FUJITSU
shaping tomorrow with you

第2回国際ナレッジグラフ推論チャレンジ（IKGRC2024）開催報告

2024年3月9日

鵜飼 孝典(富士通株式会社＆産総研)

※本資料には、一部応募作品その他の図が掲載されています

提供データ

FUJITSU

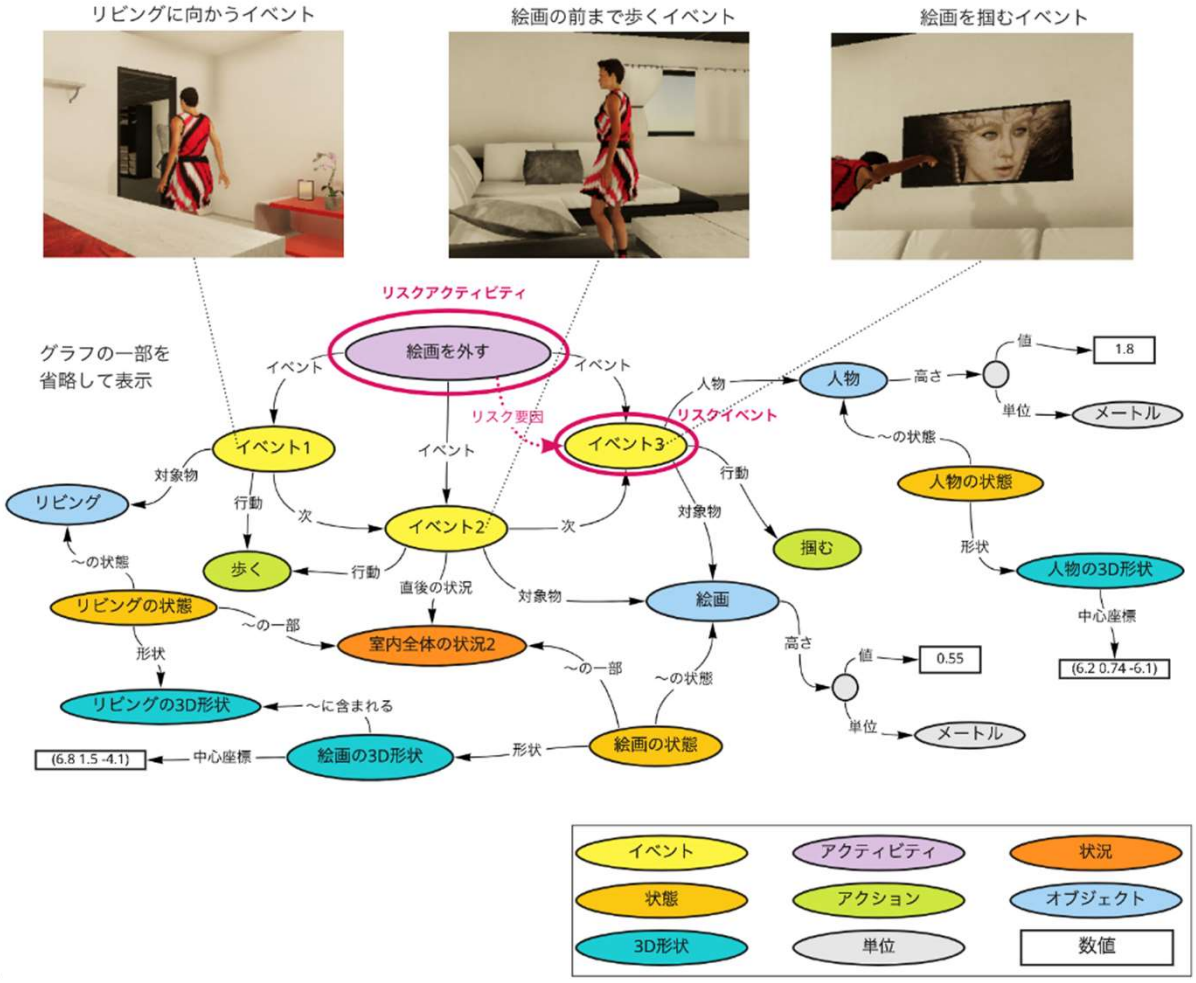
- 動画
  - mp4形式
  - 706種類の行動シナリオ：合計3535個の動画
  - 1つの行動シナリオにつき，異なる部屋の間取り（scene）で最小1～最大7パターンのデータ
  - ActionGenomeが提供するデータフォーマットでシーングラフデータを提供（新規）
- ナレッジグラフ
  - RDF形式
  - 動画に対応する706個のナレッジグラフ
  - N-TRIPLEデータを提供（新規：グラフニューラルネットワーク、埋め込み向け）
- 台本データ
  - txt形式
  - 動画とナレッジグラフを生成するための元データ
  - 行動のタイトルと簡単な文章説明を含む

