

実世界にグラウンドされた 自然言語理解の これまでとこれから

理化学研究所AIP
JST さきがけ

栗田修平



革新知能統合研究センター
Center for Advanced Intelligence Project



第7回 Language and Robotics 研究会

まず最初に



この場を用意してくださったみなさま、特に、
LangRobo研究会幹事の一人ひとりに厚く御礼申し上げます 🙇



Shuhe Kurita

PhD of Informatics, Kyoto University

RIKEN AIP, JST PRESTO

New York University (Visiting Researcher)

CURRENT INTEREST:

Grounded Visual / 3D and Language Understandings.

shuhe.kurita@riken.jp

 @ShuheKurita

SELECTED PUBLICATIONS

- ScanQA: 3D Question Answering for Spatial Scene Understanding @ **CVPR2022**
Daichi Azuma (*), Taiki Miyanishi (*), Shuhe Kurita (*), Motoaki Kawanabe (*):Eq. Cont.
- Generative Language-Grounded Policy in Vision-and-Language Navigation with Bayes' Rule,
Shuhe Kurita and Kyunghyun Cho, @**ICLR2021**
- Reconstructing neuronal circuitry from parallel spike trains,
Ryota Kobayashi, Shuhe Kurita, Anno Kurth, Katsunori Kitano, Kenji Mizuseki, Markus Diesmann,
Barry J Richmond, Shigeru Shinomoto @ **Nature Communications (2019)**
- Multi-Task Semantic Dependency Parsing with Policy Gradient for Learning Easy-First Strategies,
Shuhe Kurita and Anders Søgaard, @**ACL2019**
- Neural Adversarial Training for Semi-supervised Japanese Predicate-argument Structure Analysis,
Shuhe Kurita, Daisuke Kawahara and Sadao Kurohashi, @**ACL2018**
- Neural Joint Model for Transition-based Chinese Syntactic Analysis,
Shuhe Kurita, Daisuke Kawahara and Sadao Kurohashi, @**ACL2017** Selected **Out-standing paper!**

目次

1. はじめに
 1. Embodied AI : 身体に基づいたAI
 2. Generative language grounded policy (GLGP) on VLN
 3. SayCan: Do As I Can, Not As I Say (Google Robotics & Everyday Robotics, 2022)
2. 言語の役割
 1. 名付けによる世界の分節 (F. Saussure)
 2. 言語による指示と実世界との対応付け (L. Wittgenstein)
3. 言語モデル
 1. 歴史 (Markov process, n-gram, RNN)
 2. PaLM & Big-BENCH
 3. Chain of thought
 4. 基盤モデルたち (省略)
4. グランディングと参照表現理解
 1. 参照表現を理解する
 2. テキストに紐付いた物体検出
 3. 3D世界に対応づいた言語理解
5. 言語を使って指示を出す
 1. SHRDLU (T. Winograd, 1971)
 2. ナビゲーション : Vision & Language Navigation (VLN)
 3. プランニング : AI2THORとSayCan
6. まとめ : Language & Robotics のこれから
 1. 言語指示が有効なのはどのような課題か？
 2. 言語理解技術がどのようにロボティクスを助けるか？

はじめに : Embodied AI – 身体に基づいたAI

R2R Nav. on Matterport 3D



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

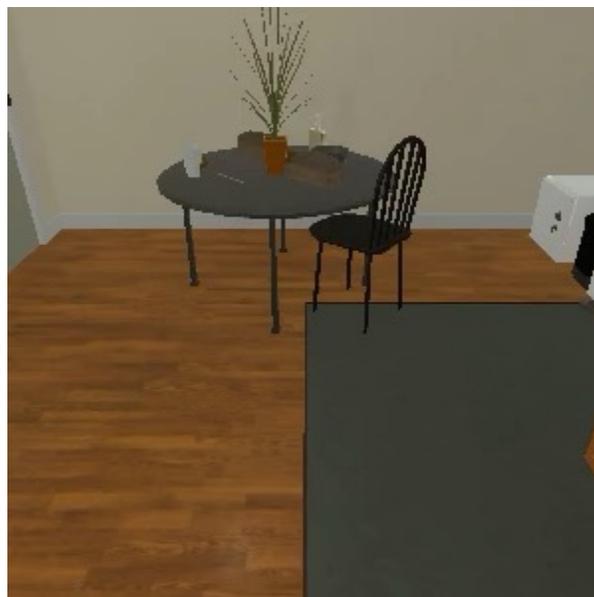
R2Rナビゲーションタスク

仮想のエージェントが写実的な仮想の室内環境中を指示文章に従って行動する。R2Rは視覚、言語および動作の情報を統合しなければ解くことが困難である。

- 仮想環境では、視覚・言語理解・動作の全てをモデルに包括的に学習させる
- 人間が身につけている常識的な知識を仮想環境データセットで再現できる
- 現実世界よりも優れた試行速度・多様性



ALFRED on AI2-THOR



ALFRED

仮想環境の内部で指示に従って簡単な家事をこなす。家具や家電などの一般的な知識、食材など登場する物体の常識的な性質、課題を解くために求められる操作の計画性、などが求められている。

To heat a cup as well as place it in the fridge.

Turn around and walk across the room and look up to face the cabinet above the left side of the microwave.

Pick up the cup from the cabinet above the left side of the microwave.

Step to the right and look down to face the microwave.

Heat the cup in the microwave and remove it.

Turn around and walk across the room and then turn right and step forward to face the fridge.

Place the cup in the fridge.

はじめに：Embodied AIへの関心の高まり

深層学習と3Dモデリング技術の発達により広い分野の国際会議で論文が出ている：

2017

R2R Navigation (Seq2Seq)

VLN

RoomNav [2017]

R2R [2017; CVPR2018]

2018

Speaker-follower + Beam decoding [NeurIPS2018]

Reinforced cross-modal matching [CVPR2018]

仮想環境上でのQA

EQA [CVPR2018], IQA [CVPR2018]

2019

EnvDrop + data aug. [NAACL2019]

Self-monitoring [ICLR2019]

Analysis on visual features of VLN [ACL2019]

FAST decoding [CVPR2019]

A joint pretraining for vision and language [EMNLP2019]

BERT for VLN [EMNLP2019]

Google Street View上でのVLN

Touchdown [CVPR2019]

複雑経路VLN

R4R [ACL2019]

対話的VLN

CVDN [RL2019] Help! ANNA [EMNLP2019]

2020

A joint pretraining with VLN [CVPR2020]

BabyWalk [ACL2020]

物体操作が可能なEAI

ALFRED [CVPR2020]

Habitat

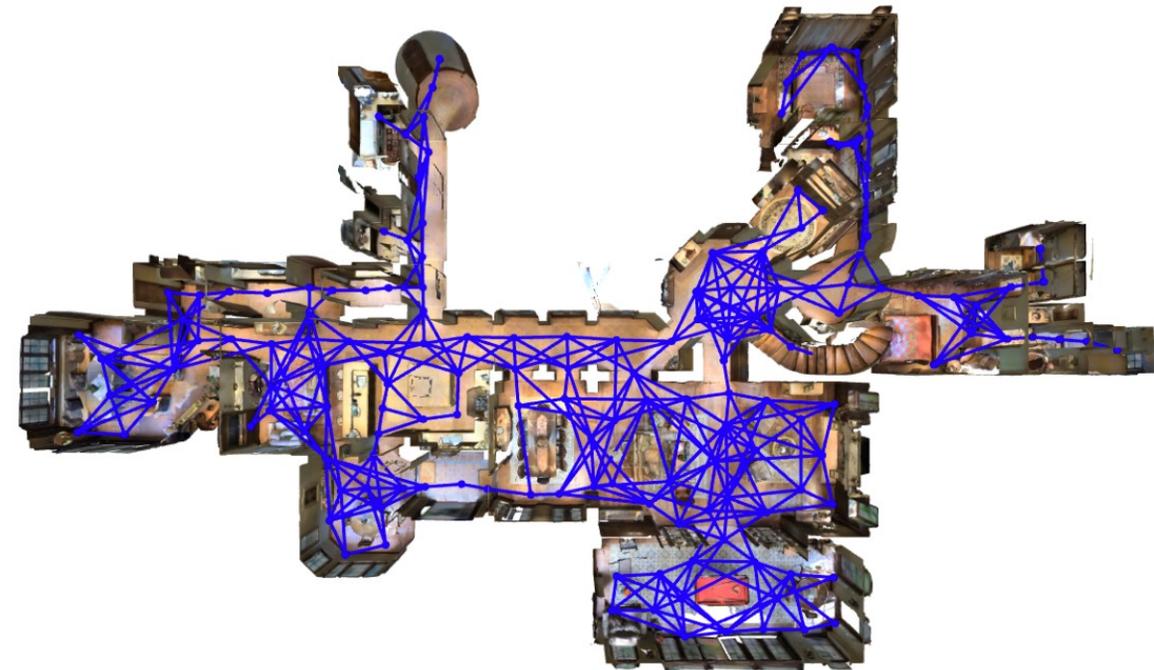
[ICCV2019]

はじめに：視覚と言語によるナビゲーション (VLN)



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

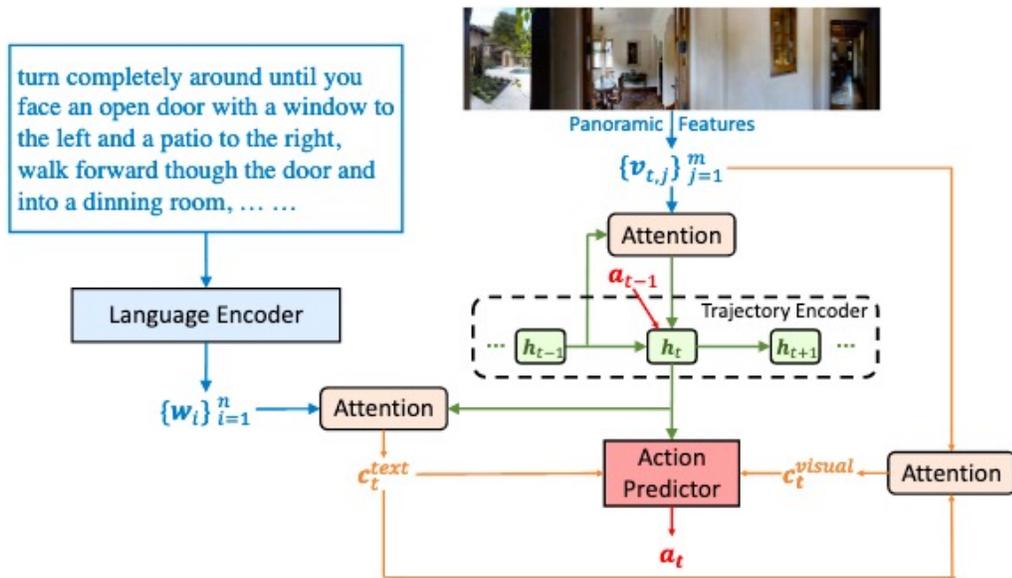
- A navigation task for a robotic agent to reach the goal place following the textual instruction.
- Photorealistic environment.
Based on real houses, corrected with 3D scans.
- The dataset was released in 2017.
- Possible application to the robotic navigation



VLNで言語モデルを利用した解決法



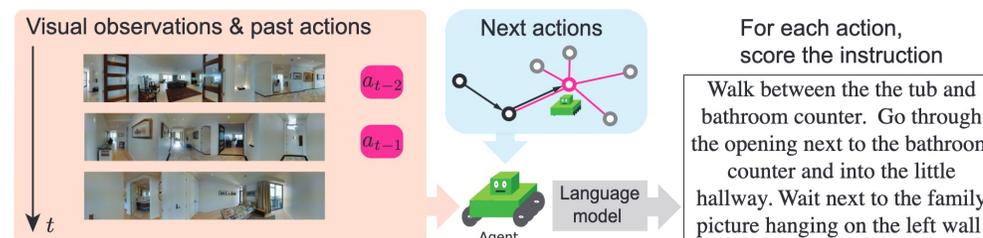
既存手法



Reinforced Cross-modal Matching [Wang et al. 2018]

提案手法

深層学習モデル(ニューラルネット)を視覚・動作からの条件付き言語モデルとする



視覚・動作情報 → ニューラルネット → 指示文章

ニューラルネット（言語モデルを内部に含む）は画像から文生成（スコアリング）を行う

画像・テキストのいずれが欠けてもタスクを解けない

➡ 見たことがない環境で、精度が落ちづらい！

言語モデルを仮想環境でのデータ作成や拡張に使用する研究はこれまでも存在した。

[Fried et al. 2018], [Magassouba et al. 2019]

言語モデルをVLNタスク解決に直接用いる手法は世界初

R4Rデータセットでも有効性を確認、

こんな当たり前のアプローチのどこがダメなのか？

モデルへの入力が視覚情報とテキストの双方だと、Vision & Language共通の問題として、

深層学習モデルが画像・テキストのいずれかの情報に大きく依存してしまうことがある。

➡ 見たことがない環境では、精度が落ちやすい！

Generative language-grounded policy (GLGP)

言語指示 X 、環境（視覚情報や過去に動作情報） h_t としたとき

VLNの先行研究では $p(a_t | h_t, X)$ をモデルしている。

本研究では、どの動作も同一のpriorを持つこと $p(a_t | h_t) = 1/|A|$ (A は動作集合) を仮定して

$$p(a_t | h_t, X) = \frac{p(X | a_t, h_t) p'(a_t | h_t)}{\sum_{a'_t \in \mathcal{A}} p(X | a'_t, h_t) p'(a'_t | h_t)} = \frac{p(X | a_t, h_t)}{\sum_{a'_t \in \mathcal{A}} p(X | a'_t, h_t)}$$

のように、言語モデル $p(X | a_t, h_t)$ を利用する。
言語モデル $p(X | a_t, h_t)$ は、次の動作から指示 X の尤度を計算することで、動作選択する。

$$L = - \sum_{t=1}^T \left\{ \log p(X | a_t, h_t) + \log \sum_{a'_t \in \mathcal{A}} p(X | a'_t, h_t) \right\}$$

Visual observations & past actions

↓ t



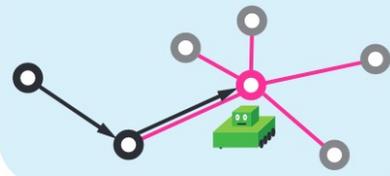
a_{t-2}



a_{t-1}



Next actions

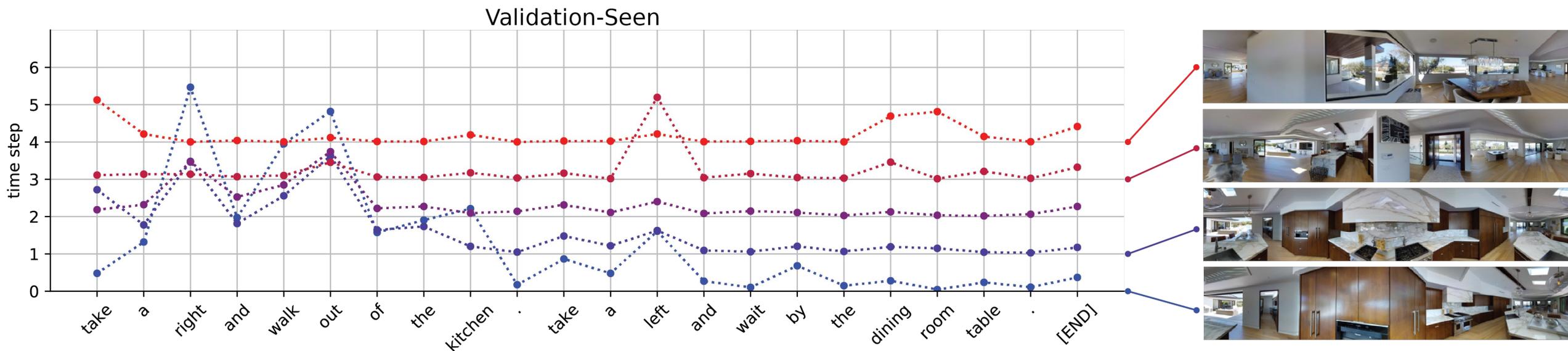


Language model

For each action,
score the instruction

Walk between the the tub and bathroom counter. Go through the opening next to the bathroom counter and into the little hallway. Wait next to the family picture hanging on the left wall.

