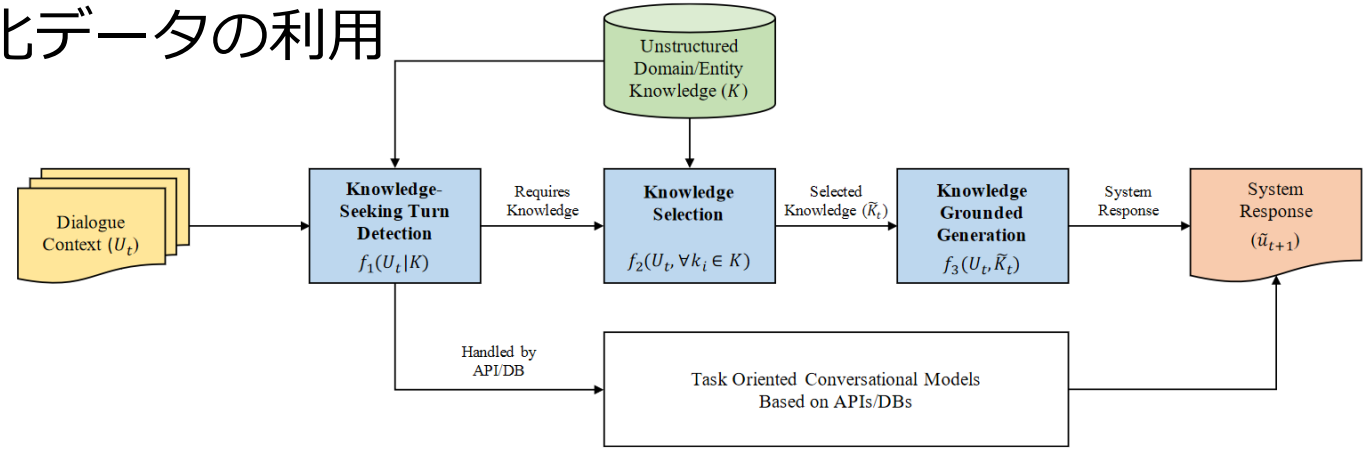
◆定義されたオントロジー
→オープンな知識リソースの利用

- 定義されたオントロジー (DSTC1-4)
- 知識グラフ上での推論 (DSTC5)
- 非構造化データの利用 (DSTC9)



◆ドキュメント知識と対話の接地（DSTC10）

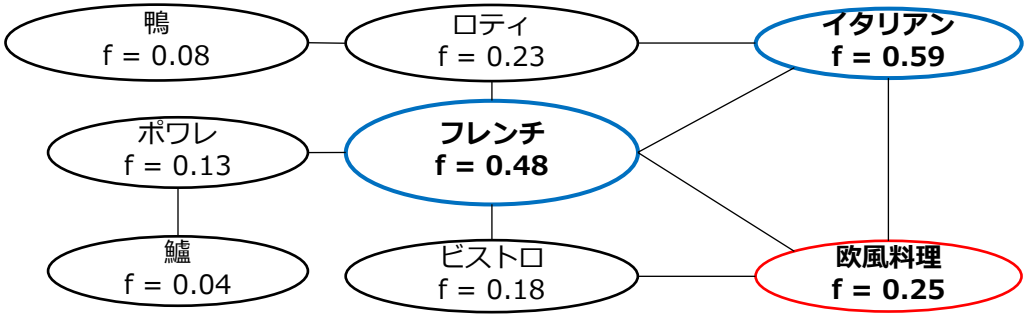
- 非構造化データの利用



知識グラフ上の推論

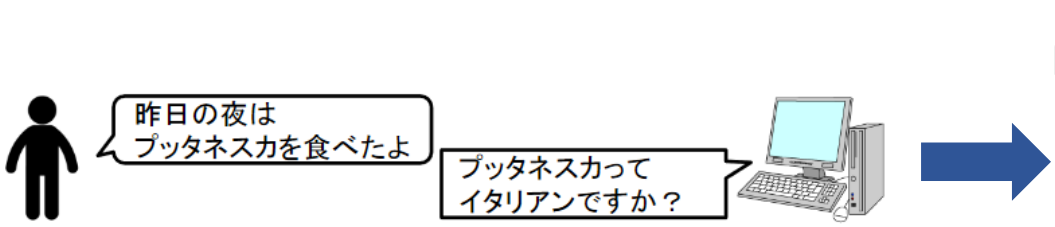
◆構造化された知識の利用 [Murase19]

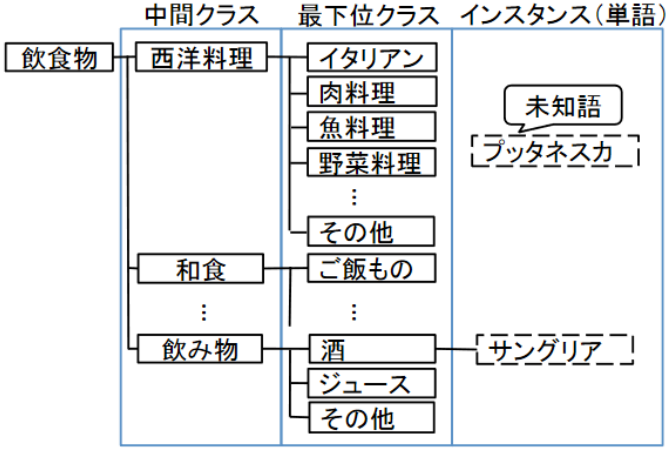
- ラベル伝搬、ベイジアンネットワークなどの利用



- 対話コンテキストを利用しない場合湧き出しも多い

◆対話による知識グラフの拡張 [大野18]





実世界知識の利用: Visual dialogue

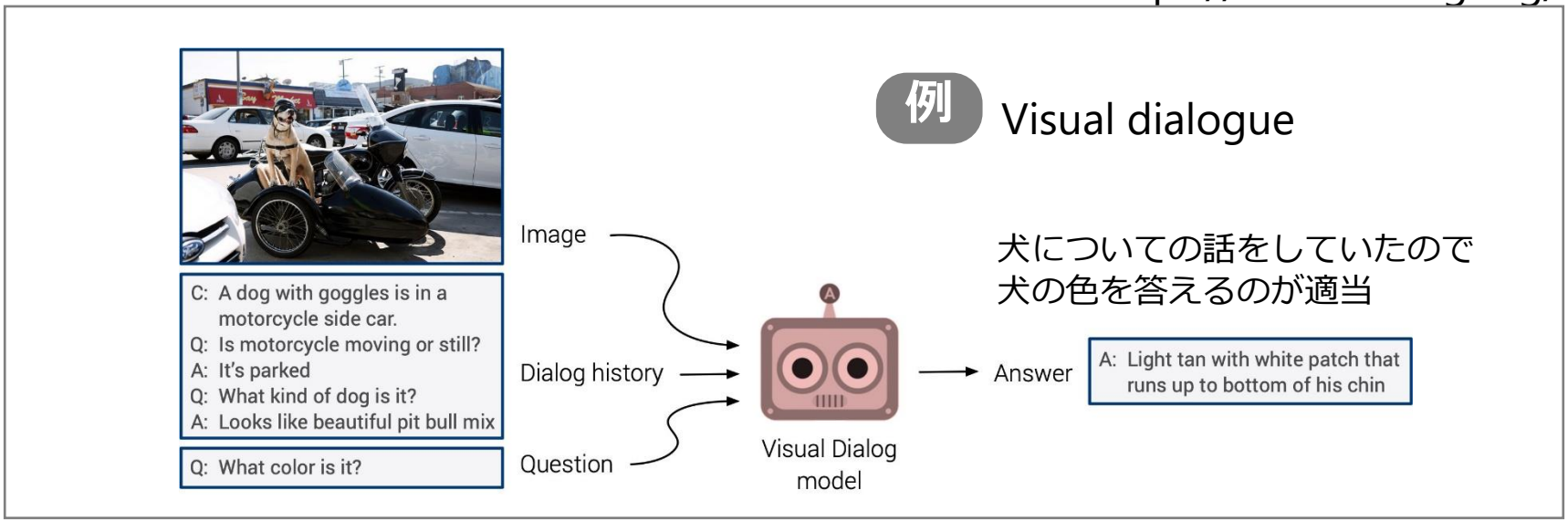
◆**画像・動画の内容と対話履歴を使った対話 [Alamri19] (DSTC7)**