- エピソード
  - 上記のアクティビティデータを3～7つ並べたもの
    - 一日の活動をイメージしたもの

提供されるデータセットの特徴

FUJITSU

- 仮想空間内でアバターの行動を表現するイベント系列と付随する環境とのインタラクション情報

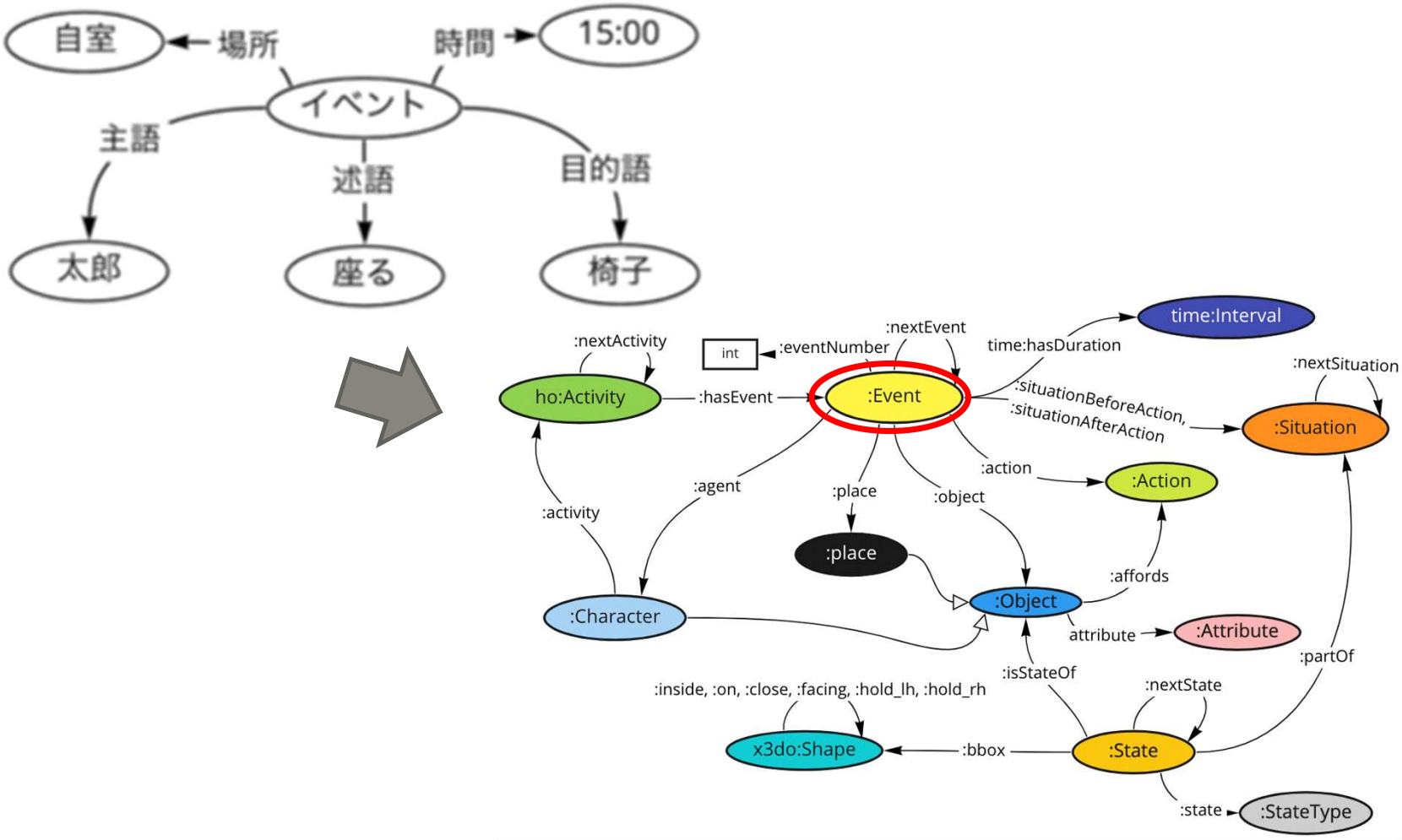


2

知識グラフ

FUJITSU

- ナレッジグラフを時空間的に生起する行動の記述に使用



問題

FUJITSU

- Q1：それぞれの部屋に何回入ったか？
  - アクティビティとアクティビティの間の移動は、最短で行われたとする
- Q2：それぞれのアクションを何回行ったか？
- Q3：キッチンに入った後行ったアクションはなにか
  - キッチンに入っていなければ、空
- Q4：キッチンに最初入る前行ったアクションはなにか
  - キッチンに入っていなければ、空
- Q5：いつ、どこでなにを持ったか
- Q6：初期から10秒後に何を行っているか？
- Q7：初期状態と10秒後のモノとモノの関係を抽出する
  - アクティビティとアクティビティの間の時間は 0 とする
- Q８：モノの状態変化を抽出する
  - 初期状態と20秒後で位置、状態が変わったものを抽出し、初期の位置、状態と20秒後の位置、状態を回答する