言語指示に紐付けられた判断の可視化



Take a right and walk out of the kitchen. Take a left and wait by the dining room table.

可能な動作集合が与えられた際に、言語モデルによる指示文章のトークンひとつひとつへのスコアリングを利用し、特に動作ごとに各トークンの予測がどのように変化するかに着目して可視化した。

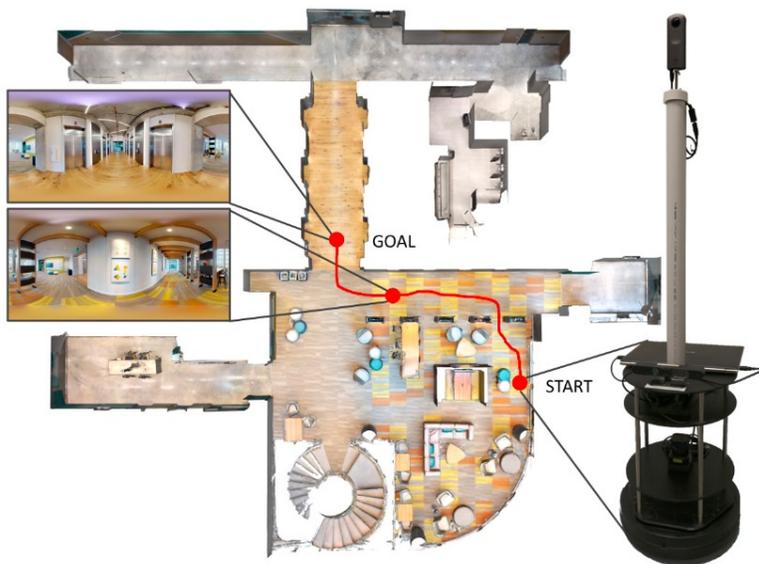
$$S(w_k) = - \sum_{a_t \in \mathcal{A}} q(a_t, w_k) \log_{|\mathcal{A}|} q(a_t, w_k),$$
$$q(a_t, w_k) = \frac{p(w_k | a_t, h_t, w_{:k-1})}{\sum_{a_t \in \mathcal{A}} p(w_k | a_t, h_t, w_{:k-1})}$$

VLNのロボットのナビゲーションへの応用

VLNモデルを現実世界のロボット上で動作させることは出来るのか？ **- Yes!**

Sim-to-Real Transfer for Vision-and-Language Navigation

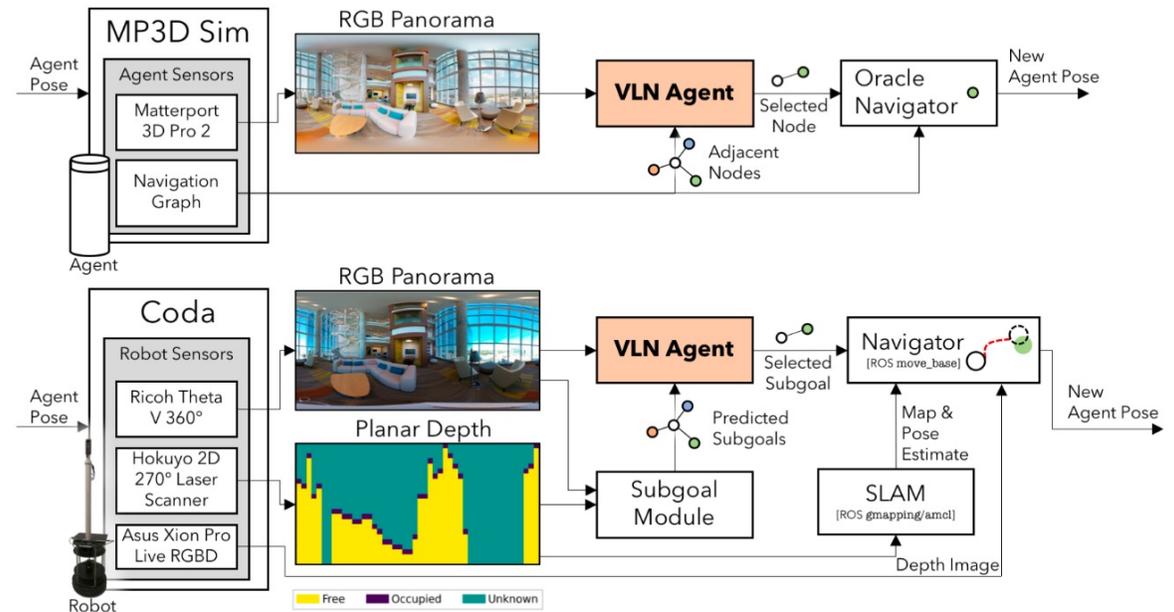
P. Anderson *et al.* (2020) – *The same first author with the original VLN paper.*



Go between the first and second bookshelves, turn to your left and walk straight down the hallway, then you should turn to your right at the hallway with the elevators and stop when the fire extinguisher box is on your left.

Walk between the two bookshelves, turn left, walk past the last set of pictures on the wall, turn right and wait by the elevators.

Turn left and head toward and past the blue bookcase. Turn left again and walk down the long hallway until you get to the opening on the right. Turn right and head toward the elevators and you're there.



Possible future direction:

Integration of high-level machine-learning agents and low-level robotic manipulation.

はじめに：ロボットへの言語での指示

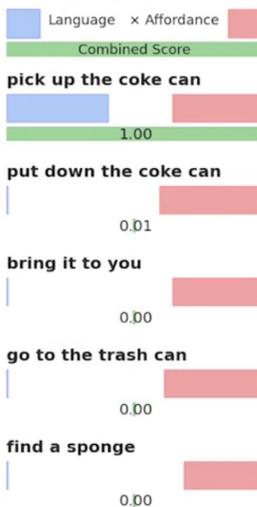
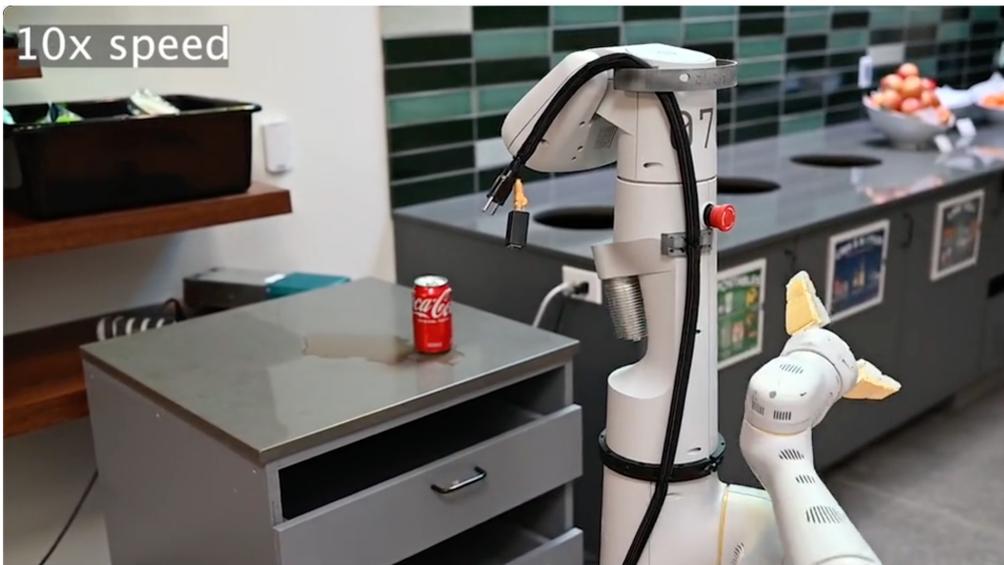
Google Robotics

SayCan: Do As I Can, Not As I Say (2022)

私が言ったようにはなく
私ができるように動作しなさい

C.f. Do as I say, not as I do:

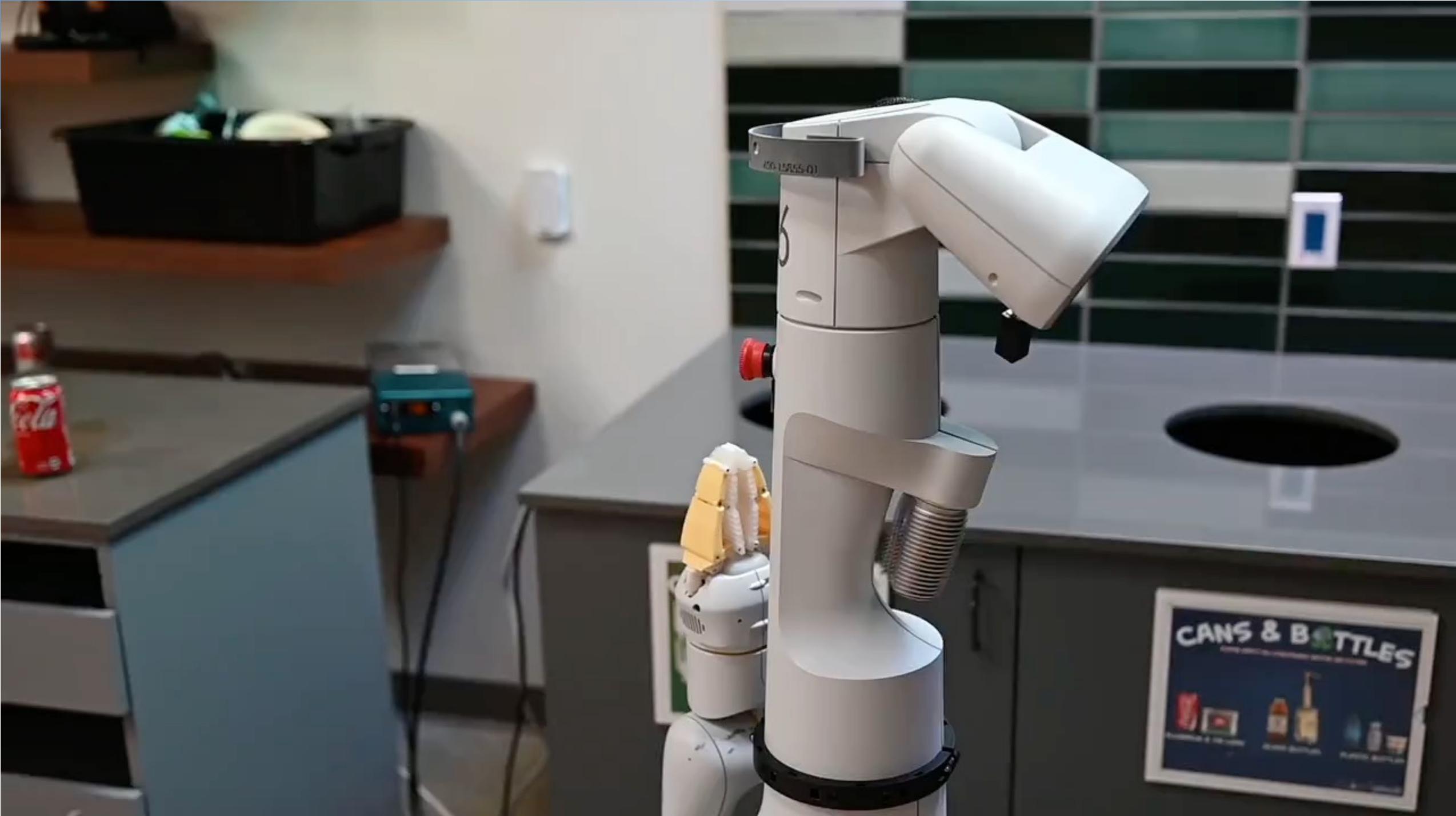
“私が教えるようには行動しなさい、
私がするようにではなく”
ジョン・セルデン「茶話」1654年



INPUT: I spilled my coke on the table, how would you throw it away and bring me something to help clean?

ROBOT:

I spilled my coke on the table, how would you throw it away and bring me something to help clean?
Robot: I would: 1. find a coke can, 2. ___



はじめに：ロボットへの言語での指示

Robotics at Google & Everyday Robots SayCan: Do As I Can, Not As I Say (2022)

私が言ったようではなく
私ができるように動作しなさい

C.f. Do as I say, not as I do:

“私が教えるように行動しなさい、
私がするようにではなく”
ジョン・セルデン「茶話」1654年

I spilled my drink, can you help?

GPT3

You could try using a vacuum cleaner.

LaMDA

Do you want me to find a cleaner?

FLAN

I'm sorry, I didn't mean to spill it.

(a) Large Language Models (LLMs)

I spilled my drink, can you help?

LLM

"find a cleaner"
"find a sponge"
"go to the trash can"
"pick up the sponge"
"try using the vacuum"

Value Functions

"find a cleaner"
"find a sponge"
"go to the trash can"
"pick up the sponge"
"try using the vacuum"



SayCan

"find a sponge"
"go to the trash can"
"pick up the sponge"
"try using the vacuum"



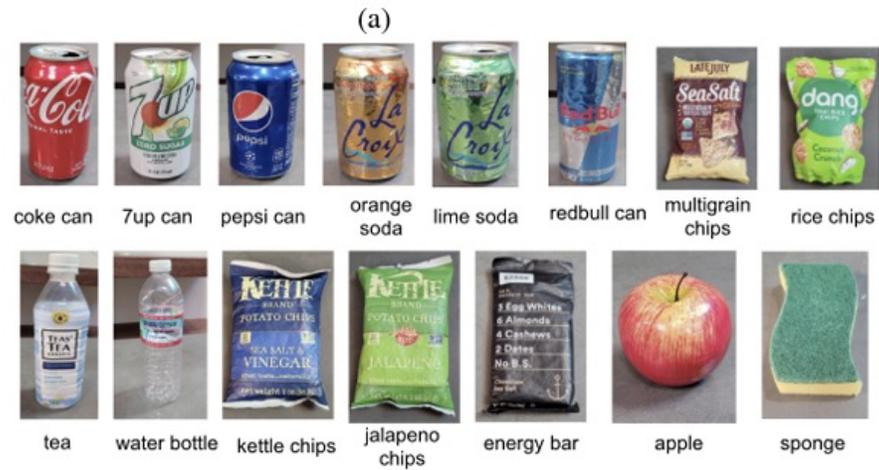
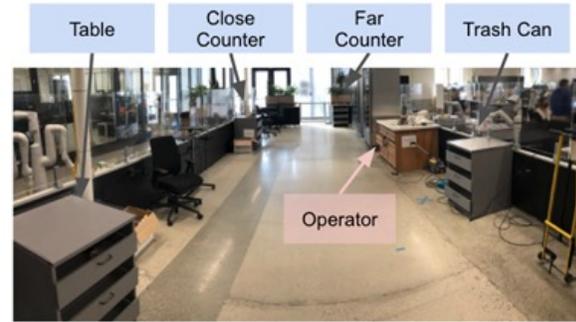
(b) SayCan

INPUT: I spilled my coke on the table, how would you throw it away and bring me something to help clean?

ROBOT: I would

1. find a coke can
2. pick up the coke can
3. go to trash can
4. put down the coke can
5. find a sponge
6. pick up the sponge
7. go to the table
8. put down the sponge
9. _____

SayCan: Do As I Can, Not As I Say (2022)



Find an apple

Find a coke

Find a sponge

Pick up the apple

Pick up the coke

...

Place the apple

Place the coke

Go to the table

Go to the counter

511 preset skill-set
+ text-description
+ value function

101 task instructions in
7 classes (via MTurk)

Figure 5: The experiments were performed in an office kitchen (a) and a mock office kitchen mirroring this setup (b), with 5 predicted locations and 15 objects (c). The robot is a mobile manipulator with a 7 degree-of-freedom arm and a gripper with policies trained from an RGB observation (d).