- VQA の拡張 → 単なる VQA と比較して対話履歴で反応が変わる

◆**画像、質問、コンテキストそれぞれの利用が必要**

- 知識と観測の対応を取る必要がある

https://video-dialog.com/
https://visualdialog.org/



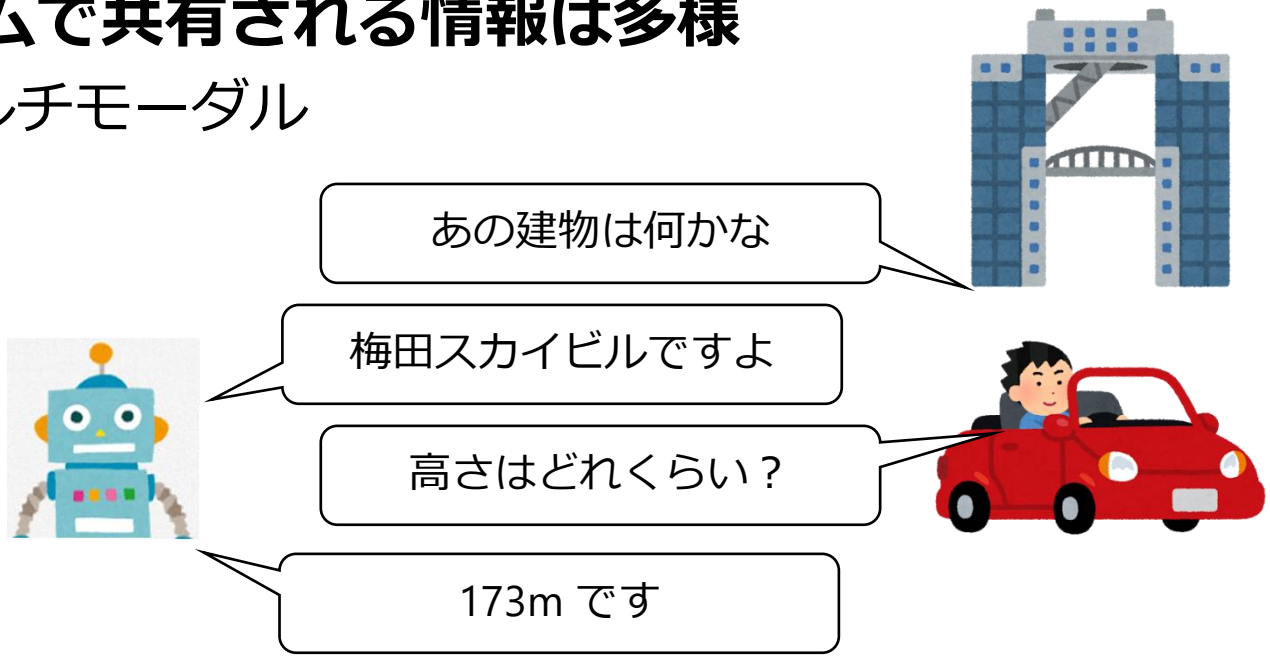
ユーザが置かれた状況を考える

◆**車載カメラ・マイクでの対話 [Misu14]**

- 人間が見ているものを検出し対話コンテキストとして利用
- 実世界情報を用いた対話・質問応答など
- 地図情報なども利用可能

◆**ユーザとシステムで共有される情報は多様**

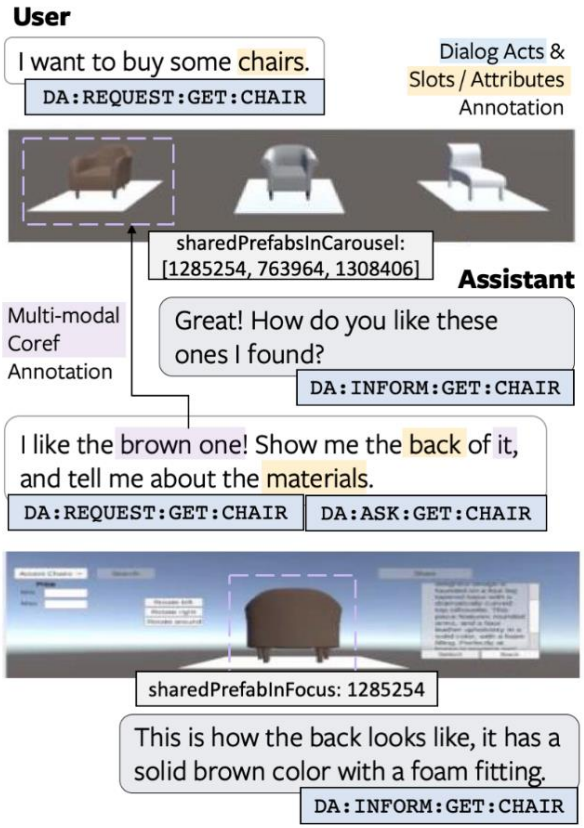
- 多くの場合マルチモーダル



仮想空間と対話

◆バーチャル空間での買い物を想定した対話 [Beirami20] (DSTC9, 10)

- ユーザが指示する物体が具体的にどれか確認をしながら対話を進める
- 言語・画像を用いたユーザの意図理解
- 対話による聞き返し
- 必要な意図に応じた発話生成



https://github.com/facebookresearch/simmc

DSTC

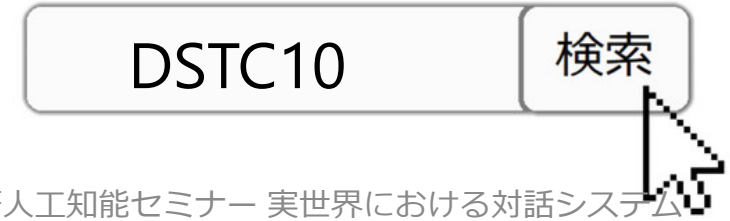
◆対話システムに関連する国際コンペ（毎年5月スタート）

- ●今年も9月末まででチャレンジが開催中
  - ~~●今からの参加も歓迎します~~
- ●様々な対話タスクが提案されコンペ形式でスコアを競う
- ●過去のチャレンジで利用されたデータの配布

https://sites.google.com/dstc.community/dstc10/past-challenges

DSTC9 (2020)