タスク

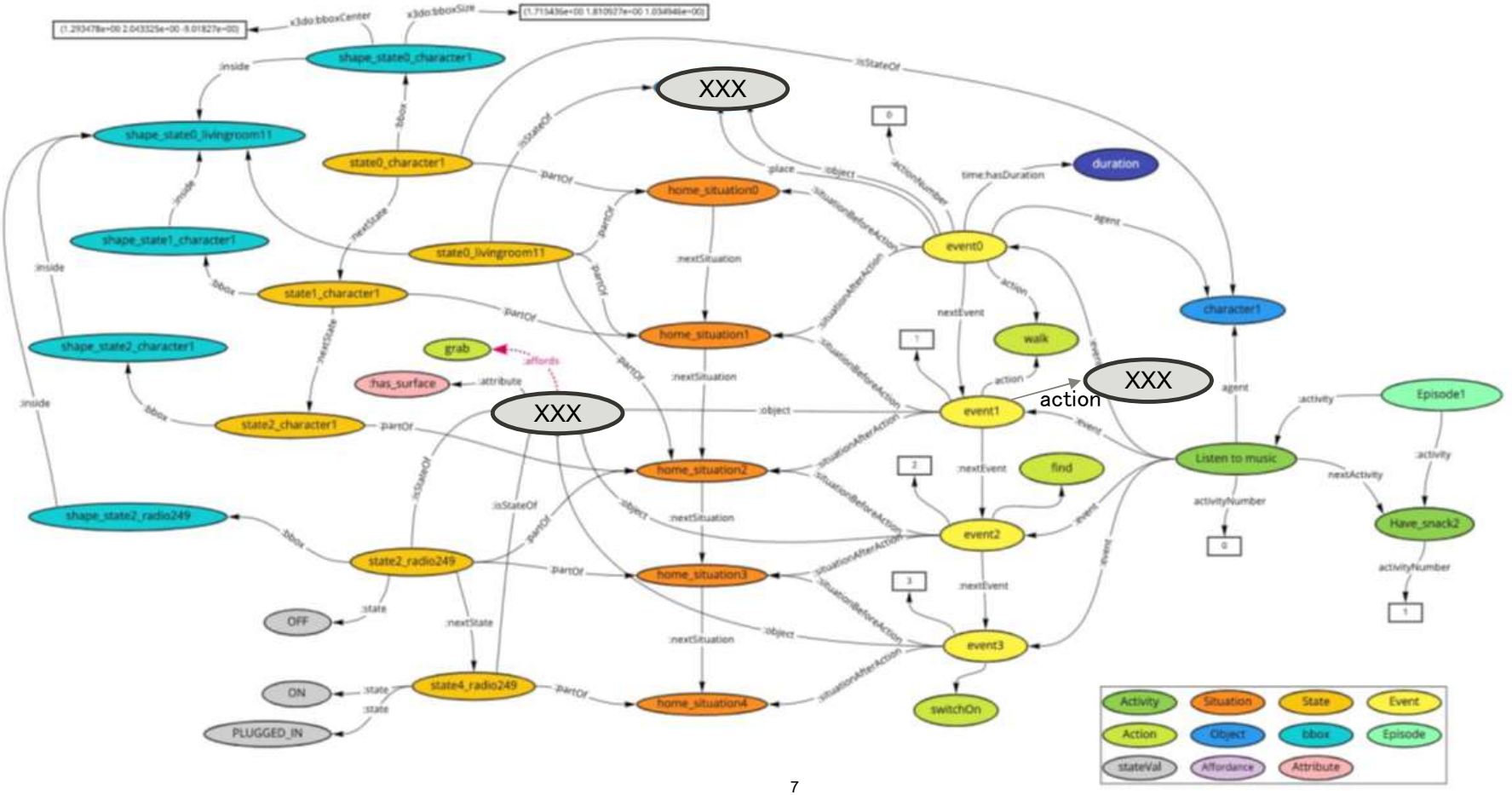
FUJITSU

- 生活行動を表現する動画、ナレッジグラフなどで構成されるマルチモーダルデータセットに対して設定されている設問に答える
  - フェーズ1：完全版データセットを対象に検索で答えられる
  - フェーズ2：欠損版データセットを対象に検索だけでは答えられない

欠損データについて

FUJITSU

- Placeを置き換える
  - 置き換える：Xでなにかわからないもの。ほかのものに置き換えるわけではない。
  - 20%, 50%, 100%
  - 今いる場所を置き換える
- さらにAction を置き換える
  - 20%, 50%, 100%
  - URLを置き換え、ラベルを削除する
- さらにObject (TargetObject..)を置き換える
  - 20%, 50%, 100%
  - URLを置き換え、ラベルを削除する（何かはわからないけど、なにかあることはわかる）
  - 動作にかかわるものだけを対象にする
  - 置き換えたObjectはステータス、リレーションも削除する


XXX
XXX
XXX
action
shape_state0_character1
shape_state0_livingroom11
shape_state1_character1
shape_state2_character1
shape_state2_radio249
state0_character1
state0_livingroom11
state1_character1
state2_character1
state2_radio249
state4_radio249
home_situation0
home_situation1
home_situation2
home_situation3
home_situation4
event0
event1
event2
event3
duration
walk
grab
has_surface
find
switchOn
character1
Listen to music
Episode1
Have_snack2
OFF
ON
PLUGGED_IN
Activity
Situation
State
Event
Action
Object
bbox
Episode
stateVal
Affordance
Attribute

スケジュール

FUJITSU

- 2023年12月22日　応募締め切り
- 2024年1月22日　追加資料締め切り
- 2024年2月5-7日 International Conference on Semantic Computing 併設のワークショップとして最終発表会を開催
- 2024年3月9日　Best　Prizeを発表