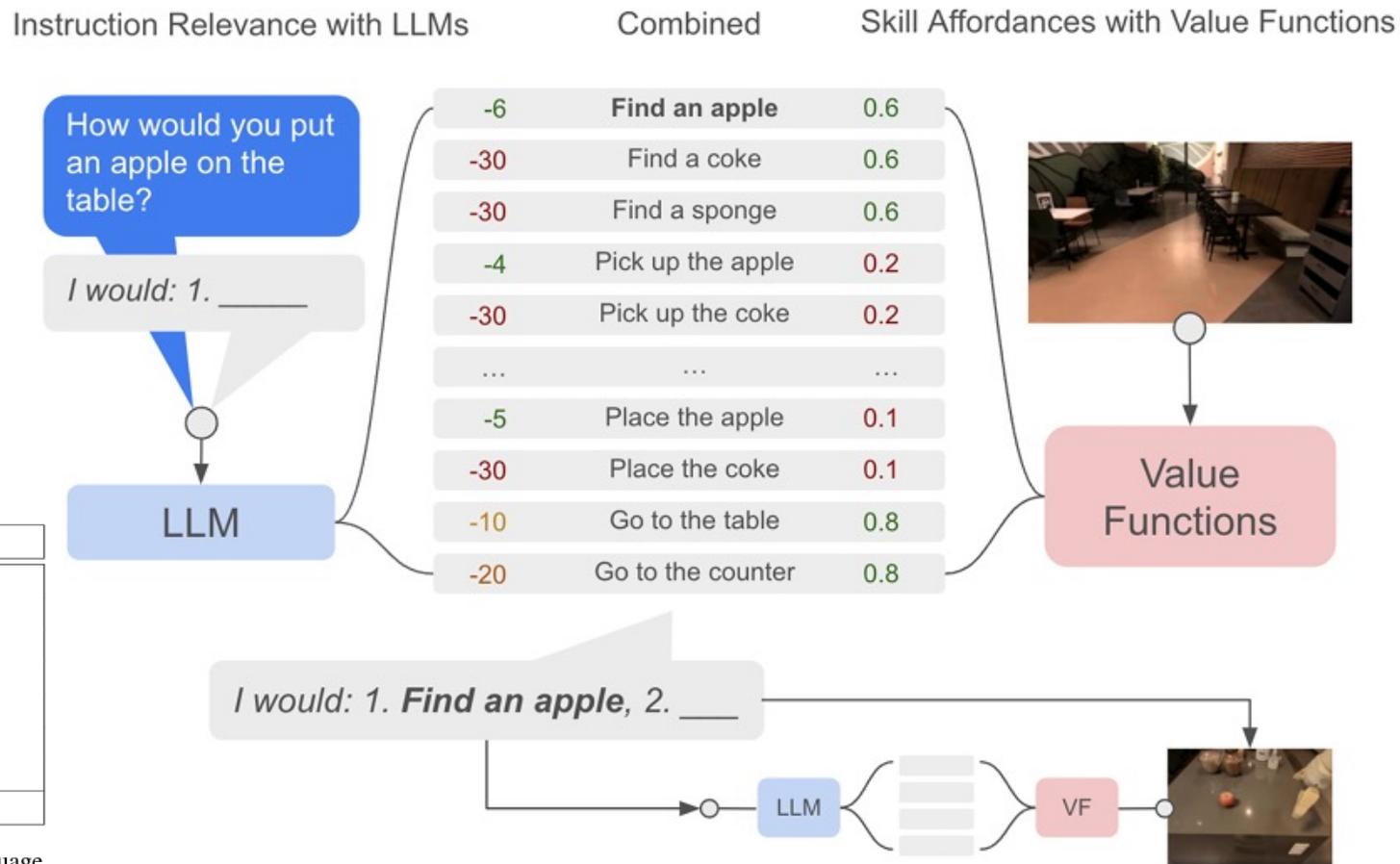
SayCan: Do As I Can, Not As I Say (2022)

- 大規模言語モデル (LLM) に既存行動を prompting として入力し、次の行動を skill-set の中から選択する
- 一方で、環境から各skill毎に計算される value関数の値と合わせることで次の行動を決定する

(あれ...GLGPとどこか似ている... 🤔)

Family	Num	SayCan	No VF	Generative	BC (NL)	BC (USE)
NL Single Primitive	15	67%	33%	40%	0%	60%
NL Nouns	15	60%	53%	40%	0%	0%
NL Verbs	15	80%	53%	87%	0%	0%
Structured Language	15	100%	20%	73%	0%	0%
Embodiment	11	64%	18%	0%	0%	0%
Crowd Sourced	15	73%	60%	47%	0%	0%
Long-Horizon	15	47%	20%	0%	0%	0%
Total	101	70%	38%	43%	0%	9%

Table 4: Planning rate across SayCan and various ablations. *No VF* uses the maximum language score skill from the LLM, *Generative* uses a generative LLM and then projects to the nearest skill via USE embeddings, *BC (NL)* uses the low-level policy with the natural language instruction, and *BC (USE)* uses the low-level policy with the natural language instruction projected to the nearest skill via USE embeddings.



はじめに：ロボットへの言語での指示

ちなみに、必要な将来動作の予測についてGPT-3も結構賢い？

I spilled my drink, can you help?

GPT3

You could try using a vacuum cleaner.

LaMDA

Do you want me to find a cleaner?

FLAN

I'm sorry, I didn't mean to spill it.

(a) Large Language Models (LLMs)

I spilled my drink, can you help?

LLM

"find a cleaner"

"find a sponge"

"go to the trash can"

"pick up the sponge"

"try using the vacuum"

Value Functions

"find a cleaner"

"find a sponge"

"go to the trash can"

"pick up the sponge"

"try using the vacuum"



SayCan

"find a sponge"

"go to the trash can"

"pick up the sponge"

"try using the vacuum"



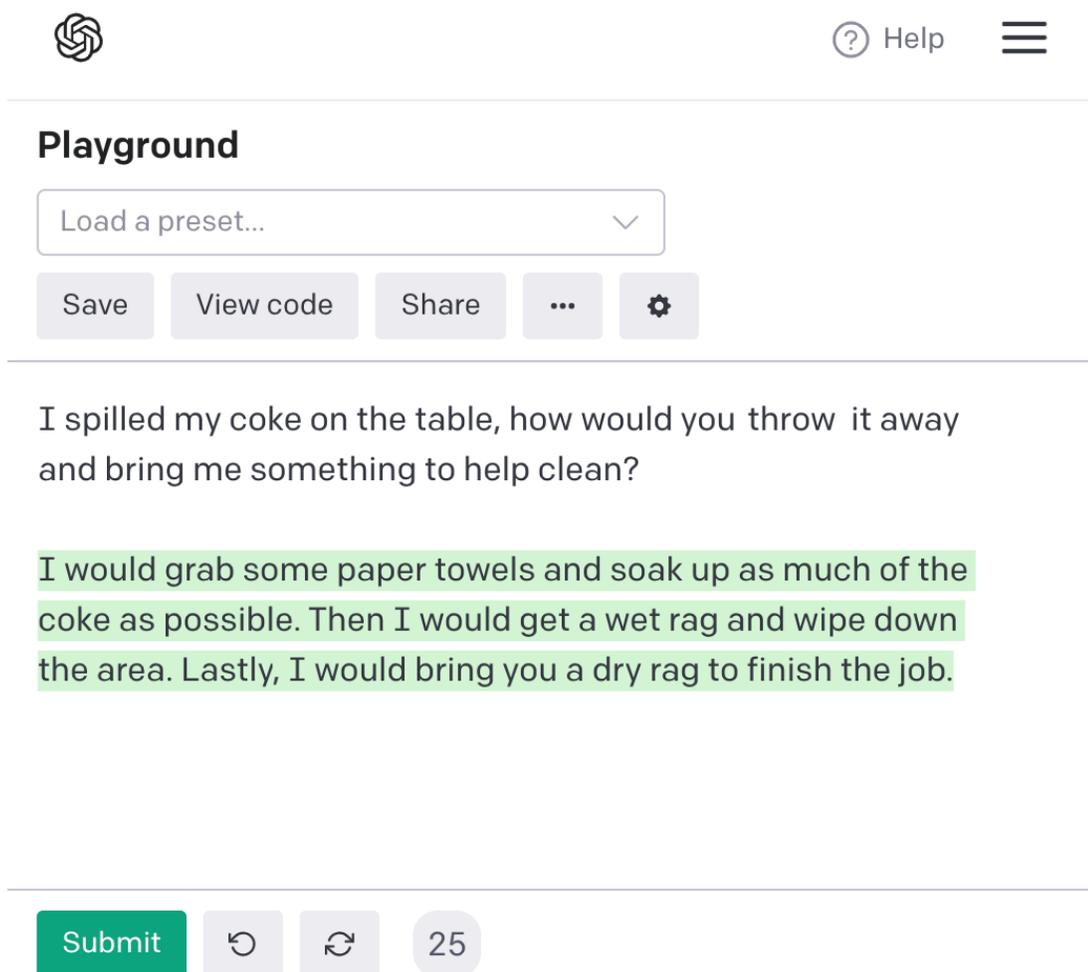
I would:

1. find a sponge
2. pick up the sponge
3. come to you
4. put down the sponge
5. done

(b) SayCan

はじめに：ロボットへの言語での指示

ちなみに、必要な将来動作の予測についてGPT-3も結構賢い？



The screenshot shows the OpenAI Playground interface. At the top, there is a logo, a 'Help' button, and a menu icon. Below this is the 'Playground' section with a 'Load a preset...' dropdown menu and buttons for 'Save', 'View code', 'Share', and a settings gear. The prompt is: "I spilled my coke on the table, how would you throw it away and bring me something to help clean?". The response is: "I would grab some paper towels and soak up as much of the coke as possible. Then I would get a wet rag and wipe down the area. Lastly, I would bring you a dry rag to finish the job." At the bottom, there is a 'Submit' button, a refresh icon, a redo icon, and a token count of 25.

Playground

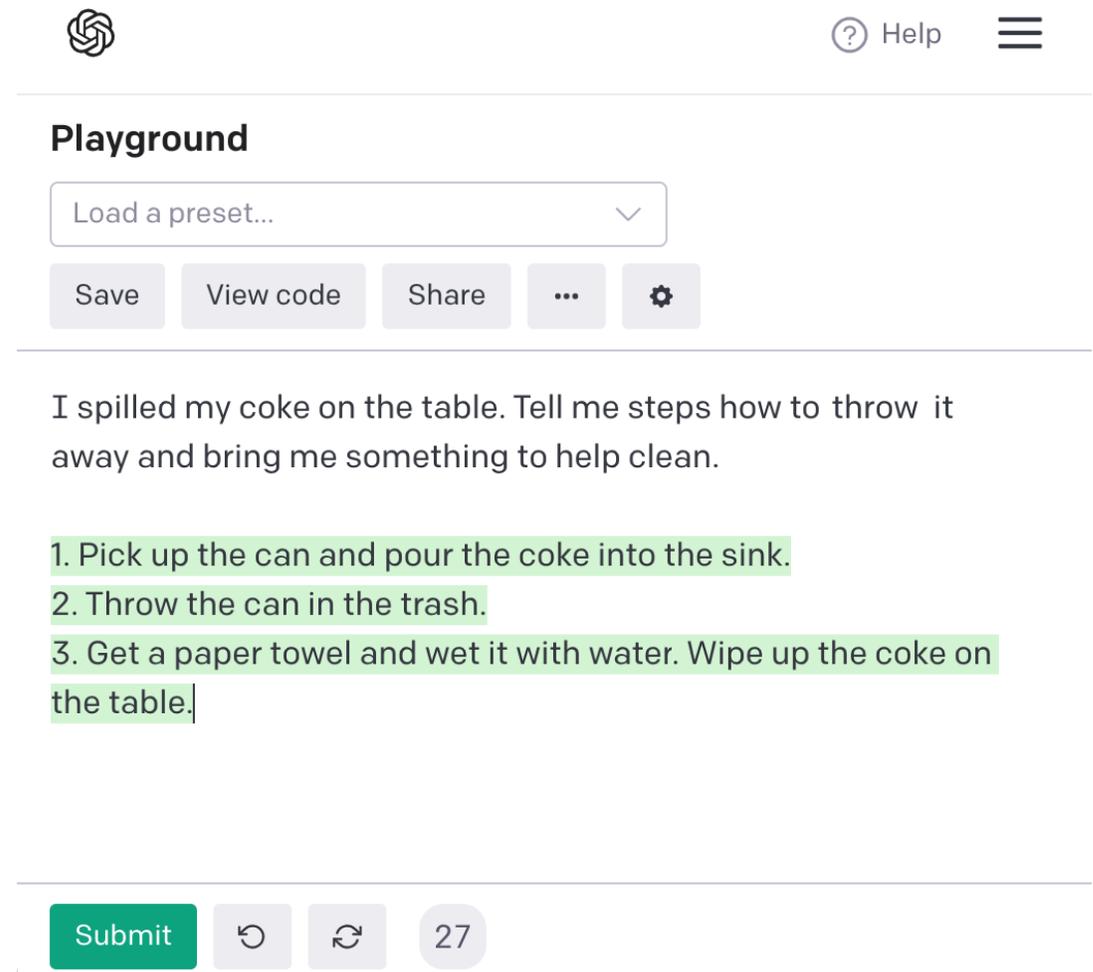
Load a preset...

Save View code Share ...

I spilled my coke on the table, how would you throw it away and bring me something to help clean?

I would grab some paper towels and soak up as much of the coke as possible. Then I would get a wet rag and wipe down the area. Lastly, I would bring you a dry rag to finish the job.

Submit ↺ ↻ 25



The screenshot shows the OpenAI Playground interface. At the top, there is a logo, a 'Help' button, and a menu icon. Below this is the 'Playground' section with a 'Load a preset...' dropdown menu and buttons for 'Save', 'View code', 'Share', and a settings gear. The prompt is: "I spilled my coke on the table. Tell me steps how to throw it away and bring me something to help clean.". The response is a list of three numbered steps: "1. Pick up the can and pour the coke into the sink.", "2. Throw the can in the trash.", and "3. Get a paper towel and wet it with water. Wipe up the coke on the table." At the bottom, there is a 'Submit' button, a refresh icon, a redo icon, and a token count of 27.

Playground

Load a preset...

Save View code Share ...

I spilled my coke on the table. Tell me steps how to throw it away and bring me something to help clean.

1. Pick up the can and pour the coke into the sink.
2. Throw the can in the trash.
3. Get a paper towel and wet it with water. Wipe up the coke on the table.

Submit ↺ ↻ 27

目次

1. はじめに
 1. Embodied AI : 身体に基づいたAI
 2. Generative language grounded policy (GLGP) on VLN
 3. SayCan: Do As I Can, Not As I Say (Google Robotics & Everyday Robotics, 2022)
2. 言語の役割
 1. 名付けによる世界の分節 (F. Saussure)
 2. 言語による指示と実世界との対応付け (L. Wittgenstein)
3. 言語モデル
 1. 歴史 (Markov process, n-gram, RNN)
 2. PaLM & Big-BENCH
 3. Chain of thought
 4. 基盤モデルたち (省略)
4. グランディングと参照表現理解
 1. 参照表現を理解する
 2. テキストに紐付いた物体検出
 3. 3D世界に対応づいた言語理解
5. 言語を使って指示を出す
 1. SHRDLU (T. Winograd, 1971)
 2. ナビゲーション : Vision & Language Navigation (VLN)
 3. プランニング : AI2THORとSayCan
6. まとめ : Language & Robotics のこれから
 1. 言語指示が有効なのはどのような課題か？
 2. 言語理解技術がどのようにロボティクスを助けるか？

言語の役割：

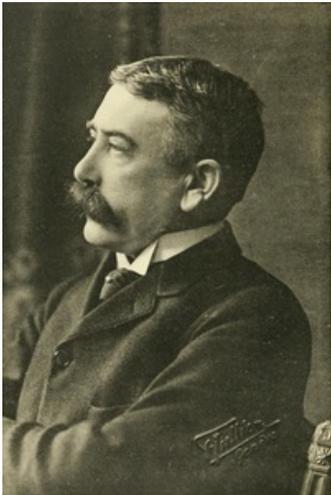
• 言葉には世界を分節する作用がある

例：虹の色は何色ですか？

7色、3色、4色、...

雪国には雪を表す言葉が多い

綿雪、粉雪、細雪、乾雪、湿雪、べと雪、吹雪、ドカ雪、
根雪、雪庇、雪塊、風雪、雪煙、着雪、氷柱、樹氷、...