- Four Parallel Tracks
  - Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access: Task Description, Resources, Overview Paper
  - Multi-domain Task-oriented Dialog Challenge II: Task Description, Resources, Overview Paper
  - Interactive Evaluation of Dialog: Task Description, Resources, Overview Paper
  - SIMMC: Situated Interactive Multi-Modal Conversational AI: Task Description, Resources, Overview Paper
- Challenge Organizers: Yun-Nung (Vivian) Chen, Chulaka Gunasekara, Luis Fernando D'Haro, Seokhwan Kim, Abhinav Rastogi
- Venue: DSTC9 Workshop @ AAAI-21
- Challenge Website



閑話休題: Embodied dialogue

◆人間は実世界の情報を観測し、そこから推論し、自身の主体に基づいて行動をしている

- 対話システムにこの機能を実現できるか？

◆一人称的主体をどう構築するか

- 「与えられた問題を解く」から「自分で設定した問題を解く」へ
- 人間の一人称的主体はどのような形で存在するか？心とは？

自己紹介

◆吉野 幸一郎 (Koichiro YOSHINO)

- 2005-2009 慶大 環境情報学部 学士（石崎研）
  - 2008 NTT-CS研実習生（東中先生・現名大）
- 2009-2015 京大 情報学研究科 修士・博士・PD（河原研）
  - 音声対話システム、音声認識、自然言語処理
  - 第一回 IPSJ-ONE 登壇者
- 2015-2020 NAIST 情報科学研究科 助教（中村研）
  - 音声対話システム、自然言語処理、DS
  - 2016-2020 JSTさきがけ「社会システムデザイン」
- **2020- 理研 GRP チームリーダー**
  - **音声対話ロボット、自然言語処理**

最近の取り組み

理研GRP

◆理化学研究所は文部科学省所管の自然科学系総合研究所

◆ガーディアンロボットプロジェクト (GRP) は
ロボットの基礎研究を目的として2020年に発足

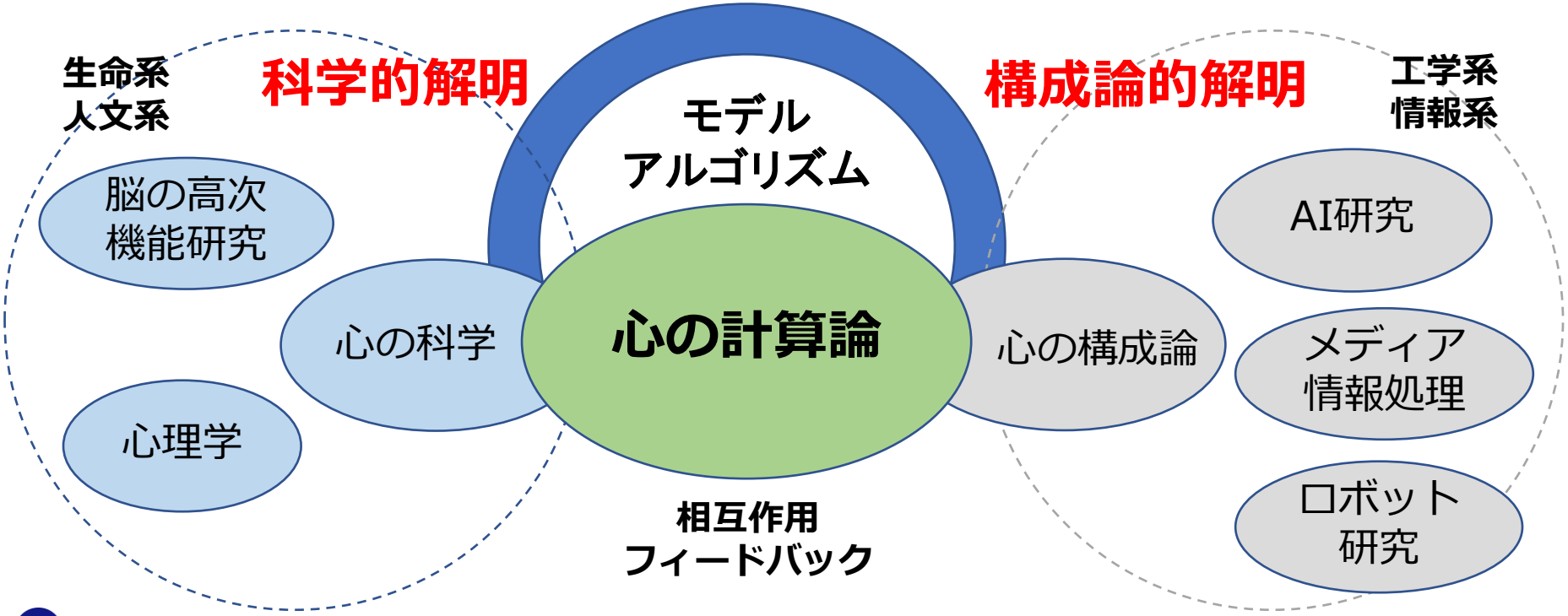
◆社会的ミッション

- 特定の人間と生活を共にし、表情や動作、環境を観測、認知して対話し、人間の行動を予測して、さりげなく支援する
  - ユーザの主体性を重視した支援
  - 人間中心システム

GRPのミッション

◆学術的ミッション＋工学的ブレークスルー

- 人間のこころのメカニズム（知覚、認識、記憶、思考、注意、運動制御、感情、社会性）を計算論的に解明し、ロボット実装を通じて構成論的に実証する。



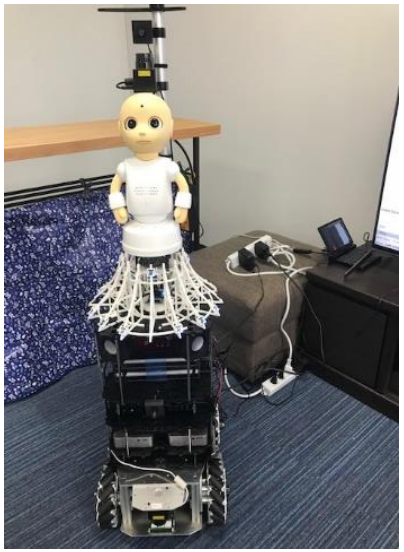
現在開発しているロボット

◆外骨格型

◆装着して立ち上がり支援を行う

◆椅子などの環境型との連動

◆自律型



◆自律動作して支援タスクを行う

◆アームなどで物をつかむ

◆対話型

◆人間に近い顔の自由度を持つ

◆表情表出などの研究を行う

開発中の試作機

※動画のため、このファイル形式では表示できません。
下記URLよりご確認ください。

https://youtu.be/W3c7mzC1G60

研究の目標

◆目的・主体にもとづいて動作する対話ロボットの構築

- ロボットの存在意義からロボット自身が主体的に目的を決定
  - 見回りの仕事を優先する、ユーザの依頼を聞く
  - 対話の目標、動作の目標から主体的な決定
- 目的に応じたロボット自身の行動最適化
  - 何の情報を取りにいくか
  - どういう行動にどういう情報が必要か