応募作品紹介

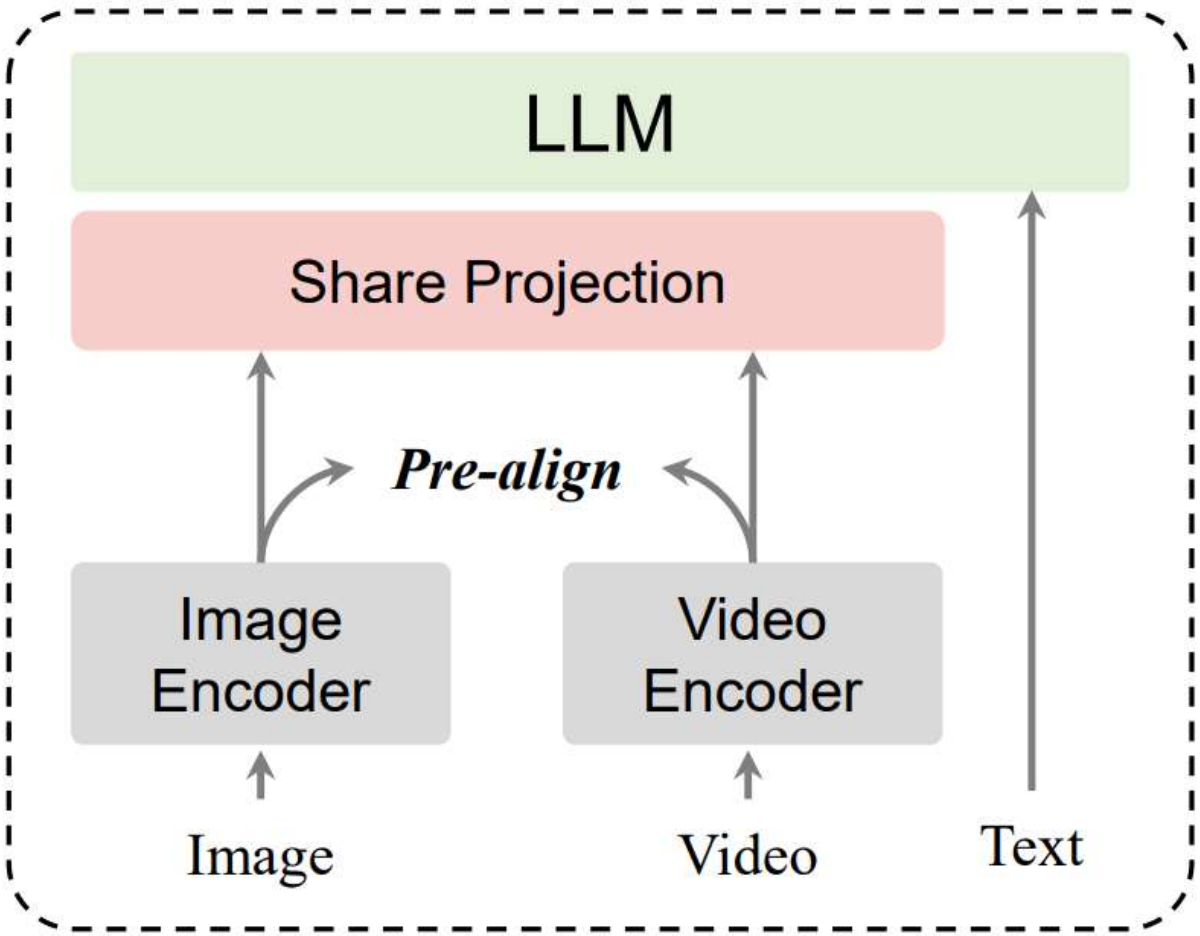
FUJITSU

- Zero-Shot Query Experiments in Knowledge Graph Reasoning Challenge for Older Adults Safety
- Event Prediction in Event-Centric Knowledge Graph Using BERT
- Prediction of Actions and Objects through Video Analysis Using Stepwise Prompt
- Prediction of actions and places by the time series recognition from images with Multimodal LLM

Video-LLaVa

FUJITSU

- LLM（Vicuna7B）に動画、画像、テキストの組み合わせを追加学習させたモデル



Zero-Shot Query Experiments in Knowledge Graph Reasoning Challenge for Older Adults Safety

FUJITSU

- チャレンジのタスク1で設定されている8つの質問について、Video-LLaVaを用いたZero-shotの問い合わせを行った。
- 単純に質問を入力する他、Zero-shot Chain-of-ThoughtとEmotionPromptの2種類のプロンプトを用いて実験を行った。
- Zero-shotでVideo-LLaVaにできるだけ詳細な説明を出力させた。
  - 詳細な説明が得られた後，そこから設定されている質問に答えることを想定している。
  - 3つの動画を対象とした予備実験を行い，8つの質問のうち1問で正解を導いた。
  - 単純なZero-shotではタスク1に正答することが困難であることを示し、今後の課題について議論した。

FUJITSU

Input the questions as they are

- Movie：Read book1_0.mp4 (a character rear view)
  - [WALK] <livingroom> (1), [WALK] <book> (1), [GRAB] <book> (1), [WALK] <sofa> (1), [SIT] <sofa> (1), [READ] <book> (1)