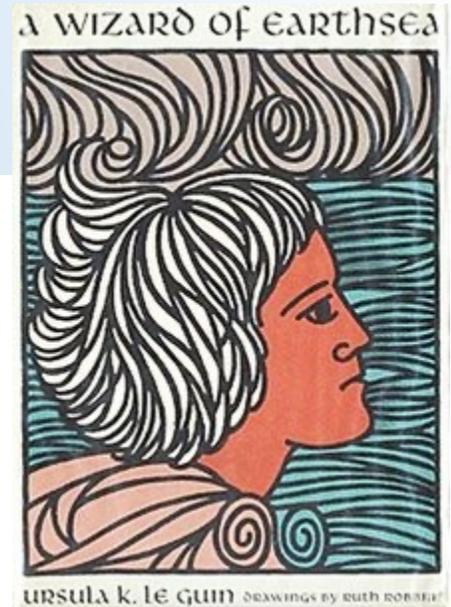
近代言語学の父
ソシュール
(1857-1913)

話し言葉の音韻では、ある音と他の音との差異だけが意味を持つ。
同様に、意味についても、言葉により言葉の指し示す対象が世界からの
差異をもつ。（切り取られ方は言葉により規定される。）

➡ 世界から、テキストによって名前を付けられた物体を分節認識できないか？

テキスト条件付き物体検出器 (桂&栗田 MIRU2022)

やっていることは
MDETR (Kamath et al. NYU 2021)の応用なのですが...



ル・グウィンの小説
『ゲド戦記』

人や物には「真の名前」が存在し
知ることによってその事物を操ることができる

言語の役割：

• 言語による指示

「哲学探求」の言語ゲーム

親方Aは石材建築を行なう。石材には角、柱、板、梁がある。弟子BはAに石材を渡さねばならないが、その順番はAがそれらが必要とする順番である。この目的のため、二人は「角」「柱」「板」「梁」という単語からなる言語を使用する。

Aはこれらを叫ぶ。Bは、それらの叫びに応じて、もっていくよう覚えた石材を持っていく。これを完全に原初的な言語と考えよ。

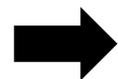
傍目には、「角」や「柱」と言われてもなんのことだかわからないしかし、親方と弟子の間では、明瞭な指示（角石を持ってこい等）に対応する

（言語ゲームの解釈は難しいが）ウィトゲンシュタインは、言葉の意味はその言葉の使用される状況と意図伝達によって規定されると考えたのではないか。同じ言葉でも言語ゲームが異なれば全く違う意味を持ちうる：わかりやすい例：指示語や代名詞は言語ゲームにより全く違う対象を指す



ウィトゲンシュタイン
(1889-1951)

『論理哲学論考』
『哲学探求』



この例では、かなりクローズドな世界での指示を考えていたようにも見える...
(実際には広範なコミュニティを含みうる)

目次

1. はじめに
 1. Embodied AI : 身体に基づいたAI
 2. Generative language grounded policy (GLGP) on VLN
 3. SayCan: Do As I Can, Not As I Say (Google Robotics & Everyday Robotics, 2022)
2. 言語の役割
 1. 名付けによる世界の分節 (F. Saussure)
 2. 言語による指示と実世界との対応付け (L. Wittgenstein)
3. 言語モデル
 1. 歴史 (Markov process, n-gram, RNN)
 2. PaLM & Big-BENCH
 3. Chain of thought
 4. 基盤モデルたち (省略)
4. グランディングと参照表現理解
 1. 参照表現を理解する
 2. テキストに紐付いた物体検出
 3. 3D世界に対応づいた言語理解
5. 言語を使って指示を出す
 1. SHRDLU (T. Winograd, 1971)
 2. ナビゲーション : Vision & Language Navigation (VLN)
 3. プランニング : AI2THORとSayCan
6. まとめ : Language & Robotics のこれから
 1. 言語指示が有効なのはどのような課題か？
 2. 言語理解技術がどのようにロボティクスを助けるか？

言語モデル：歴史

- $$p(X) = \prod_i p(w_i | w_0, \dots, w_{i-1})$$
$$= p(w_L | w_0, \dots, w_i, \dots, w_{L-1}) p(w_{L-1} | w_0, \dots, w_i, \dots, w_{L-2}) \cdots p(w_1 | w_0) p(w_0)$$



Markov, A. A



Shannon, C. E

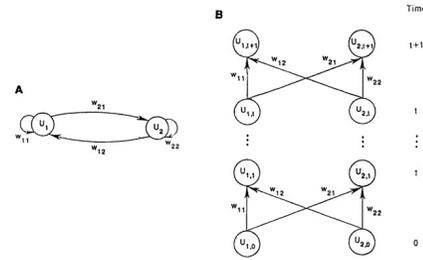


FIGURE 17. A comparison of a recurrent network and a feedforward network with identical behavior. A: A completely connected recurrent network with two units. B: A feedforward network which behaves the same as the recurrent network. In this case, we have a separate unit for each time step and we require that the weights connecting each layer of units to the next be the same for all layers. Moreover, they must be the same as the analogous weights in the recurrent case.

Rumelhart's RNN (1985)

Markov, A. A. 1913 Essai d'une recherche statistique sur le texte du roman "Eugene Onegin" illustrant la liaison des epreuve en chain **(bi-gram, tri-gram)**

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal, 27(3):379–423. Continued in the following volume. **(n-gram)**

Rumelhart, David E; Hinton, Geoffrey E, and Williams, Ronald J (Sept. 1985). Learning internal representations by error propagation. **(RNN)**

言語モデル: PaLM & BIG-bench dataset

Beyond the Imitation Game benchmark (BIG-bench)

PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery* Sharan Narang* Jacob Devlin*

Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham

Hyung Won Chung Charles Sutton Sebastian Gehrmann Parker Schuh Kensen Shi

Sasha Tsvyashchenko Joshua Maynez Abhishek Rao† Parker Barnes Yi Tay

Noam Shazeer† Vinodkumar Prabhakaran Emily Reif Nan Du Ben Hutchinson

Reiner Pope James Bradbury Jacob Austin Michael Isard Guy Gur-Ari

Pengcheng Yin Toju Duke Anselm Levskaya Sanjay Ghemawat Sunipa Dev

Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus

Denny Zhou Daphne Ippolito David Luan† Hyeontaek Lim Barret Zoph

Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick

Andrew M. Dai Thanumalayan Sankaranarayanan Pillai Marie Pellat Aitor Lewkowycz

Erica Moreira Rewon Child Oleksandr Polozov† Katherine Lee Zongwei Zhou

Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta† Jason Wei

Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

- PaLMはdecoder-only large language model論文である
- BIG-Benchは途方もない巨大科学論文に見えるが公開 github への pull request を介して既存のものを含め204個のタスクを集めたもの

BEYOND THE IMITATION GAME: QUANTIFYING AND EXTRAPOLATING THE CAPABILITIES OF LANGUAGE MODELS

Alphabetic author list:*

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholami-davoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakas, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Cameron Dour, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khoshabi, Daniel Levy, Daniel Mosegué González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engelfu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, François Chollet, Frieda Rong, Gaurav Mishra, Genia Idris Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovich-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shelpin, Hinrich Schütze, Hirum Yakuza, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocořík, Jana Thompson, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Berant, Jörg Froberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chaiyafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqi, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátys Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starrit, Michael Strube, Michał Śwędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mímee Xu, Mirac Suzgun, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhddeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Castanes, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Huth, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pезhshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qianlong Chen, Rabih Banjade, Rachel Eita Rudolph, Raefer Gabriel, Rahel Habacker, Ramón Risco Delgado, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelev, Ronan LeBraz, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Aasadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima (Shammie) Debnath, Siamak Shakeri, Simon Thormeyer, Simon Thormeyer, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Sumner Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Timothy Telleen-Lawton, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, Zivi Wu

言語モデル : PaLM

- 5400億パラメータの 118層 Decoder-only Transformer

Model	Layers	# of Heads	d_{model}	# of Parameters (in billions)	Batch Size
PaLM 8B	32	16	4096	8.63	256 → 512
PaLM 62B	64	32	8192	62.50	512 → 1024
PaLM 540B	118	48	18432	540.35	512 → 1024 → 2048

+ SwiGLU Activation, + Parallel Layers, + Multi-Query Attention

+ RoPE Embeddings

+ 7800億 tokens の多言語 (人工言語含む) 学習データ

Total dataset size = 780 billion tokens

Data source	Proportion of data
Social media conversations (multilingual)	50%
Filtered webpages (multilingual)	27%
Books (English)	13%
GitHub (code)	5%
Wikipedia (multilingual)	4%
News (English)	1%

言語モデル：PaLM

- 一定以上のパラメータを持つ言語モデルはそれまでに持たなかった問題解決能力を創発する???

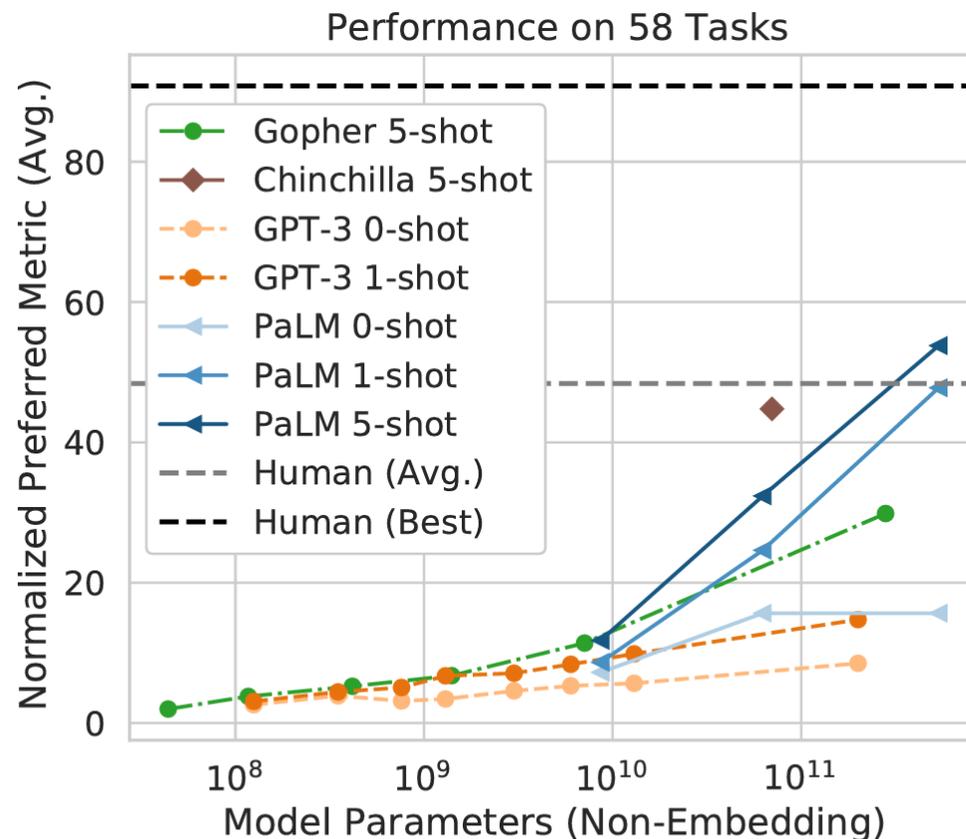


Figure 3: BIG-bench evaluation of PaLM. (left)

言語モデル：PaLM

- 一定以上のパラメータを持つ言語モデルはそれまでに持たなかった問題解決能力を創発する???

イベント間グラフ推論

- **goal_step_wikihow** - The goal is to reason about goal-step relationship between events. Example:
Input: In order to "clean silver," which step should be done first? (a) dry the silver (b) handwash the silver
Answer: (b) handwash the silver

論理推論

- **logical_args** - The goal is to predict the correct logical inference from a passage. Example:
Input: Students told the substitute teacher they were learning trigonometry. The substitute told them that instead of teaching them useless facts about triangles, he would instead teach them how to work with probabilities. What is he implying? (a) He believes that mathematics does not need to be useful to be interesting. (b) He thinks understanding probabilities is more useful than trigonometry. (c) He believes that probability theory is a useless subject.
Answer: (b) He thinks understanding probabilities is more useful than trigonometry.

英語ことわざ

- **english_proverbs** - The goal is to guess which proverb best describes a text passage. Example:
Input: Vanessa spent lots of years helping out on weekends at the local center for homeless aid. Recently, when she lost her job, the center was ready to offer her a new job right away. Which of the following proverbs best apply to this situation? (a) Curses, like chickens, come home to roost. (b) Where there is smoke there is fire (c) As you sow, so you shall reap.
Answer: (c) As you sow, so you shall reap.

動作・イベントなどの論理的な順番

- **logical_sequence** - The goal is to order a set of "things" (months, actions, numbers, letters, etc.) into their logical ordering. Example:
Input: Which of the following lists is correctly ordered chronologically? (a) drink water, feel thirsty, seal water bottle, open water bottle (b) feel thirsty, open water bottle, drink water, seal water bottle (c) seal water bottle,

ナビゲート

- **navigate** - The goal is to follow a set of simple navigational instructions, and figure out where you would end up. Example:
Input: If you follow these instructions, do you return to the starting point? Always face forward. Take 6 steps left. Take 7 steps forward. Take 8 steps left. Take 7 steps left. Take 6 steps forward. Take 1 step forward. Take 4 steps forward.
Answer: No

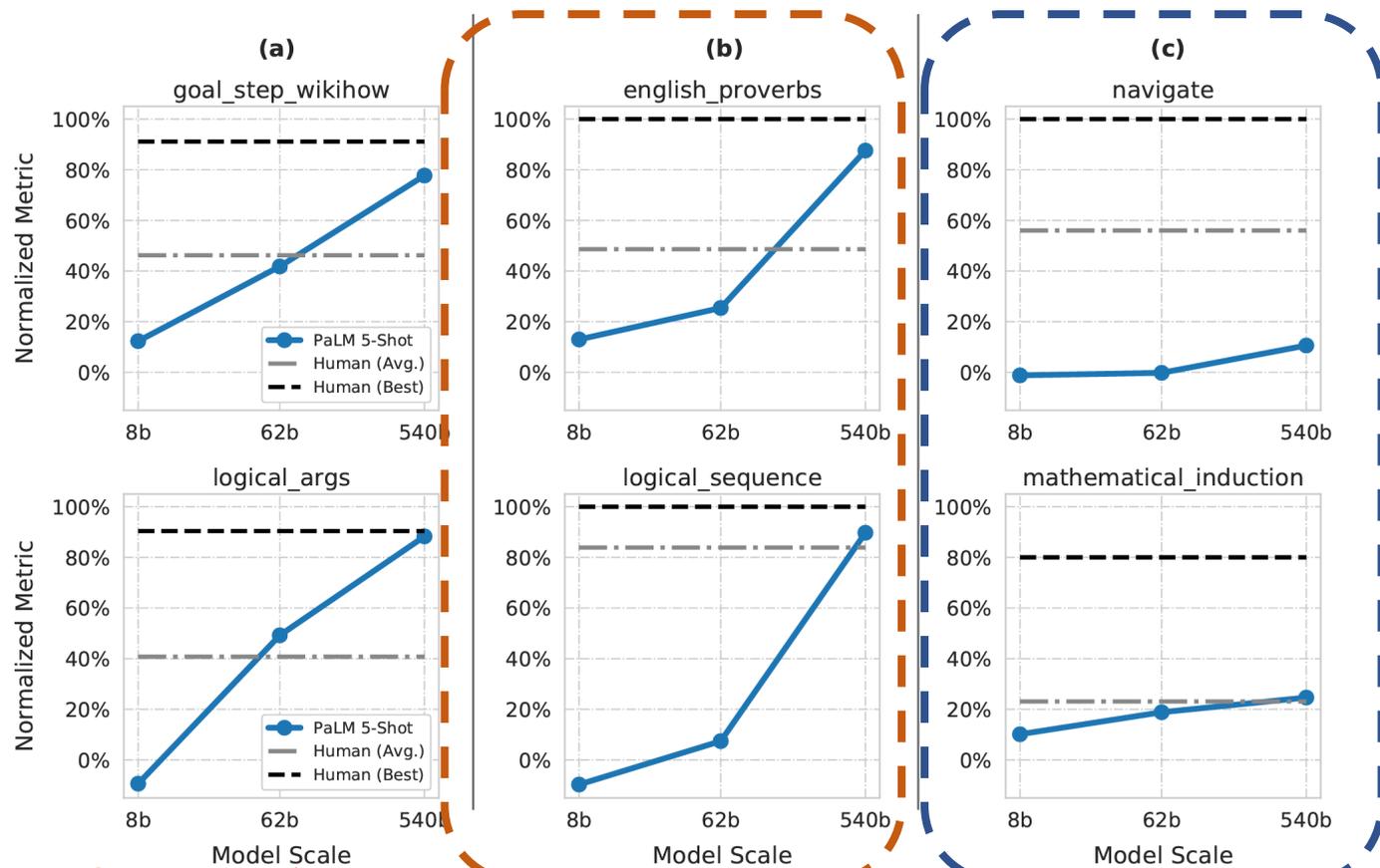
数学的推論

- **mathematical_induction** - The goal is to perform logical inference mathematical induction rules, even if they contradict real-world math. Example:
Input: It is known that adding 2 to any odd integer creates another odd integer. 2 is an odd integer. Therefore, 6 is an odd integer. Is this a correct induction argument (even though some of the assumptions may be incorrect)?
Answer: Yes

(from BIG-bench dataset)

言語モデル：PaLM

- 一定以上のパラメータを持つ言語モデルはそれまでに持たなかった問題解決能力を創発する??