◆目的・主体にもとづいた知識の取得・選択・推論

- 必要な情報の取捨選択
  - 知識を取得するための行動も選択する
- 記憶の形態、記憶や経験から知識を構築する
- 構築した知識を世界知識と組み合わせる

研究における課題

◆インタラクションを通じて**知識を蓄える**
- 動作的知識（身体性に基づく知識）
- 対話を通じた知識
- 世界知識（因果・相関関係）

◆インタラクション・行動に必要な**知識を利用する**
- どの**知識を利用するか選択**する
- どの**観測を利用するか選択**する（多様なセンサー情報）

◆必要な**知識の推論**を行い**行動を決定**する
- 既に得た知識から必要な新たな知識を推論する

まさに一般人工知能の機能 (The Last Problem)

なぜ一般人工知能が難しいか

◆**Common Sense の壁**

- 常識を**システムの身体性・環境にあわせて**記述する必要

◆**常識を獲得するためのコストが高い**

- ある状況をロボットに経験させれば学習させることはできる
- ただし**あらゆる状況をシステムに学習させる**ことは非現実的

◆**常識的推論がそもそも難しい**

- 常識を**明示的に教える**ことの難しさ（常識は説明されない）

◆**記号接地（シンボルグラウンディング）の問題**

- 知識表現中の記号と、それが指す実世界の実体との接続

実世界に紐づいた知識推論システムが必要

実世界の対話システムに必要な能力

◆人間の生活空間で動作するシステムには**想像力**が必要
- ユーザから与えられる**曖昧な指示**
- 生活空間に関わる様々な知識
- 自身の行動によって何がもたらされるか**予測する能力**

◆これらを全てシステムに経験させるのは現実的ではない
- **テキストとして記述された知識との融合**

実世界の対話システムに必要な能力

◆観測結果から**推論を行い状況を認識する**能力
- 観測内容と知識の接続と推論

◆認識した状況から**自身が行うべき行動を決定する**能力
- ユーザの要求が曖昧な場合は聞き返す
- 将来のユーザ満足度を最大化する → 気の利く行動・先回り
  - ユーザの要求によっては断ったりする