Question	Generated text	Evaluation
How many times did he enter each room?	He entered each room twice.	X
How many times did he perform each action?	He performed each action twice.	X
What action did he take after entering the kitchen?	He sat down on the couch after entering the kitchen.	X
What action did he take before entering the kitchen for the first time?	He walked into the living room before entering the kitchen for the first time.	X
When, where, and what did he pick up?	He picked up a remote control from the couch and walked into the living room.	X
What are he doing 10 seconds after the start?	He is walking into the living room.	X
Extract the relationship between objects at the initial state and after 10 seconds.	The man is walking into the living room.	X
Extract changes in the state of objects from the initial state to 20 seconds later.	The man is walking into the living room.	X

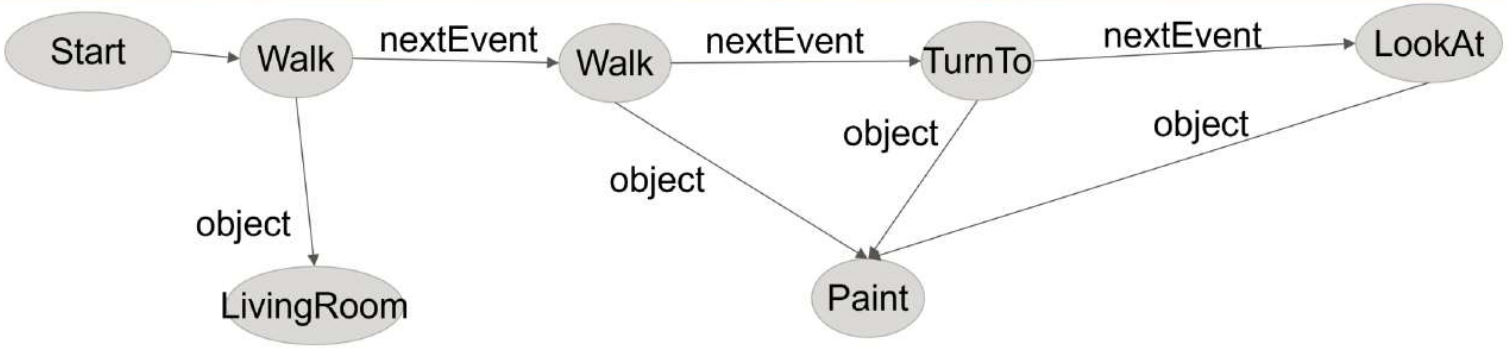
6

Event Prediction in Event-Centric Knowledge Graph Using BERT

FUJITSU

- チャレンジのタスク2に取り組み、不完全なナレッジグラフからアクションを推定して補完する手法を提案した。
- 提供されているイベント中心ナレッジグラフを典型的なナレッジグラフの構造に変換し、グラフ上をランダムウォークすることでシーケンスを生成する。
- 不完全なナレッジグラフでは一部のエンティティがマスクされているため、生成されたシーケンス中にマスクされている単語が出現する。
- このマスクされた単語をBERTを用いて予測することで、ナレッジグラフの欠損エンティティの補完を行う。
- 既存のナレッジグラフの埋め込みモデル（TransE、 ComplEx）を使用したベースラインを上回った。

Prediction by applying a hole-filling problem



- Generating graph walks
  - Start→nextEvent→Walk
  - Walk→object→LivingRoom
  - Start→nextEvent→Walk→object→LivingRoom→nextEvent→Walk
  - Walk→object→Paint→nextEvent→TurnTo→object→Paint
  - TurnTo→object→Paint→nextEvent→LookAt

Prediction of Actions and Objects through Video Analysis Using Stepwise Prompt

FUJITSU

- タスク1
  - ナレッジグラフのグラフ検索により質問に応答するアプローチ
  - 個別にSPARQLクエリを設計することで正確の回答を得た。
- タスク2
  - CLIP、GPT-4V、Video-LLaVaを使用し、欠損している動作(Action)とオブジェクト(Object)を動画から補完する複数のアプローチを提案した。
  - 動画からActionやObjectが欠損している時間のフレームを抽出
  - CLIPを用いて各フレームにおけるAction、Objectを推論
  - 次にGPT-4Vを用いて画像フレームとプロンプトにより欠損しているActionを予測
  - GPT-4Vの代替モデルとしてVideo-LLaVAを用いた欠損データの推論能力の検証実験を行った。
  - 結果として、Video-LLaVAが欠損データの予測に適しているという結論を得た。