不連続な増加が見られる

ナビゲートと 数学的推論は540Bでも
難しいらしい

Figure 5: 5-shot evaluations on six individual tasks with interesting scaling characteristics. For each task, the

言語モデル：PaLM

- Chain of thought prompting (Google Research, 2022)

Standard prompting

Input: Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: The answer is 11.

...

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
A:

Model output: The answer is 50. ❌

Chain of thought prompting

Input: Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?
A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

...

Q: The cafeteria had 23 apples. If they used 20 to make lunch and bought 6 more, how many apples do they have?
A:

Model output: The cafeteria had 23 apples originally. They used 20 to make lunch. So they had $23 - 20 = 3$. They bought 6 more apples, so they have $3 + 6 = 9$. The answer is 9. ✅

Model+Technique	Accuracy
PaLM 540B+chain-of-thought+calculator	58%
PaLM 540B+chain-of-thought	54%
PaLM 540B w/o chain-of-thought	17% 37% ↑
PaLM 62B+chain-of-thought	33%
GPT-3+finetuning+chain-of-thought+calculator	34%
GPT-3+finetuning+chain-of-thought+calculator+verifier	55%

“思考”の途中経過を例としてモデルに入力することで、モデルに途中経過を出力させながら正解出力へと導くことができる。



PaLM論文の結論で著者らは、（単なる理解でなく）生成する能力こそが、カテゴリ分類や回帰タスクのような言語生成を必要としないタスクにも有効でありうる、と主張している。

シンボリックな情報を予測することで精度が向上したとする実験論文は昔からあるが、ここまで明確に実験で示したことがすごい

※ Chain of thought については、最近、東大松尾研&Googleから興味深い論文が出ています:

“Large Language Models are Zero-Shot Reasoners”, (Kojima et al. 2022)

言語モデル : Chain of thought

- Chain of thought prompting (Google Research, 2022)

Math Word Problems (free response)

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Math Word Problems (multiple choice)

Q: How many keystrokes are needed to type the numbers from 1 to 500?
Answer Choices: (a) 1156 (b) 1392 (c) 1480 (d) 1562 (e) 1788

A: There are 9 one-digit numbers from 1 to 9. There are 90 two-digit numbers from 10 to 99. There are 401 three-digit numbers from 100 to 500. $9 + 90(2) + 401(3) = 1392$. The answer is (b).

CSQA (commonsense)

Q: Sammy wanted to go to where the people were. Where might he go?
Options: (a) race track (b) populated areas (c) desert (d) apartment (e) roadblock

A: The answer must be a place with a lot of people. Race tracks, desert, apartments, and roadblocks don't have a lot of people, but populated areas do. So the answer is (b).

StrategyQA

Q: Yes or no: Would a pear sink in water?

A: The density of a pear is about 0.6 g/cm^3 , which is less than water. Thus, a pear would float. So the answer is no.

Date Understanding

Q: The concert was scheduled to be on 06/01/1943, but was delayed by one day to today. What is the date 10 days ago in MM/DD/YYYY?

A: One day after 06/01/1943 is 06/02/1943, so today is 06/02/1943. 10 days before today is 05/23/1943. So the answer is 05/23/1943.

Sports Understanding

Q: Is the following sentence plausible? "Joao Moutinho caught the screen pass in the NFC championship."

A: Joao Moutinho is a soccer player. The NFC championship is part of American football, not soccer. So the answer is no.

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Last Letter Concatenation

Q: Take the last letters of the words in "Lady Gaga" and concatenate them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

Coin Flip (state tracking)

Q: A coin is heads up. Maybelle flips the coin. Shalonda does not flip the coin. Is the coin still heads up?

A: The coin was flipped by Maybelle. So the coin was flipped 1 time, which is an odd number. The coin started heads up, so after an odd number of flips, it will be tails up. So the answer is no.

言語モデル : Chain of thought

- Chain of thought prompting (Google Research, 2022)

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.
Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

coin. Is th
A: The co
So the co
is an odd
heads up
flips, it w
is no.

DA (commonsense)

入力では **something to eat** とは全く言っていない！
(推論としては完全に間違いである！)

しかし「フルーツじゃないものを持ってきて」と言われて、
食べられないものを持ってくる人はどのくらいいるだろうか？

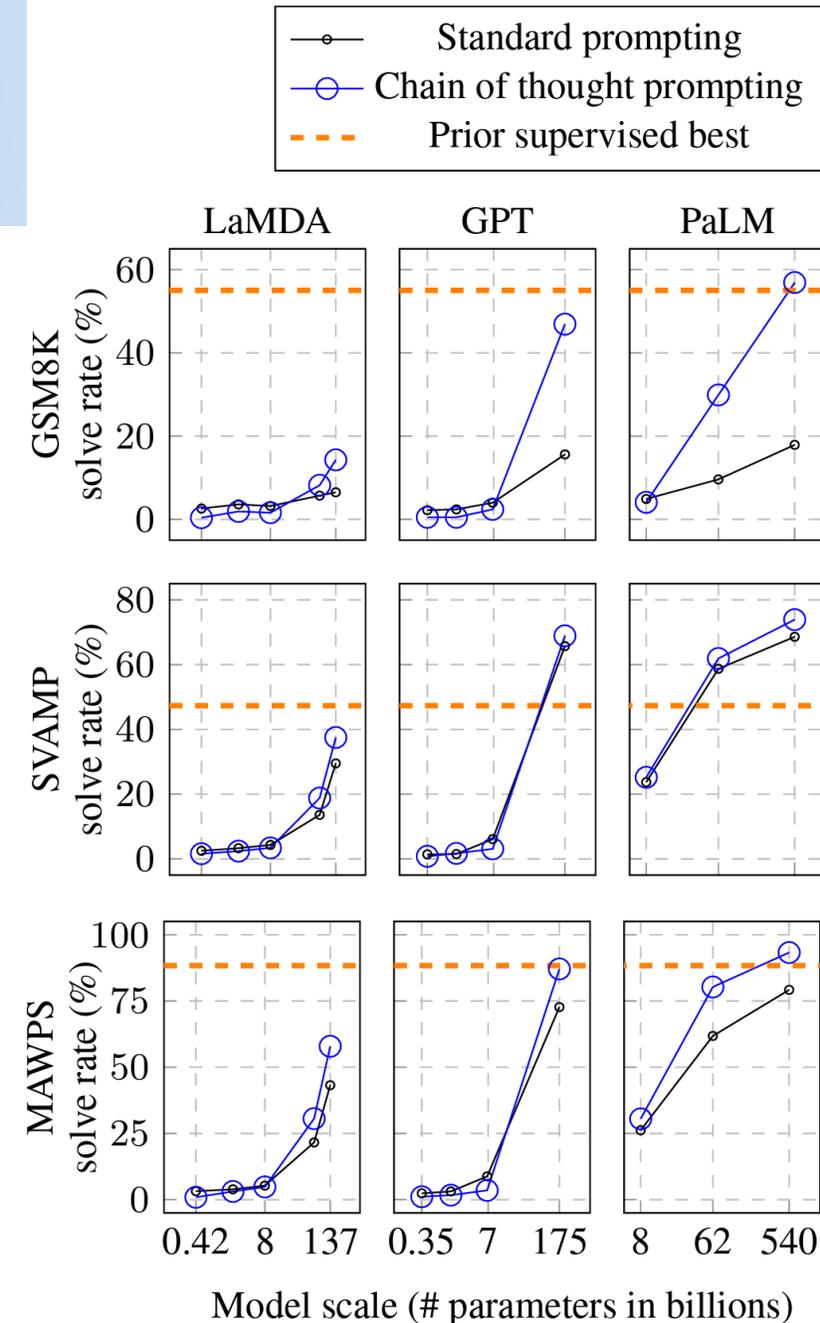
指示文での論理 (の欠陥) を
"常識"により override している？

(実際には、SayCanに登場する objects は食品関係に
強く bias されており、それを学習しているだけかもしれない)



言語モデル：Chain of thought

- Chain of thought prompting (Google Research, 2022)
 - ただし、chain of thoughtそのものもモデルサイズによって創発するものらしい
 - 小さなモデルでchain of thoughtの効果を得ることは難しいかもしれない (有害ですらありえるケースも)
 - シンボル情報を生成して後の解析に使う
アイデアは、今に始まったことでもない
 - 例えばNLPの transition-based parser あたりでもシンボル情報を生成して後の解析に使っています
 - もっとも、そういう環境では、得られた情報はむしろ“error-propagation”等としてノイズの原因扱いされることが多かった気がします
- そもそも、PaLMは汎用選手であり例えば機械翻訳や読解理解のような個別タスクでは専用選手であるモデルで教師あり学習したほうが効率がよいだろう



目次

1. はじめに
 1. Embodied AI : 身体に基づいたAI
 2. Generative language grounded policy (GLGP) on VLN
 3. SayCan: Do As I Can, Not As I Say (Google Robotics & Everyday Robotics, 2022)
2. 言語の役割
 1. 名付けによる世界の分節 (F. Saussure)
 2. 言語による指示と実世界との対応付け (L. Wittgenstein)
3. 言語モデル
 1. 歴史 (Markov process, n-gram, RNN)
 2. PaLM & Big-BENCH
 3. Chain of thought
 4. 基盤モデルたち (省略)
4. グランディングと参照表現理解
 1. 参照表現を理解する
 2. テキストに紐付いた物体検出
 3. 3D世界に対応づいた言語理解
5. 言語を使って指示を出す
 1. SHRDLU (T. Winograd, 1971)
 2. ナビゲーション : Vision & Language Navigation (VLN)
 3. プランニング : AI2THORとSayCan
6. まとめ : Language & Robotics のこれから
 1. 言語指示が有効なのはどのような課題か？
 2. 言語理解技術がどのようにロボティクスを助けるか？

参照表現を理解する

- 参照表現とは，ある対象を指し示すテキスト中の表現である
- 栗田は大学院ではテキスト中での参照表現を中心に解析していた
 - これらを特に照応解析，省略解析等と呼ぶ
 - 文に対する文法的な解析や意味的な解析とも関連が深い分野です

	Predicate	NOM	ACC	DAT
(1) タクシーが _{NOM} 客を _{ACC} 駅に _{DAT} 送った。 takushi-ga kyaku-wo eki-ni okutta. A taxi carried passengers to the station.	送った okutta sent/carried	タクシー takushi taxi	客 kyaku passenger	駅 eki station
(2) その列車は荷物を _{ACC} 運んだ。 sono ressha-wa nimotsu-wo hakonda. The train also carried baggages.	運んだ hakonda carried	[列車 ressha train]	荷物 nimotsu baggage	NULL
(3) タクシーが _{NOM} 客を _{ACC} 乗せたとき事故に _{DAT} 巻き込まれた。 takushi-ga kyaku-wo noseta toki jiko-ni makikomareta. When the taxi picked up passengers, it was involved in the accident.	乗せた noseta picked up 巻き込まれた makikomareta was involved	タクシー takushi taxi [タクシー takushi taxi]	客 kyaku passenger NULL	事故 jiko accident
(4) この列車に <u>は</u> 乗れません。 kono ressha-ni-wa noremasen. You can not take this train.	乗れません noremasen can not take	[あなた anata you]	NULL	列車 ressha train

参照表現を理解する

- テキストには，外部への参照も当然に存在する
 - 照応・省略解析では，“著者”や“読者”を解析に入れる [萩行2014]
 - 固有表現解析では，“科学技術振興機構”や“栗田修平”のような固有表現を，テキストから抽出した固有表現言及に対応付ける。

理研の栗田らは，JSTの研究費で研究を行い，ICLR2021で発表を行った。

理化学研究所

栗田修平

科学技術振興機構

International Conference on
Learning Representations, 2021

- この技術を応用することで，Wikipedia等処理し，大規模な知識ベースを精度良く構築できる



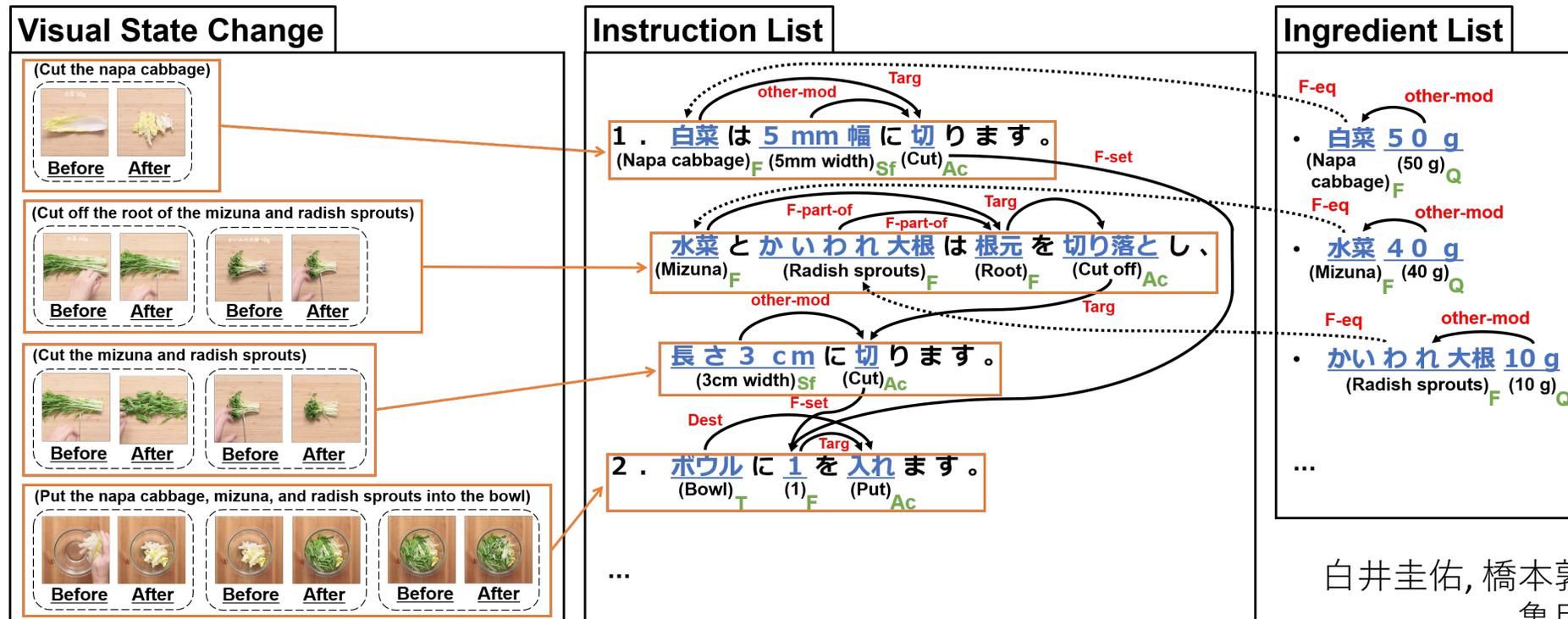
森羅 SHINRA

Wikipedia 構造化プロジェクト



参照表現を理解する

- レシピテキスト中の参照表現
 - レシピテキストには、調理動作と調理対象の素材への言及が出現する
 - “素材が調理動作によりどのように変化するか？”に着目し、レシピテキストに紐付いたアノテーションを行った