◆自身の意図・行動を**正しく把握し説明する**能力
- 言語を用いた説明能力
- Explainable AI

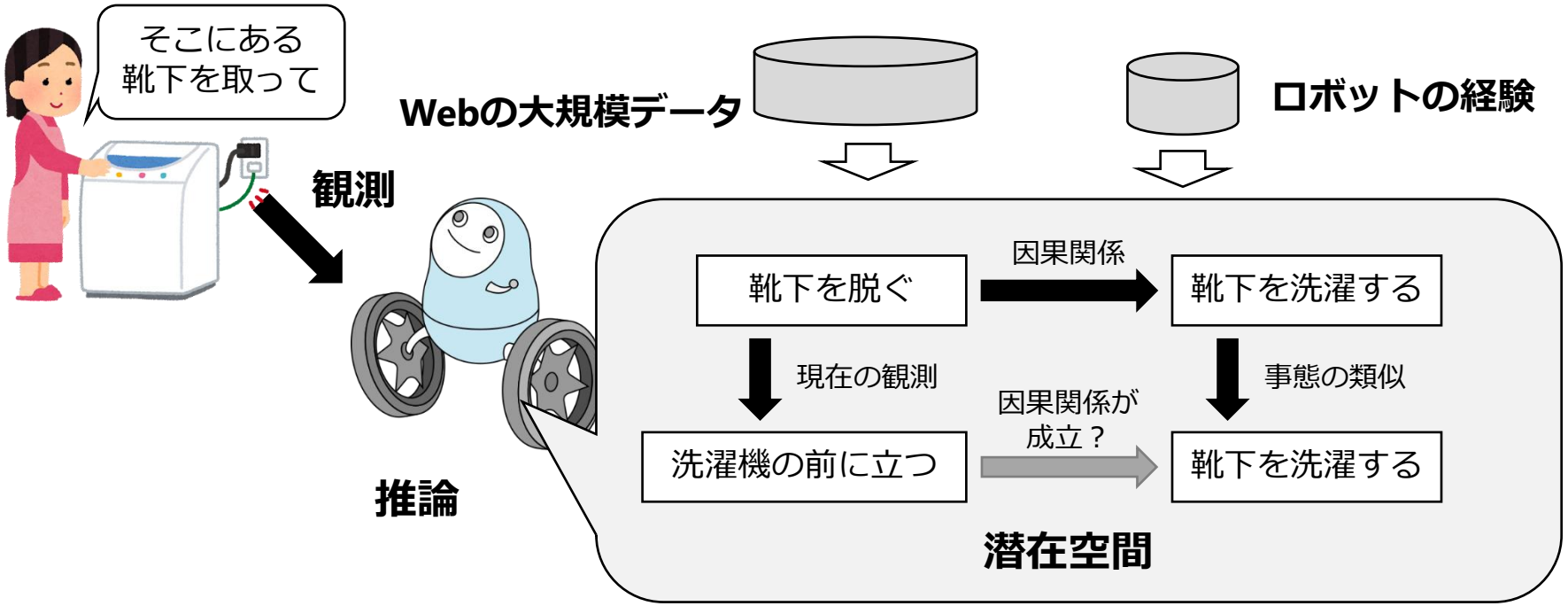
観測結果から推論を行い状況を認識する

◆観測したイベントの**知識に対する紐づけ**

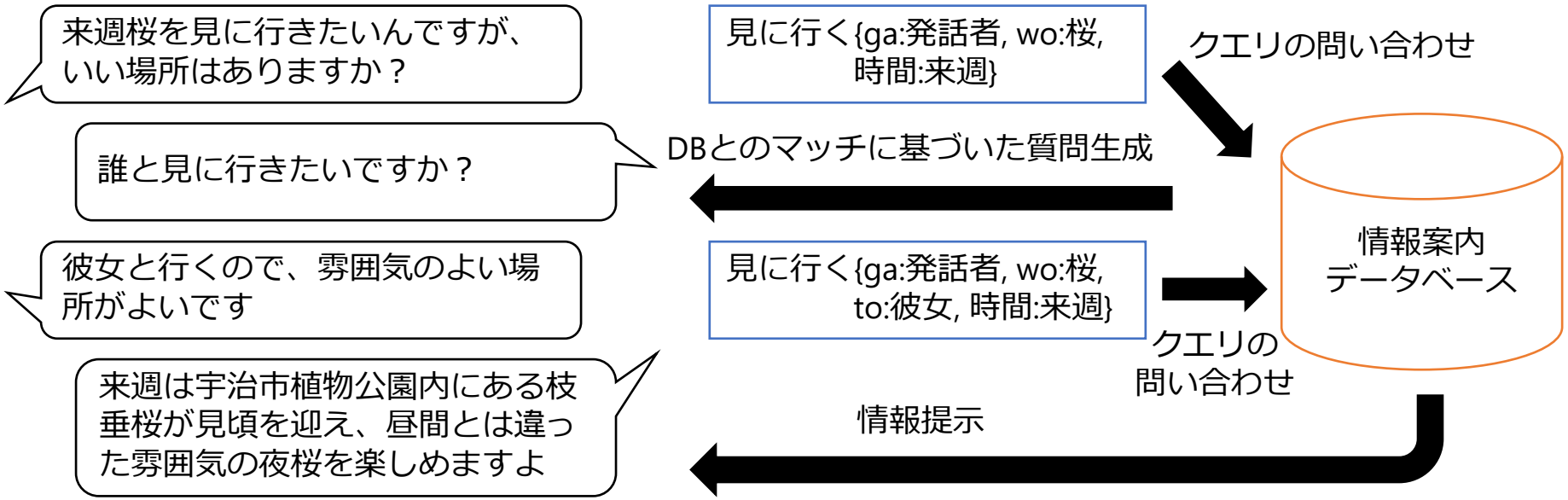
- 述語（動詞・事態性名詞）を中心としたイベントの記述

◆**イベント間関係を用いた推論**

- 因果関係、支持、攻撃などの関係利用、スパースネスの解消



これまでの取り組み: イベントに基づく理解



◆意味解析（項構造解析）に基づく発話情報の理解

- 項構造解析も漸進的に行う

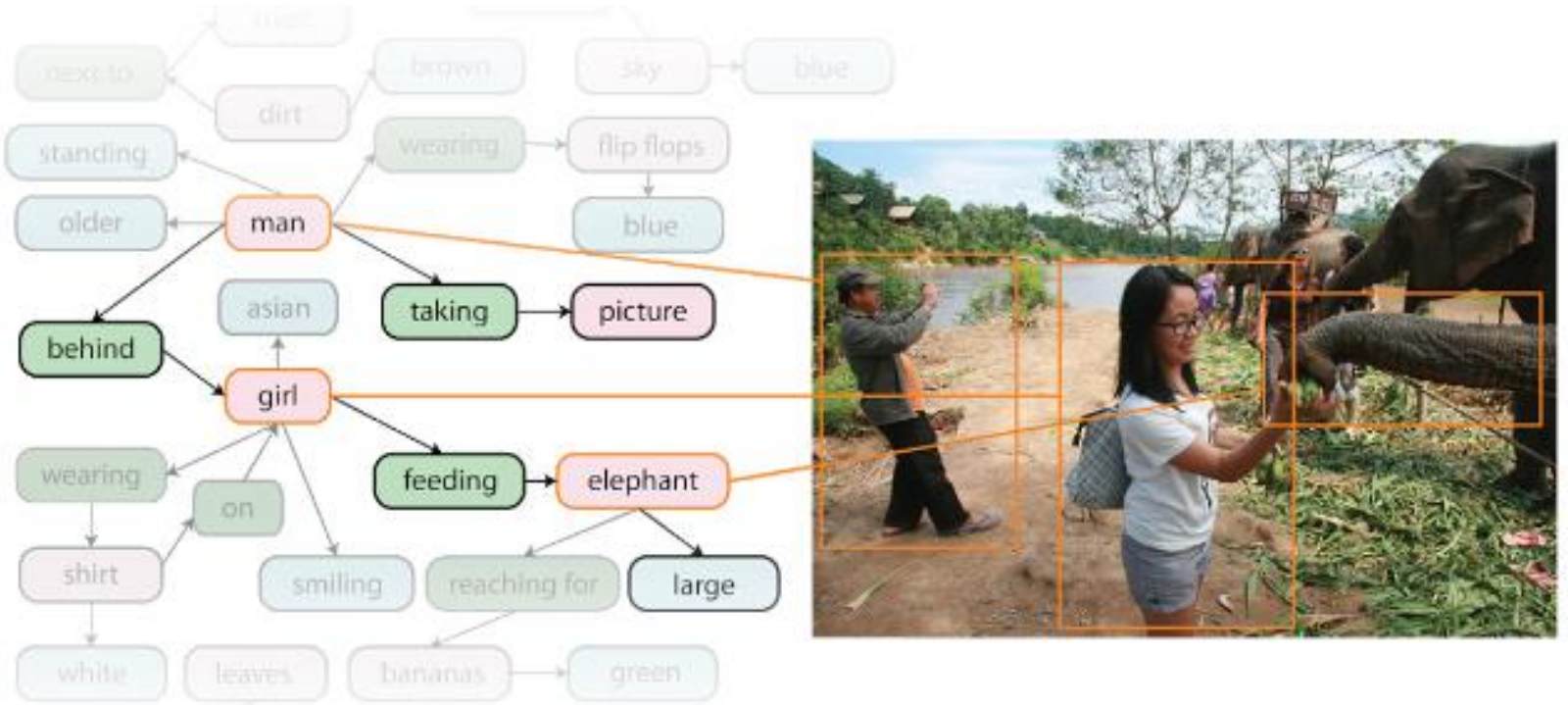
◆項構造 → 述語を中心とする事態表現に着目

- イベントを基盤とする言語理解フレームを動的に生成

他のモーダル情報からのイベント理解

◆シーングラフなどの併用

- ●視覚情報から得られる情報の利用
- ●イベント単位でマルチモーダル情報の融合が必要



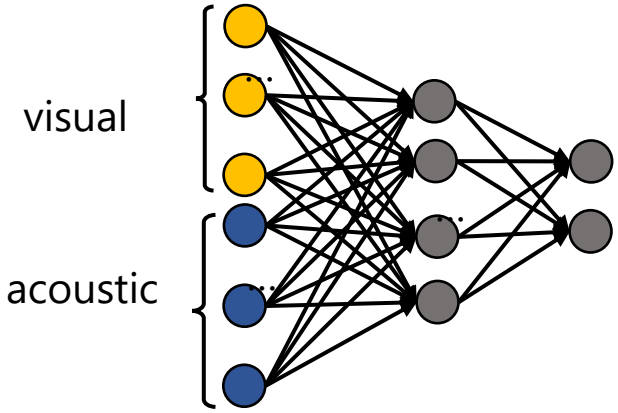
マルチモーダル情報の統合

◆なぜマルチモーダル情報の活用が容易になったか？

- ニューラルネットワークの功績

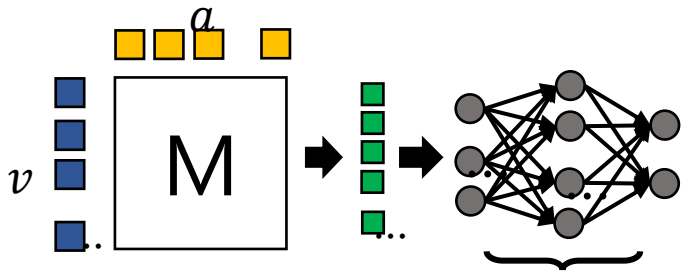
1. **同一ネットワークでのインテグレーションが容易**
   - 異なるモダリティからの情報を同じ目的関数で学習できる
   - 目的に応じた様々な組み合わせ方が可能
2. **特徴量設計がネットワーク任せ**
   - 目的関数に応じた特徴量をネットワークが学習してくれる
   - 基本的タスクから事前学習することも可能（e.g., 言語モデル）
3. **マルチタスク学習が容易**
   - 複数の目的関数を与えることが容易
   - 目的関数の組み合わせ方、学習方法も様々

マルチモーダル情報の利用法



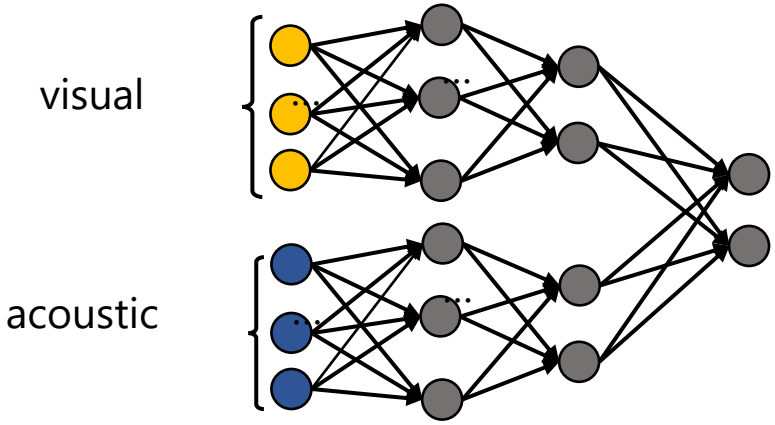
◆**Early fusion**

- 特徴量レベルでの結合



◆**Tensor fusion**

- 特徴量の外積を利用



◆**Late fusion**

- 識別器レベルでの結合

◆**これら以外にも Graph-fusion, Hierarchical fusion など様々な利用法**

◆**ただし基本的に同じ NN の枠組みで適用可能**
＝泥臭い特徴量設計が不要

- MFCC（音響特徴） → 音声波形
- SHIFT（画像特徴） → ピクセル値

マルチモーダル情報結合

◆マルチモーダル情報を用いた嘘検出タスク [Nguyen19]

◆Hierarchical fusion と Tensor fusion の併用が有効

- モダリティごとの重みとモダリティ間の関係双方が重要

		Positive label (lie)			Negative label (truth)		
	Accuracy	Precision	Recall	F1 score	Precision	Recall	F1 score
Single (visual)	0.4928	0.4095	0.3525	0.3789	0.5434	0.6026	0.5714
Single (acoustic)	0.5378	0.4747	0.5000	0.4870	0.5920	0.5673	0.5794
Multi early	0.5342	0.4701	0.4836	0.4768	0.5869	0.5737	0.5802
Multi late	0.5468	0.4794	0.3811	0.4247	0.5829	**0.6763**	0.6261
Multi hierarchy (Tian 2015)	0.5378	0.4733	0.4713	0.4723	0.5879	0.5897	0.5888
Multi TFN (Zadeh 2016)	0.5036	0.4216	0.3525	0.3839	0.5511	0.4571	0.4997
Multi Hierarchical TFN	**0.5863**	**0.5304**	**0.5000**	**0.5148**	**0.6258**	0.6538	**0.6395**

認識した状況から行うべき行動を決定する

◆多様な情報源の利用と対話プロセスの活用

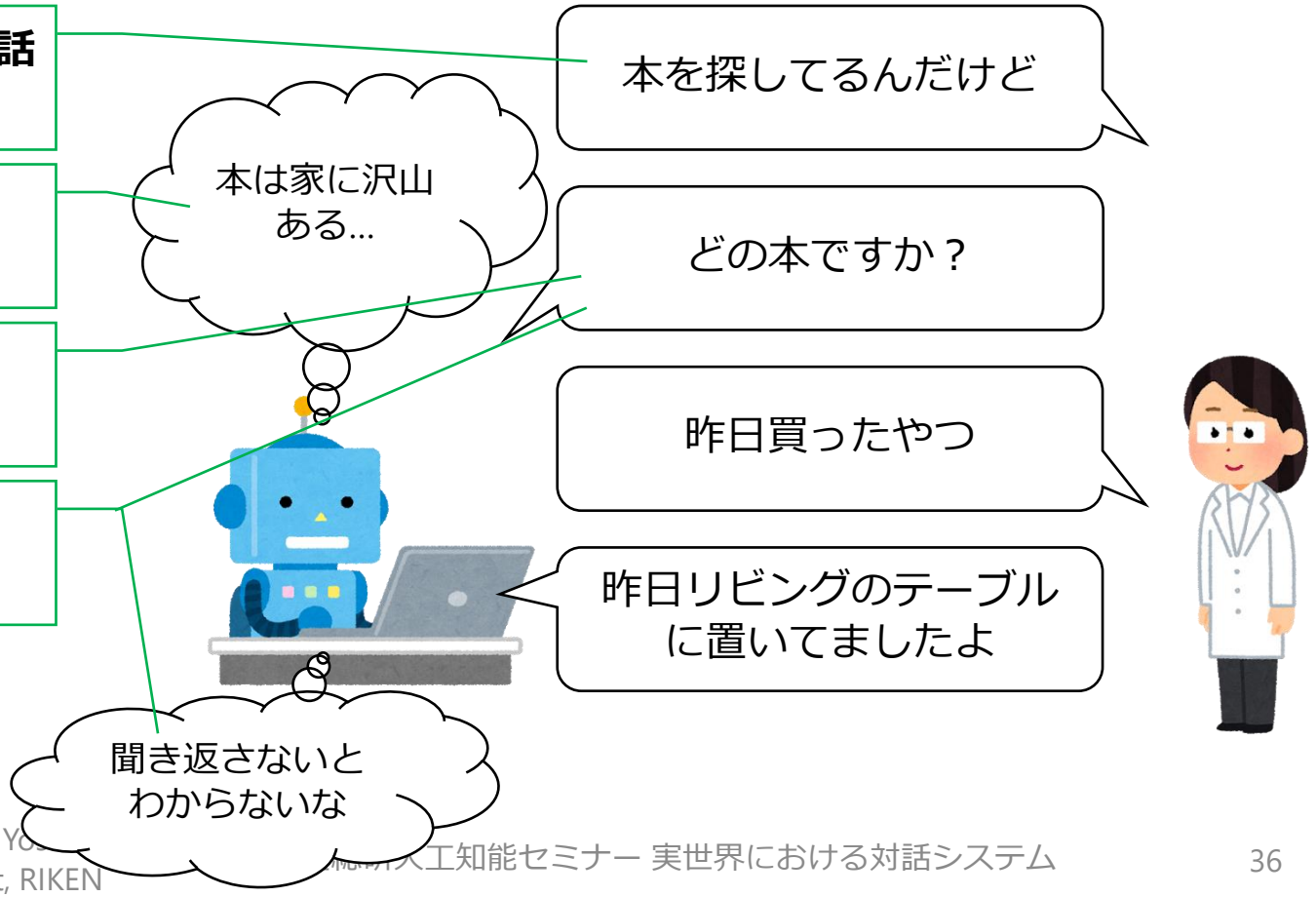
●実世界に紐づいた情報の対話的な利用

課題1: 曖昧なユーザ発話意図の認識

課題2: アクセス情報候補の推定

課題3: 曖昧性解消のための質問生成

課題4: 行動決定と生成



これまでの取り組み:
言語を用いた仮説推論

◆発話に含まれる意味内容の理解・仮説推論

被告人は車を運転していましたか？/
被告人の呼気中からアルコールは検出されましたか？

情報探索対話 (質問と応答)