白井圭佑, 橋本敦史, 牛久祥孝, 栗田修平,
亀甲博貴, 森信介 NLP2022

※ドメインは限られているが、自動調理ロボットのような現実的な応用可能性

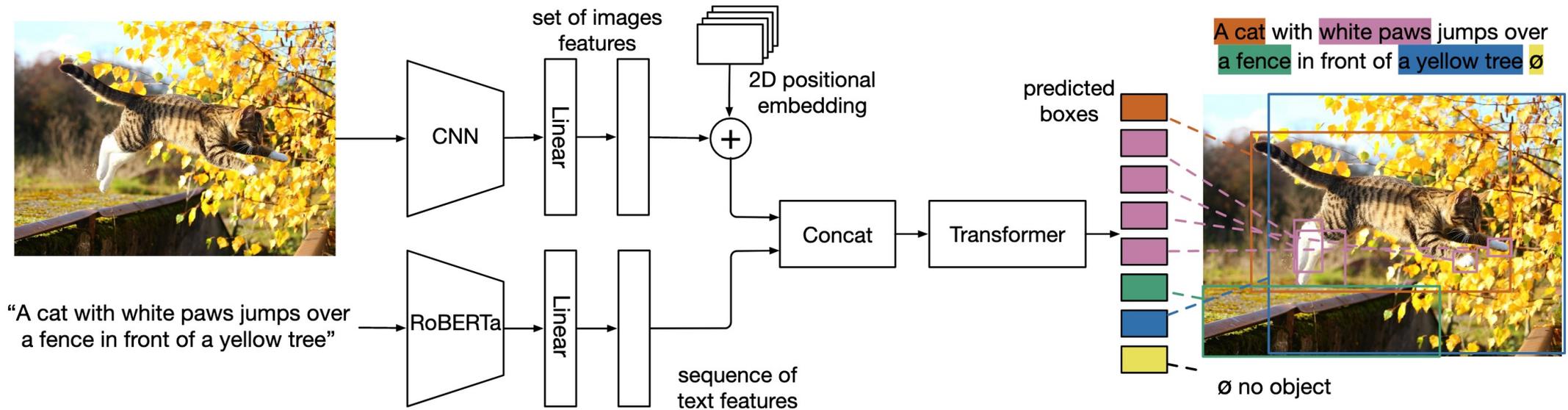
参照表現を理解する

- 画像の参照表現理解(referring expression comprehension)
 - テキストが与えられて、画像からテキストに対応した物体・人物を検出する
 - いわゆるVisual Groundingと呼ばれることも

RefClef	RefCOCO	RefCOCO+
		
<p>right rocks rocks along the right side stone right side of stairs</p>	<p>woman on right in white shirt woman on right right woman</p>	<p>guy in yellow dirbbling ball yellow shirt and black shorts yellow shirt in focus</p>

※実世界で動作させるためには、前述の3Dアノテーション等、何らかの拡張が必要だろう

画像と言語の事前学習：MDETR モデル



MDETR (Kamath *et al.* ICCV2021, NYU)

- 物体検出モデルであるDETR (DEtection TRansformer)を基に,
vision & language用にテキストエンコーダーを追加したモデル
- Soft Token Prediction 事前学習：文内の単語・熟語にマッチする物体領域を学習させる

画像と言語の対応付け事前学習

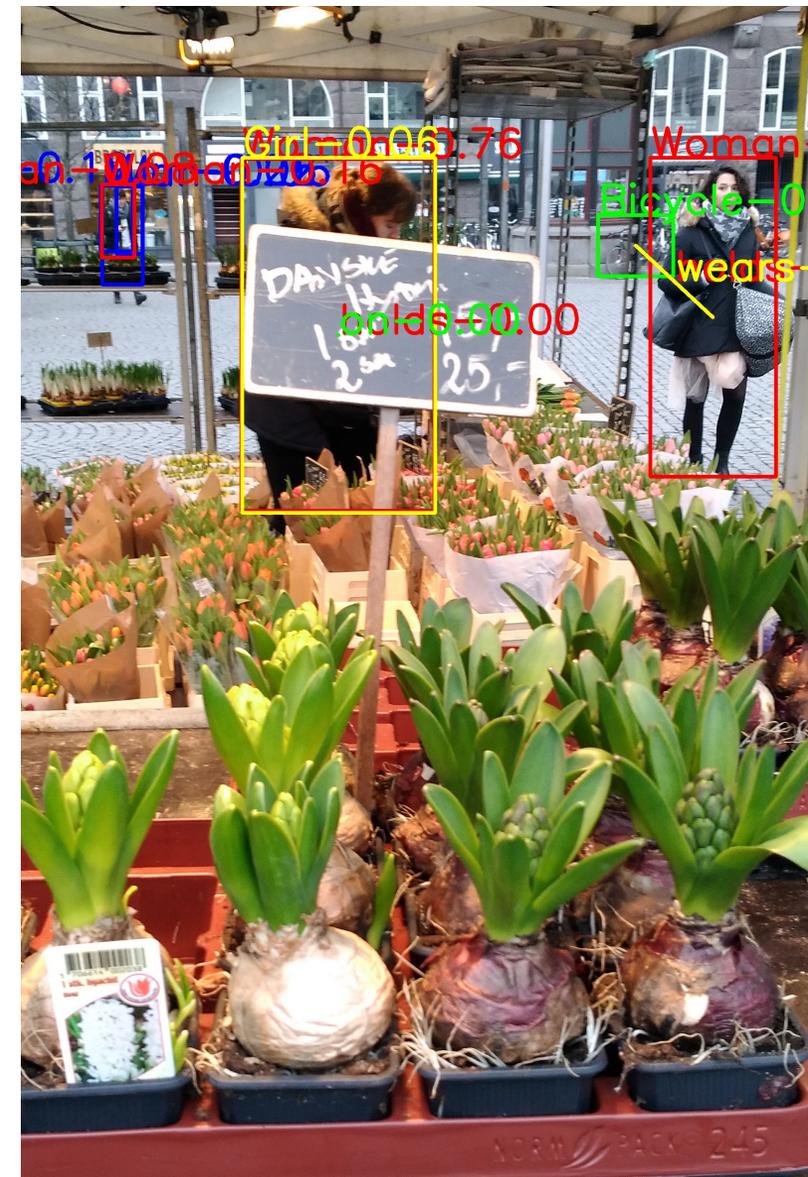
そもそも、画像と言語を対応付けるとは、どういうことか？

(例) 画像中の物体検出・属性抽出・関係抽出 (VinVL)



このように画像からdetailedなシンボル特徴をグラフ形式で抽出しようとするscene graphの試みは、自然言語文から構文木/意味/関係グラフを作るための解析器パイプラインにも似ている。

画像と言語の対応付け事前学習

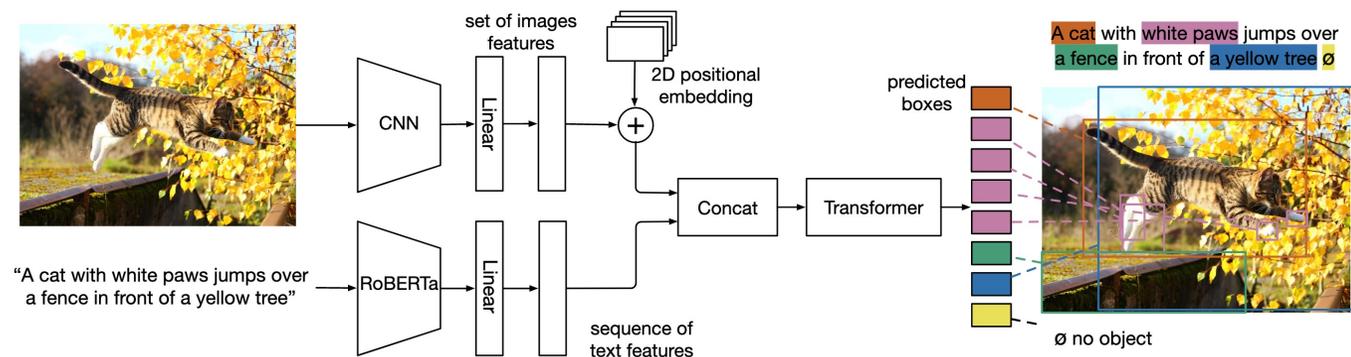
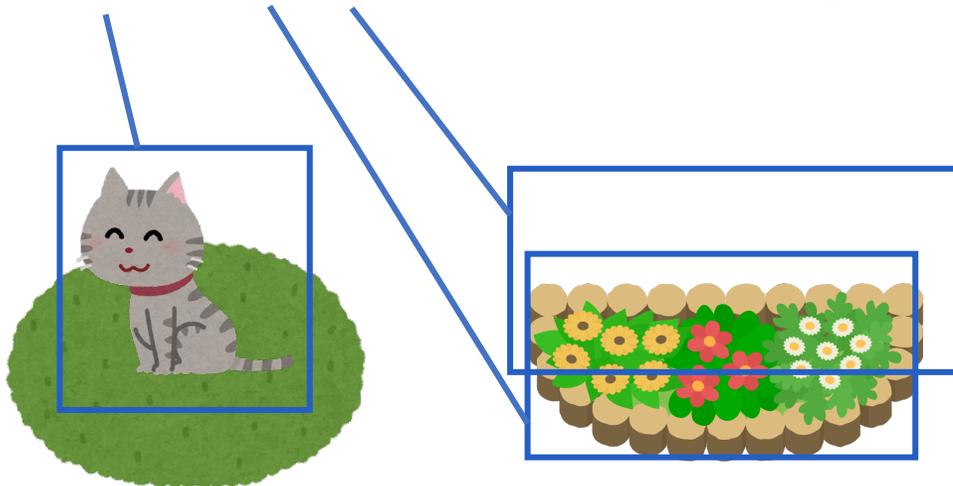


画像と言語の対応付け事前学習

言語に条件付けられた画像からの情報抽出

- 初めに、文の解析器や画像からのScene Graphは、非常に便利なものです
- **しかし、テキスト/画像からのSymbolicな情報抽出と統合には、限界があるのではないかと？**
 - 精度の限界、他モダリティとの対応付けの限界、情報粒度の問題、ill-pose性、フレーム問題、etc...
- 言語による指示文章（もしくは指示表現やキャプションの言語情報）がある状況で、**画像からのsymbolic extractionに頼らずに、言語中に含まれる表現を中心に画像と直接に対応付けできないか？**

「近所のしま猫に花壇を荒らされると困るから
しま猫が**花壇の上**に来たらメールしてね」



関連研究: MDETR [Kamath *et al.*, 2021, NYU]

テキスト条件付き物体検出器

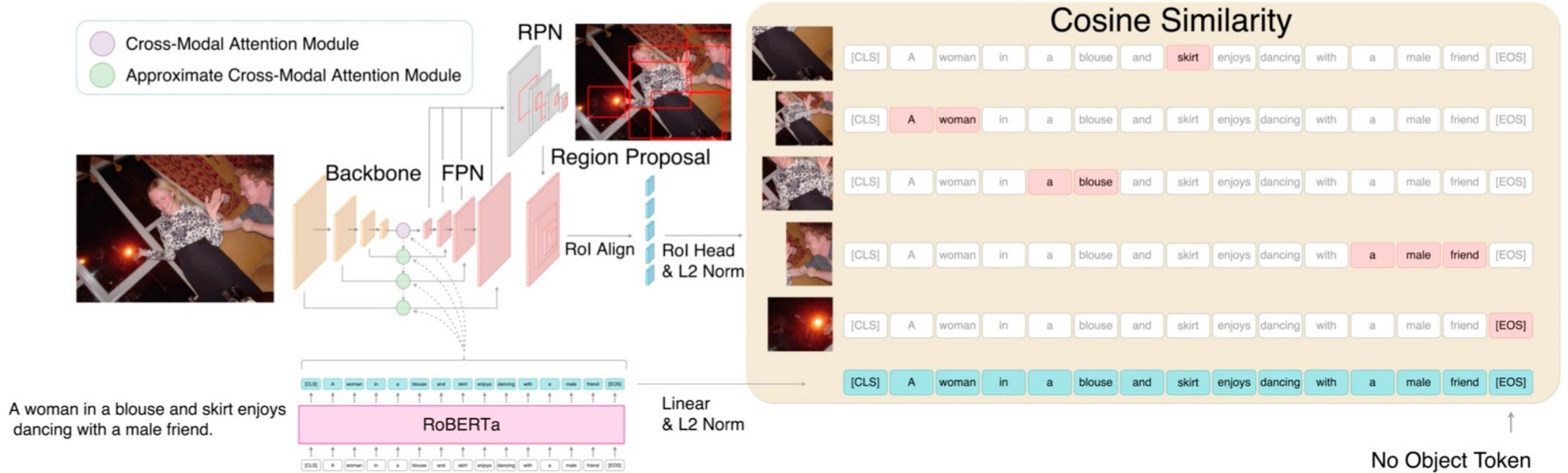


図 3 参照表現理解タスクの結果例：“the empty water glass” に対する予測が緑，正解が赤の矩形。左) 提案手法。右) MDETR.

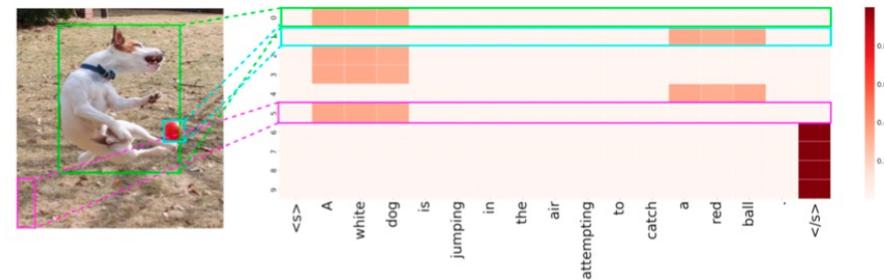


図 2 事前学習で獲得できる出力例.

3D世界に対応づいた言語理解

- 3D世界を捉える手段はいくつかある（RGBカメラ・深度センサ）
- 3D世界のデータと言語情報はまだあまり紐付いていない

(2D) Image



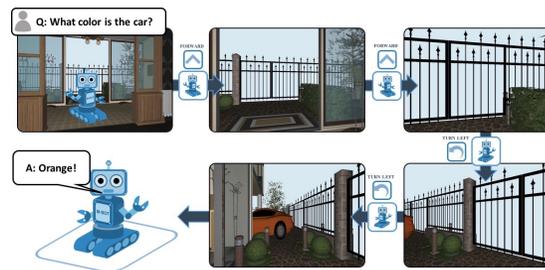
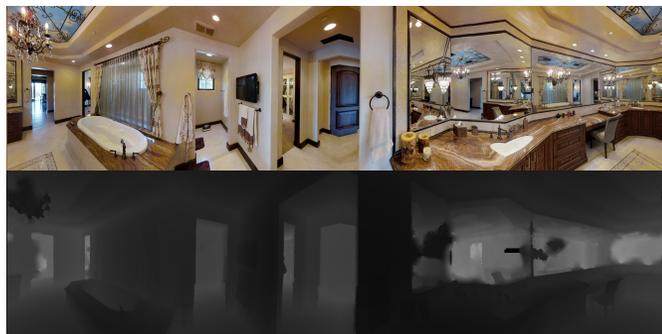
How many doors are open?