はい

「被告人は車を運転していた」かつ
「被告人の呼気中からアルコールが検出された」
ので**「被告人は飲酒運転をしていた」**

合理的な論証の構築

◆重み付き仮説推論エンジンを用いた対話による論証構築

◆深層強化学習を用いた質問戦略の最適化

深層強化学習による質問戦略効率化

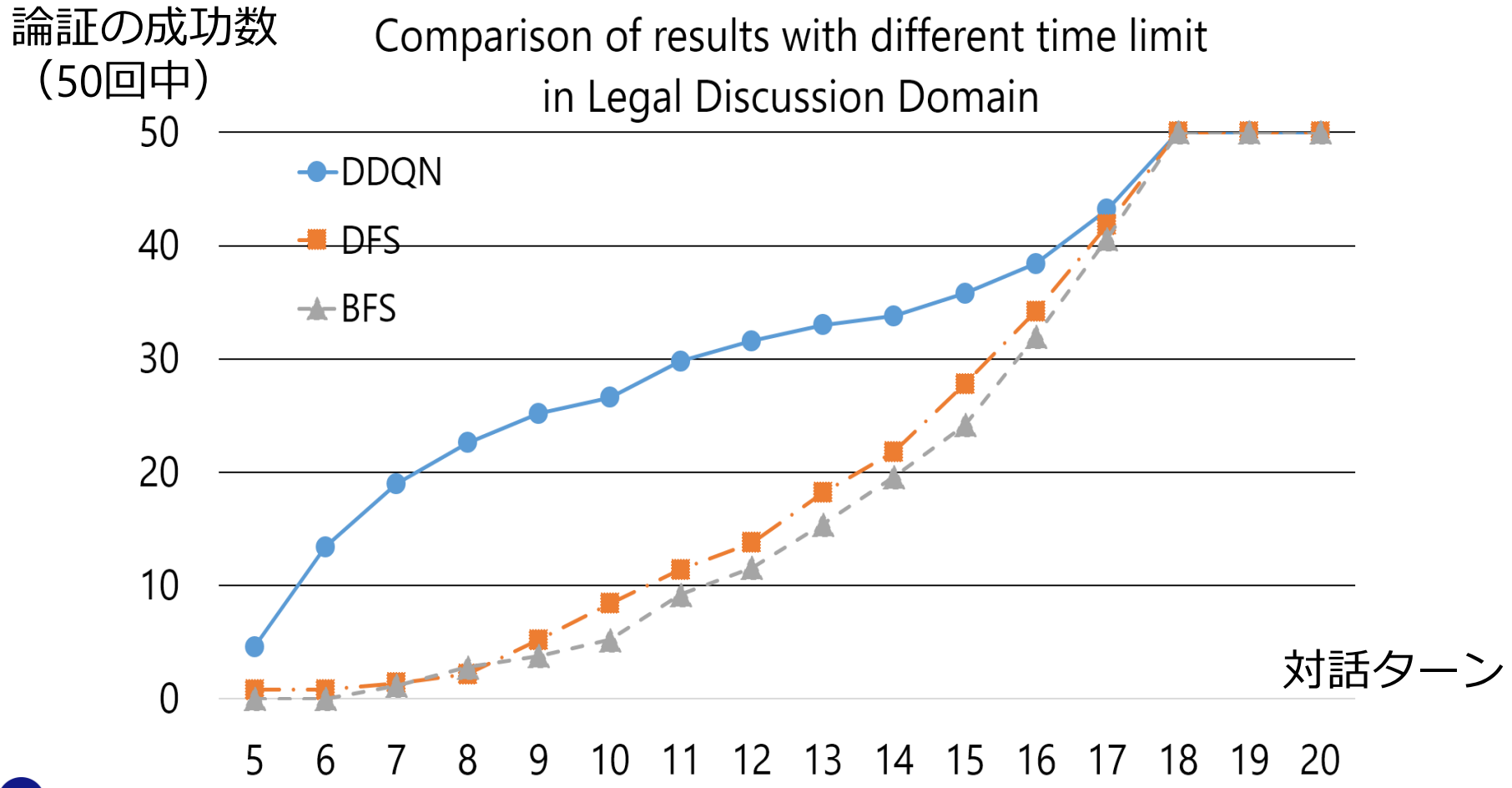
◆既存の事実収集戦略よりも効率的に事実を収集可能

- ●殺人容疑の少年が無実である根拠を収集する対話
- ●事実を収集する対話の過程をマルコフ決定過程でモデル化

turn	話者	発話	合理性
1	質問者 回答者	通りを横切った女性は殺人事件を目撃していないのですか？ はい、その女性は殺人事件を目撃していません。	0.1
2	質問者 回答者	通りにいた老人は少年が「殺してやる」と言ったのを聞いていないのですか？ はい、その老人は少年が「殺してやる」と言ったのは聞いていません。	0.4
3	質問者 回答者	その老人は嘘をついていますか？ わかりません。	0.4
4	質問者 回答者	では少年は中腰になって背の高い男性を刺していないのでしょうか？ わかりません。	0.4
5	質問者 回答者	通りを横切った女性は少年が父を刺すのを目撃していないのですね？ わかりません。	0.4
6	質問者 回答者	では少年は凶器となったナイフを購入しましたか？ いいえ、少年は凶器となったナイフを購入していません。	0.7

深層強化学習による質問戦略効率化

◆深層強化学習によりより早く効率的な事実を収集可能

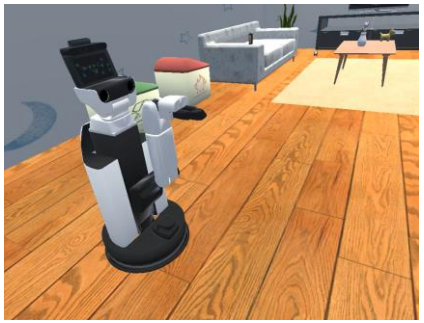


自身の意図・行動を正しく把握し説明する

◆実世界とのインタラクション

- ロボットが実空間で何を観測したか
- ロボットが実空間で何を行ったか

Input



Output

Picked up the rabbit ornament from the kitchen table.