Visual QA

画像+QA

- 一枚の画像からQA
- 独立した画像から現実世界の知識が取れるのか？

Panorama RGB (+ Depth) Image



Embodied QA

ナビゲーション+画像+QA

- 3D環境中を物体の位置まで移動してQA
- まだあまり良いデータセットがなかった

3D Point Cloud Data



3D Data QA

3D世界情報+QA

- カメラやSLAM情報から得られた情報を統合してQA

Embodied QA datasets

Embodied Question Answering CVPR2018

- QA on House3D, based on SUNCG
- Limited datasize 😞 • Licensing issue 😞

	Environments	Unique Questions	Total Questions
train	643	147	4246
val	67	104	506
test	57	105	529

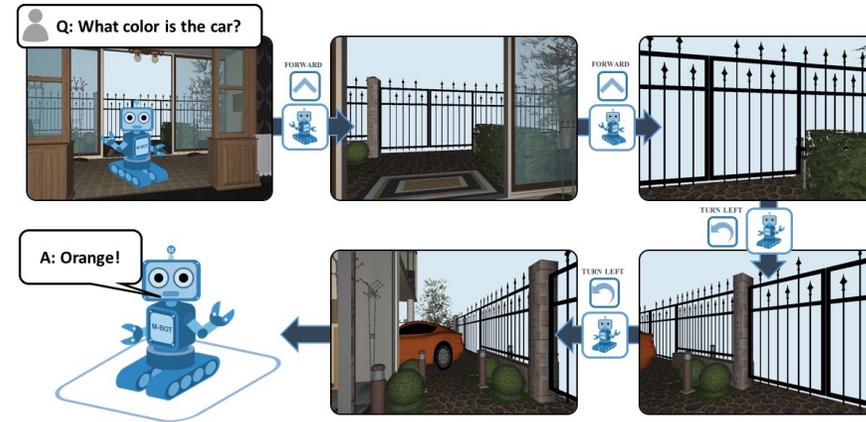
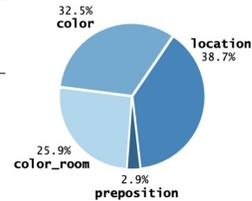


Figure 3: Overview of the EQA v1 dataset including dataset split statistics (left) and question type breakdown (right).

Generated from few templates
Limited data

Embodied Question Answering in Photorealistic Environments with Point Cloud Perception

CVPR2019

- Photorealistic EQA, based on matterport 3D scan data
- Very limited datasize 😞

	Homes	Floors	Total Qns.	Unique Qns.
train	57	102	767	174
val	10	16	130	88
test	16	28	239	112

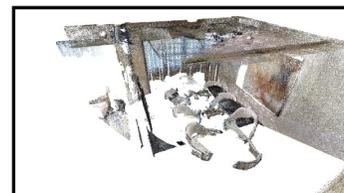
Table 1: Statistics of splits for EQA in Matterport3D



(a) RGB Panorama



(b) Mesh Reconstruction



(c) Point Cloud



(d) RGB-D Render

Generated from few templates
Very limited data

3D環境におけるQA

- 例えば "How many white chairs are in the kitchen?" という質問が与えられたエージェントは、SLAM等を利用しながらキッチン内部を探索するだろう
- 探索で見つかった椅子の色や同一性の判定が問題となる
- 従来の2D画像に頼るQAではなく将来的には、SLAM等から得られる観測情報を統合したQAが必要になるのでは？



3D モデリング上の質問応答例

- How many chairs are there on the brown table?
 - Answer: four
 - 適切な回答のためには視点を転回する必要がある
- What color is the chair?
 - この質問は物体が特定できない (不適當)

3D-QA dataset: ScanQA を作成

3D点群(3D-pcd) からQA + Object Localizationを行うタスクを提案
大規模データセットを作成した

Question + 3D-Scan

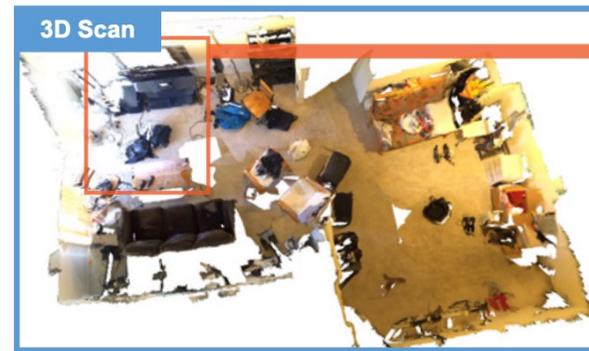


Q. Where is the medium sized blue suitcase laid?

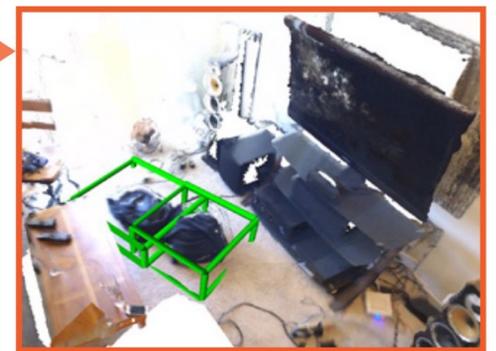
Answer + 3D-Bounding Box



A. in front of right bed



Q. What is sitting on the floor between the tv and the wooden chair?



A. 2 black backpacks

- 3D空間から質問に応じて物体位置を取得、質問応答を行うタスク
- 2Dのbaselineモデルには、広い部屋の中から正解の物体を含む画像を拾ってることが難しい

ScanQA: 3D Question Answering for Spatial Scene Understanding
Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, Motoki Kawanabe

CVPR2022

大規模 3D-QA dataset: ScanQA



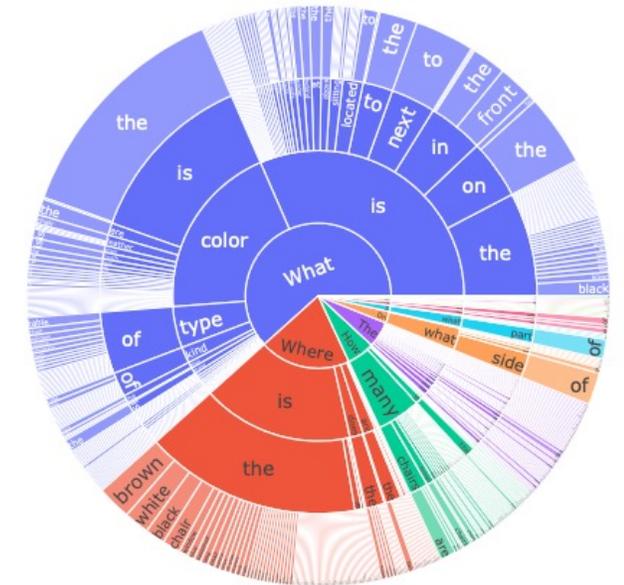
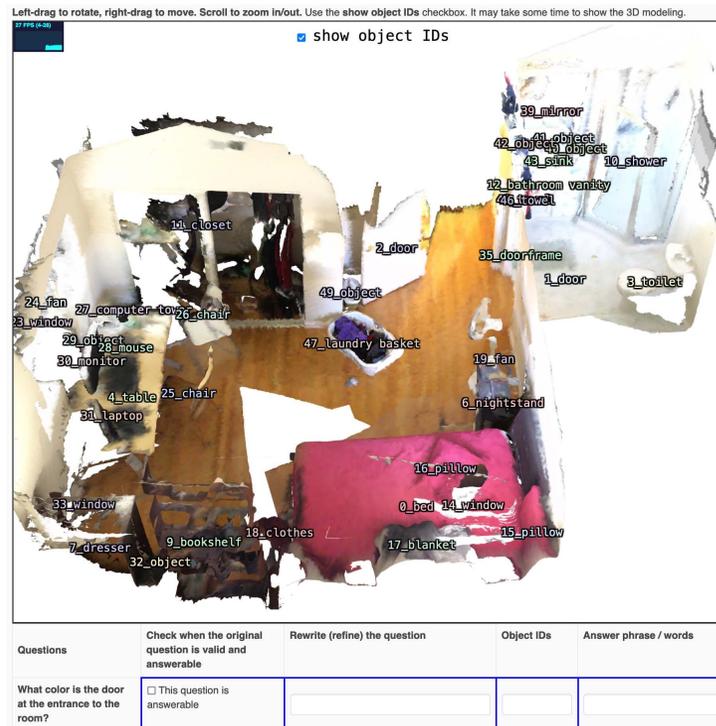
Underspecified questions

- Q: What is in the corner?
- Several objects at corners!
- Q: What color is the chair?
- Three chairs at the scene!

Valid questions

- Q: What is over the chair beneath the blackboard?
- Answer: jacket
- Q: What color is the office chair next to the desk with a monitor?
- Answer: green

Split	# Question	# Unique question	# 3D Scenes
Train	25,563	20,546	562
Val	4,675	4,306	71
Test w/ objects	4,976	4,552	70
Test w/o objects	6,149	5,484	97
Total	41,363	32,337	800

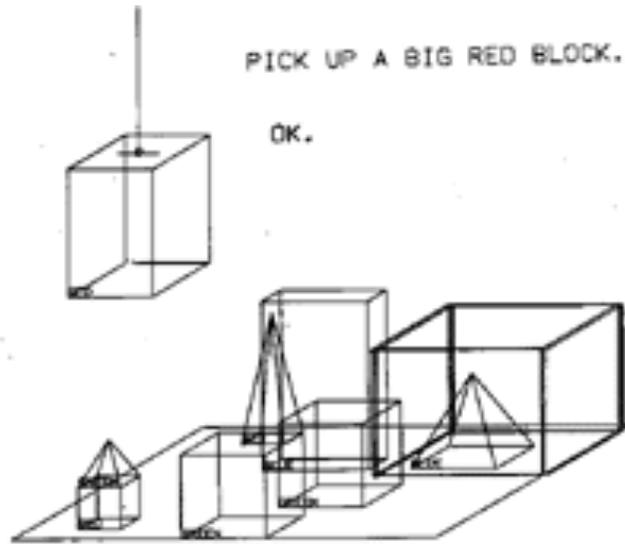


目次

1. はじめに
 1. Embodied AI : 身体に基づいたAI
 2. Generative language grounded policy (GLGP) on VLN
 3. SayCan: Do As I Can, Not As I Say (Google Robotics & Everyday Robotics, 2022)
2. 言語の役割
 1. 名付けによる世界の分節 (F. Saussure)
 2. 言語による指示と実世界との対応付け (L. Wittgenstein)
3. 言語モデル
 1. 歴史 (Markov process, n-gram, RNN)
 2. PaLM & Big-BENCH
 3. Chain of thought
 4. 基盤モデルたち (省略)
4. グランディングと参照表現理解
 1. 参照表現を理解する
 2. テキストに紐付いた物体検出
 3. 3D世界に対応づいた言語理解
5. 言語を使って指示を出す
 1. SHRDLU (T. Winograd, 1971)
 2. ナビゲーション : Vision & Language Navigation (VLN)
 3. プランニング : AI2THORとSayCan
6. まとめ : Language & Robotics のこれから
 1. 言語指示が有効なのはどのような課題か？
 2. 言語理解技術がどのようにロボティクスを助けるか？

言語を使って指示を出す

SHRDLU (Winograd, 1971)



Person: PICK UP A BIG RED BLOCK.
Computer: OK. (does it)
Person: GRASP THE PYRAMID.
Computer: I DON'T UNDERSTAND WHICH PYRAMID YOU MEAN.
Person: FIND A BLOCK WHICH IS TALLER THAN THE ONE YOU ARE HOLDING AND PUT IT INTO THE BOX.
Computer: BY "IT", I ASSUME YOU MEAN THE BLOCK WHICH IS TALLER THAN THE ONE I AM HOLDING.
Computer: OK. (does it)
Person: WHAT DOES THE BOX CONTAIN?
Computer: THE BLUE PYRAMID AND THE BLUE BLOCK.
Person: WHAT IS THE PYRAMID SUPPORTED BY?
Computer: THE BOX.
Person: HOW MANY BLOCKS ARE NOT IN THE BOX?
Computer: FOUR OF THEM.

<https://hci.stanford.edu/~winograd/shrdlu/>

Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions (Hatori *et al.* 2018, PFN)

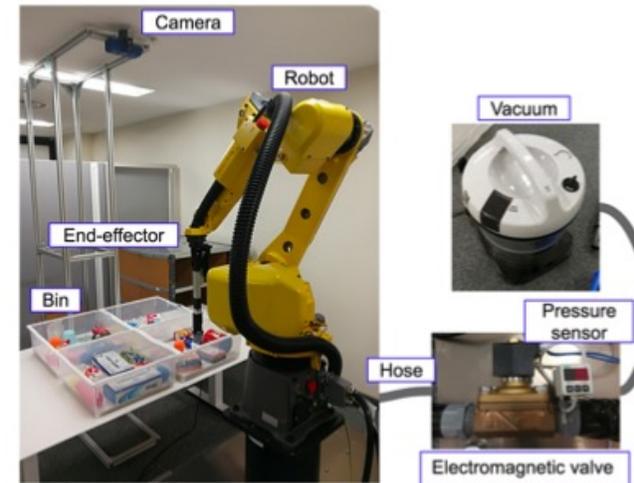
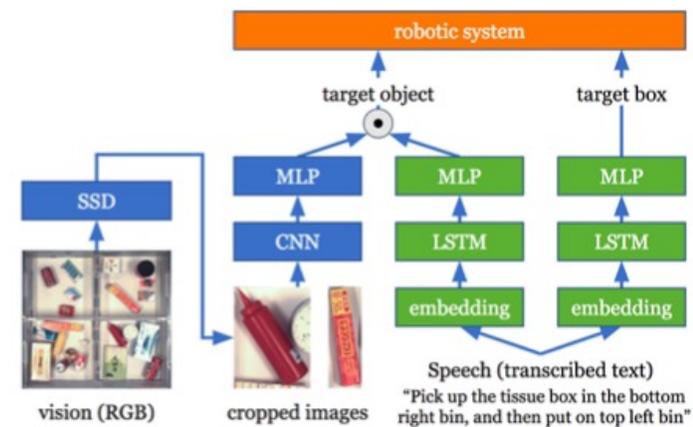


Fig. 5: Robot setup for experiments

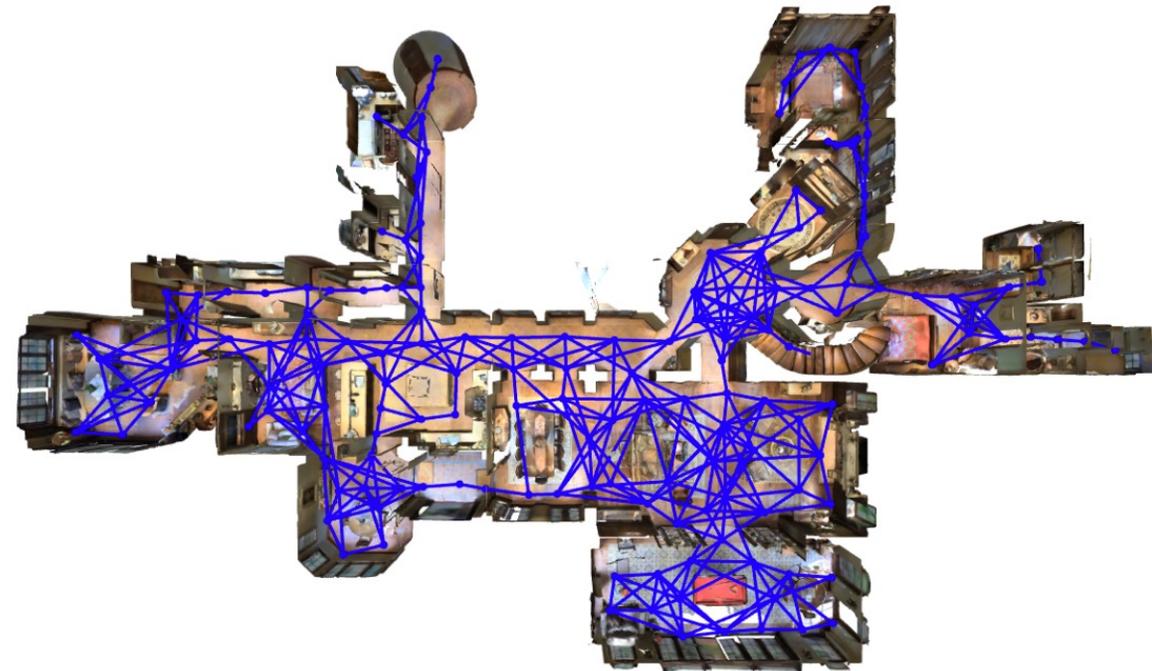


視覚と言語によるナビゲーション (VLN)



Leave the bedroom, and enter the kitchen. Walk forward, and take a left at the couch. Stop in front of the window.