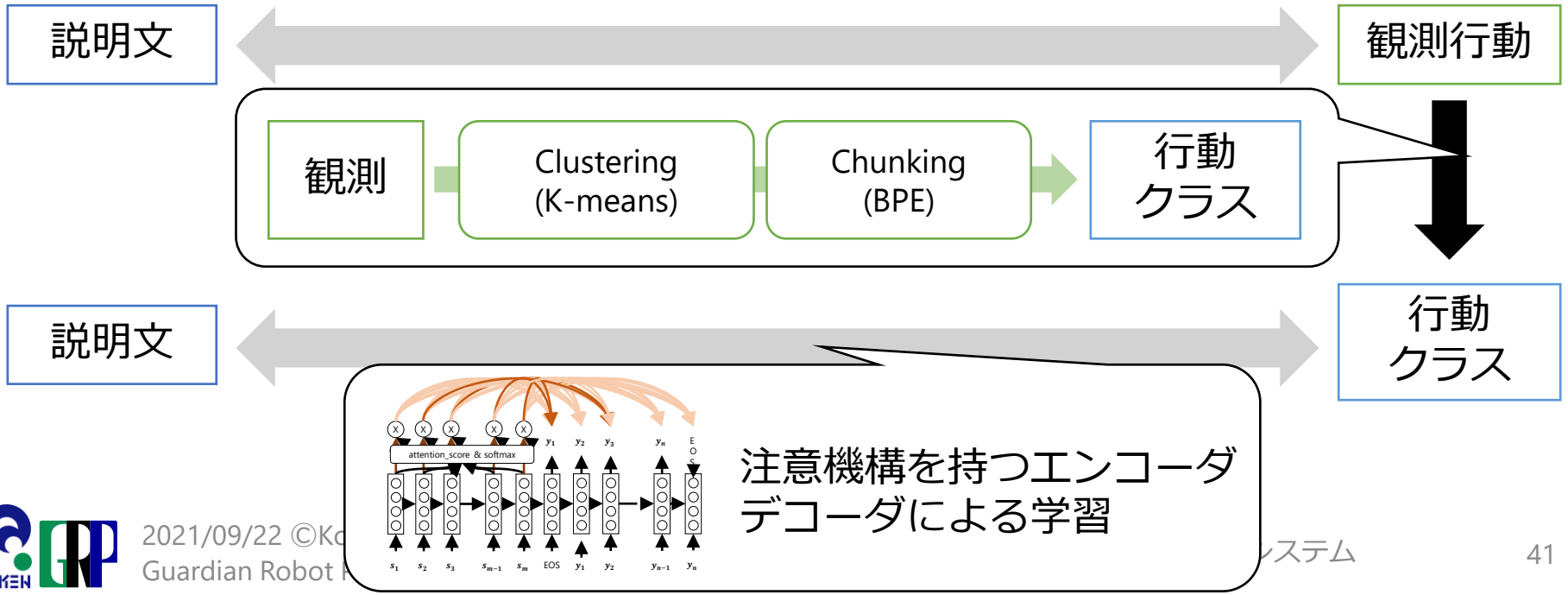
◆ロボット自身の意図をユーザから理解できるようにする

◆ロボット自身が何をやったか、意図と違ったことをやっていないか自覚する能力

マルチモーダル入力と説明文の end-to-end学習

◆入力: ロボットが一人称視点で観測可能なもの
- ●各関節の動作軌跡
- ●一人称視点からえられる動画情報

◆出力: 動作を説明するテキスト
- ●自然言語による動作結果の説明



生成結果の主観評価

Method	a	b	c	d	e	a-c
Vanilla	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
Explicit	10.6%	17.2%	37.2%	28.3%	6.7%	65.0%
Implicit	3.9%	5.0%	35.0%	56.1%	0.0%	43.9%
Hybrid	7.2%	8.9%	33.3%	43.9%	6.7%	49.4%

a. 生成文が適切にロボットの行動を説明している
b. 生成文はおおよそロボットの行動を説明している
c. 生成文は動作を適切に説明しているが、物体等に誤りがある
d. 生成文は文法的に正しいが内容は誤っている
e. 生成文は意味をなしていない

少量の学習データ（1000サンプル）でも対応学習可能

生成した説明の例

Ref.	Pick up the teapot on the floor (床のティーポットを拾って)	
Vanilla	Repeating function words (てててててててててててて)	e:9
Explicit	Pick up the sauce on the floor (床のソースを取って)	a:7, b:1, c:1
Implicit	Pick up the sauce on the table (テーブルの上のソースを取って)	b:1, c7, d:1
Hybrid	Pick up the sauce placed on the floor (床にあるソースを取って)	a:7, c:2

◆同じ色の物体が識別できていない

- 画像認識モデルと組み合わせると
  もう少し精度がよいはず

◆特に Attention を使うものは「取る」を過生成しがち

現在地とこれから

◆できるようになったこと

- ニューラルネットワークを用いた効率的なマルチモダリティ統合
  - 様々な識別問題の劇的な精度向上
- 生成ネットワークによる多様な出力
  - 言語、音声、画像、動作計画

◆今後取り組むべき課題

- 多様な入力を用いることができるシステムのインテグレーション
  - 多様なシステムが特定の目的に最適化され動作する
- 最適化目標の自発的設定
  - あらかじめ定義された最適化目標・手法に依らないシステム
- 理解・接地・推論
  - 何をもって理解できたとするのか
  - 実空間への接地、知識空間への接地、これらを用いた推論

実世界で動作する
対話システムを目指して

◆研究することは沢山ある
- ●実世界における接地、経験からの学習
- ●一般人工知能の問題はほぼ未解決
- ●NNによって一研究者が扱える範囲は各段に広がった

◆何のためのシステムか
- ●目的関数が解きたい問題を表現しているか考える

◆人間の情報処理プロセスはどうなっているのか
- ●解明されていないことも沢山ある
- ●記憶のメカニズム、知識選択の方法、推論のプロセス

◆作ってみて考える
- ●構成論的解明
- ●なぜ作ったか？を常に問いかける

参考文献

- Ma, Yi, et al. "Knowledge graph inference for spoken dialog systems." *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.
- Kim, Seokhwan, et al. "Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access." *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2020.
- Murase, Yukitoshi, Yoshino Koichiro, and Satoshi Nakamura. "Associative knowledge feature vector inferred on external knowledge base for dialog state tracking." *Computer Speech & Language* 54 (2019): 1-16.
- 大野航平, et al. "対話を通じた未知語のクラス獲得に向けた暗黙的確認の提案." 人工知能学会論文誌 33.1 (2018): DSH-E_1.
- Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., ... & Parikh, D. (2019). Audio visual scene-aware dialog. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7558-7567).
- Misu, T., Raux, A., Gupta, R., & Lane, I. (2014, June). Situated language understanding at 25 miles per hour. In Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL) (pp. 22-31).
- Crook, P. A., Poddar, S., De, A., Shafi, S., Whitney, D., Geramifard, A., & Subba, R. (2019). SIMMC: Situated Interactive Multi-Modal Conversational Data Collection And Evaluation Platform. arXiv preprint arXiv:1911.02690.
- Krishna, Ranjay, et al. "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations." *International Journal of Computer Vision* 123.1 (2017): 32.

参考文献

- Nguyen, T. T., Yoshino, K., Sakti, S., & Nakamura, S. (2020). Policy reuse for dialog management using action-relation probability. IEEE Access, 8, 159639-159649.
- Tanaka, Shohei, et al. "Conversational Response Re-ranking Based on Event Causality and Role Factored Tensor Event Embedding." *Proceedings of the First Workshop on NLP for Conversational AI*. 2019.
- 勝見久央, et al. "論証対話システムにおける情報探索対話戦略の最適化." 人工知能学会論文誌 35.1 (2020): DSI-D_1.
- Takano, W., & Nakamura, Y. (2015). Statistical mutual conversion between whole body motion primitives and linguistic sentences for human motions. The International Journal of Robotics Research, 34(10), 1314-1328.
- Yamada, T., Matsunaga, H., & Ogata, T. (2018). Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. IEEE Robotics and Automation Letters, 3(4), 3441-3448.
- Yoshino, K., Wakimoto, K., Nishimura, Y., & Nakamura, S. (2020). Caption Generation of Robot Behaviors Based on Unsupervised Learning of Action Segments. In Conversational Dialogue Systems for the Next Decade (pp. 227-241). Springer, Singapore.
- Shridhar, Mohit, et al. "Alfred: A benchmark for interpreting grounded instructions for everyday tasks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020.