- A navigation task for a robotic agent to reach the goal place following the textual instruction.
- Photorealistic environment.
Based on real houses, corrected with 3D scans.
- The dataset was released in 2017.
- Possible application to the robotic navigation

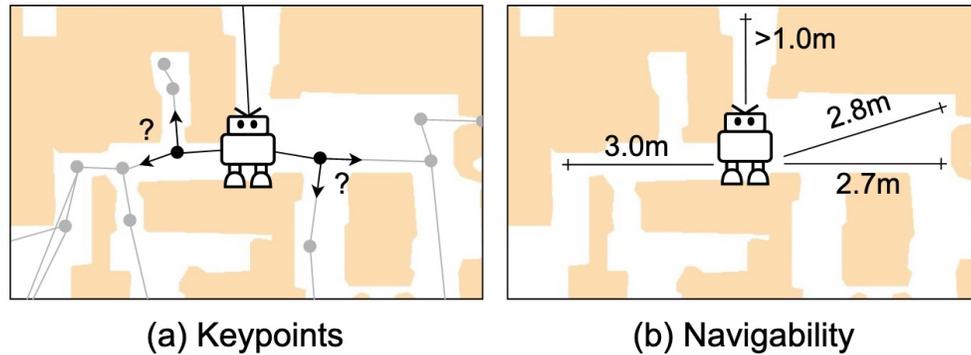


ナビゲーション: VLN から Habitat へ

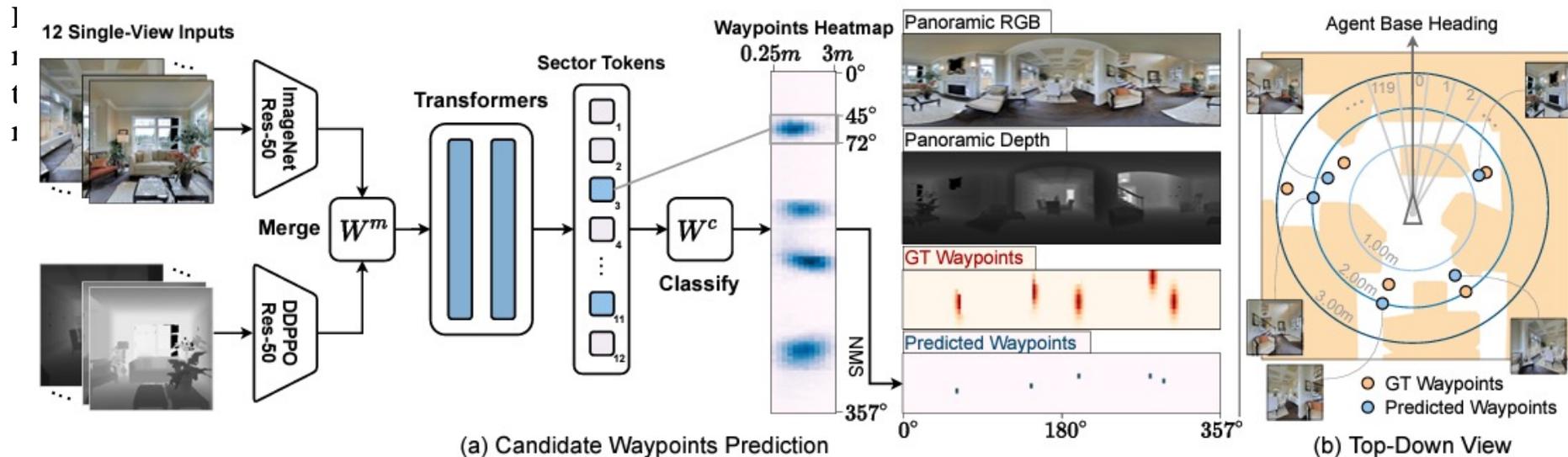


ナビゲーション: VLN-CE on Habitat

Bridging the Gap Between Learning in Discrete and Continuous Environments
for Vision-and-Language Navigation (Yicong *et al.* CVPR2022)



高いレベルのway-point選択
低いレベルでの連続動作生成



プランニング: ALFRED & SayCan



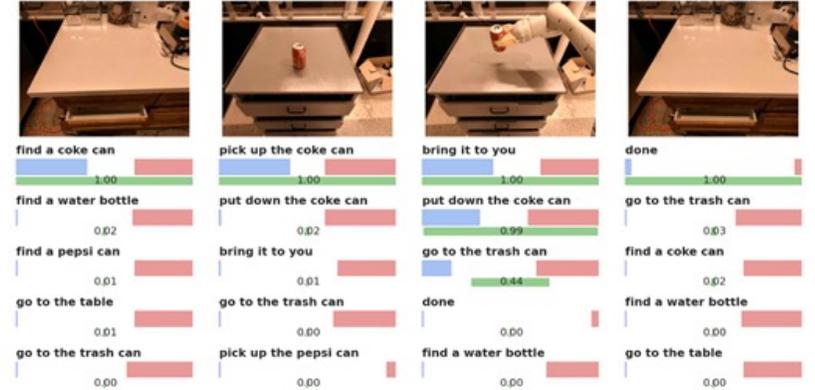
To heat a cup as well as place it in the fridge.

- Open the drawer
- Pick up a mug
- Open the microoven door
- Put the mug onto the microoven
- ...

Human: I spilled my coke, can you bring me a replacement?

Robot: I would
1. Find a coke can
2. Pick up the coke can
3. Bring it to you
4. Done

Language x Affordance
Combined Score

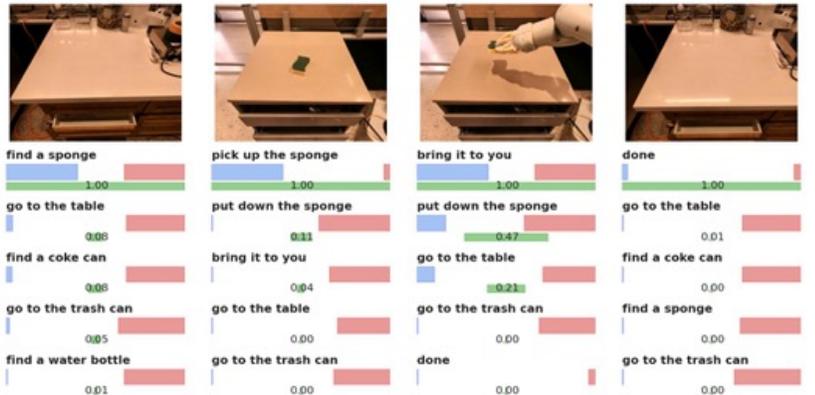


(a)

Human: I spilled my coke, can you bring me something to clean it up?

Robot: I would
1. Find a sponge
2. Pick up the sponge
3. Bring it to you
4. Done

Language x Affordance
Combined Score



(b)

現状では、実世界でできることを学習させたいのであれば、
仮想/現実ともにまだ動作/物体の集合が大きくはないという問題がある気がする
※仮想世界系では、最近3Dアセットのデータセットが充実してきている
限られた物体という制限は取り払われるかもしれないが...

目次

1. はじめに
 1. Embodied AI : 身体に基づいたAI
 2. Generative language grounded policy (GLGP) on VLN
 3. SayCan: Do As I Can, Not As I Say (Google Robotics & Everyday Robotics, 2022)
2. 言語の役割
 1. 名付けによる世界の分節 (F. Saussure)
 2. 言語による指示と実世界との対応付け (L. Wittgenstein)
3. 言語モデル
 1. 歴史 (Markov process, n-gram, RNN)
 2. PaLM & Big-BENCH
 3. Chain of thought
 4. 基盤モデルたち (省略)
4. グランディングと参照表現理解
 1. 参照表現を理解する
 2. テキストに紐付いた物体検出
 3. 3D世界に対応づいた言語理解
5. 言語を使って指示を出す
 1. SHRDLU (T. Winograd, 1971)
 2. ナビゲーション : Vision & Language Navigation (VLN)
 3. プランニング : AI2THORとSayCan
6. まとめ : Language & Robotics のこれから
 1. 言語指示が有効なのはどのような課題か？
 2. 言語理解技術がどのようにロボティクスを助けるか？

まとめ：Language & Robotics のこれから

1. 言語指示が有効なのはどのような課題か？

まず、指示とは (1) 制約であり拘束条件である、(2) ヒントにもなり、終状態を規定する。

言語による指示が非常に有効であるためには、以下の条件を満たす必要がある？

[1] 指示による制約数 \ll 動作によって可能な状態数
(環境条件)

[2] 理想とする終状態をシンボリックな制約を用いて書き下すことが現実的でない
(報酬条件)

[1]の条件が満たされていない状況は、タスクを解くために必要十分な指示を与えることが可能
直感的な操作以外での言語指示のメリットは薄いかもしれない
(2D/3Dグラウンディングはこの場合でも重要なまま残り続けるでしょう)

[2]の条件が満たされていない状況では、理想とする終状態へ向かう経路を絞り込める

はじめに：ロボットへの言語での指示

Google Robotics

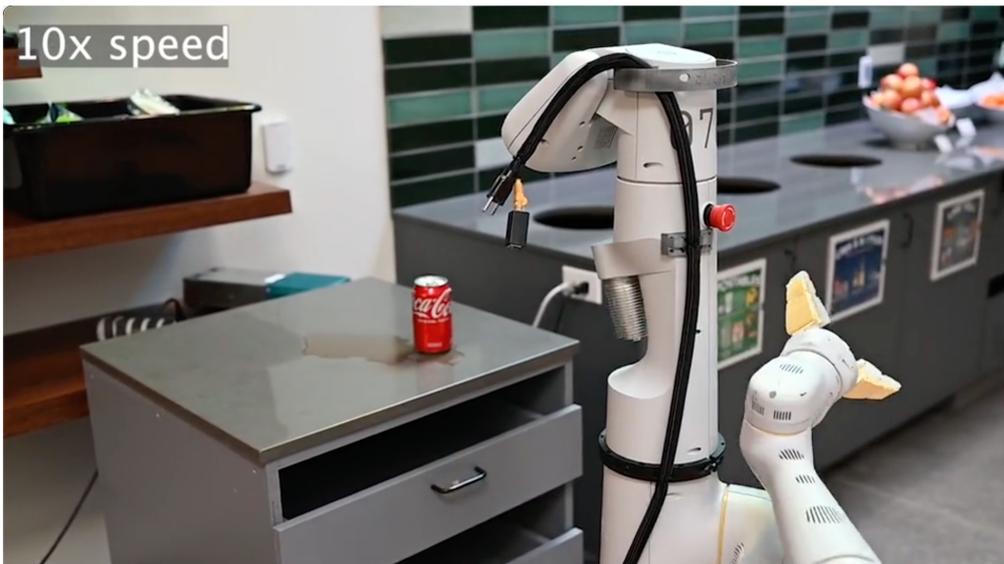
SayCan: Do As I Can, Not As I Say (2022)

私が言ったようにはなく
私ができるように動作しなさい

C.f. Do as I say, not as I do:

“私が教えるようには行動しなさい、
私がするようにではなく”

ジョン・セルデン「茶話」1654年



INPUT: I spilled my coke on the table, how would you throw it away and bring me something to help clean?

ROBOT:

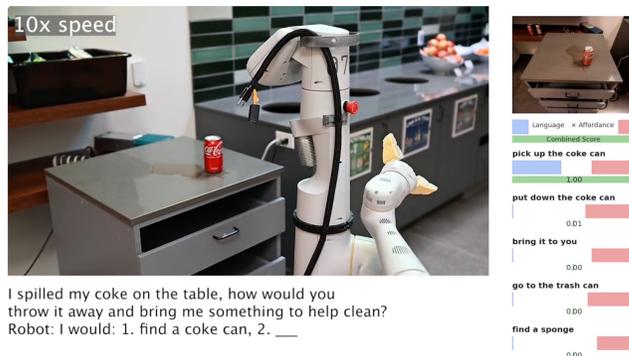
I spilled my coke on the table, how would you throw it away and bring me something to help clean?
Robot: I would: 1. find a coke can, 2. ___

まとめ：Language & Robotics のこれから

1. 言語指示が有効なのはどのような課題か？

SayCan: Do As I Can, Not As I Say (2022)

私が言ったようにではなく
私ができるように動作しなさい



ALFREDやSayCanは
言えること > やれること ？

画像生成 (DALLE-2)

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a photorealistic style in the style of Andy Warhol as a pencil drawing



DALL-E 2



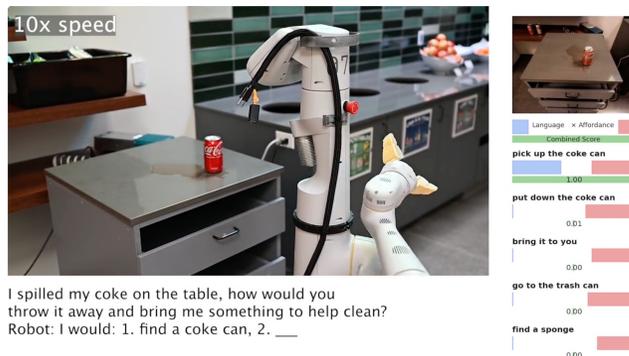
画像生成は
言えること < やれること ？

まとめ : Language & Robotics のこれから

1. 言語指示が有効なのはどのような課題か？

SayCan: Do As I Can, Not As I Say (2022)

私が言ったようにではなく
私ができるように動作しなさい



ALFREDやSayCanは
言えること > やれること ?

画像生成 (DALLE-2)

TEXT DESCRIPTION

An astronaut Teddy bears A bowl of soup

riding a horse lounging in a tropical resort in space playing basketball with cats in space

in a photorealistic style in the style of Andy Warhol as a pencil drawing



DALL-E 2



画像生成は
言えること < やれること ?

まとめ：Language & Robotics のこれから

1. 言語指示が有効なのはどのような課題か？

指示とは (1) 制約であり拘束条件である、(2) ヒントにもなり、終状態を規定する。

言語による指示が非常に有効であるためには、以下の条件を満たす必要がある（？）

[1] 指示による制約数 \ll 動作によって可能な状態数

（環境条件）

[2] 理想とする終状態をシンボリックな制約を用いて書き下すことが現実的でない

（報酬条件）

[1]の条件が満たされていない状況は、タスクを解くために必要十分な指示を与えることが可能

直感的な操作以外での言語指示のメリットは薄いかもしれない

（2D/3Dグラウンディングはこの場合でも重要なまま残り続けるでしょう）

[2]の条件が満たされていない状況では、理想とする終状態へ向かう経路を絞り込める

（将来の）実世界デバイスやロボットを自由な言語指示から動作させようとした場合
明らかに[1]と[2]をみたす課題になるだろう



現状の技術でも [1]と[2]をみたしうるLanguage & Robotics課題とは何か？

まとめ : Language & Robotics のこれから

2. 言語理解技術がどのようにロボティクスを助けるか？

SayCan (Instructing a robot)

Human: How would you bring me something that isn't a fruit?

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

Explanation: the user wants something to eat that isn't a fruit. An energy bar is not a fruit, so I will bring the user an energy bar.

Plan: 1. find(energy bar) 2. pick(energy bar) 3. find(user) 4. put(energy bar) 5. done().

them.

A: The last letter of "Lady" is "y". The last letter of "Gaga" is "a". Concatenating them is "ya". So the answer is ya.

入力では **something to eat** とは全く言っていない！
(推論としては完全に間違いである！)

しかし「フルーツじゃないものを持ってきて」と言われて、
食べられないものを持ってくる人はどのくらいいるだろうか？

to the answer is (b).

Sports Understanding

Following sentence
Joao Moutinho caught the
ball in the NFC
game.

Moutinho is a soccer player.
The NFC championship is part of
American football, not soccer. So the
answer is (a).

Flip (state tracking)

The coin is heads up. Maybelle flips
the coin. Malonda does not flip the
coin. Is the coin still heads up?

A: The coin was flipped by Maybelle.
So the coin was flipped 1 time, which
is an odd number. The coin started
heads up, so after an odd number of
flips, it will be tails up. So the answer
is no.

指示文での論理 (の欠陥) を
言語モデルが
"常識"により override している? (※)

この例では先の[2]の終状態を絞り込むために
言語モデルを使用していると言えるだろう

「言えること > やれること」タスクでも
グラウンディングや終状態の絞り込みで
LLMが役に立つ？