FUJITSU

Task 2
Action Reasoning

If action is missing and object is not missing, reasoning is done as following steps

1. From sentences combining all ***actions*** and ***object*** only those combinations that result in natural sentences are extracted using GPT-4V (in advance).

all actions **object** **create sentences** **extract natural sentences**

grab
drink
turn to
⋮

\+ glass

He grabs glass.
He drinks glass.
He turns to glass.
⋮

GPT4-V

He grabs glass.
He reads glass.

2. Inference using information about the ***objects*** and the sentences extracted in 1.

Movie of event (Viewpoint 0)

\+

prompt

From the below sentences select only one that adequately describes what the person in the video is doing .
- **He grabs glass.**
- **He drinks glass.**

Video-LLaVA

He grabs glass.

17

Prediction of actions and places by the time series recognition from images with Multimodal LLM

FUJITSU

- タスク1
  - ナレッジグラフのグラフ検索により質問に応答するアプローチ
  - 個別にSPARQLクエリを設計することで正確の回答を得た。
- タスク2
  - Detectron2を用いて画像中の人物を検出
  - その人物を強調した画像をLLaVAに入力することでActionを予測する手法を提案
  - LLaVaに画像を入力する際に時系列性が失われることに対処するため、時系列的に連続するフレーム画像を1枚に連結した画像を作成し、これを入力することでActionとPlaceを予測する手法を提案した
  - 単一のフレームを使用する手法と比較して推論精度が0.23ポイント向上し、一定の有効性を持つことが示された

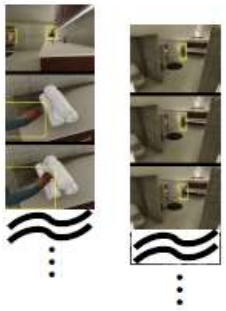
FUJITSU

Reason on images created by concatenating sequential images from the video

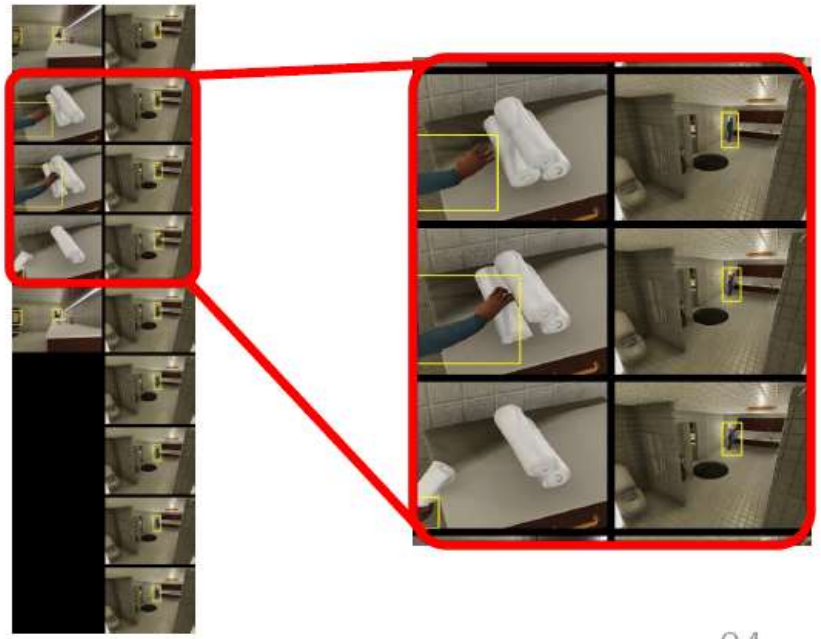
Images used

- **Single:** image at 66% of Duration

- **Vertically:** concatenated time series images



- **Matrix:** concatenated time series images



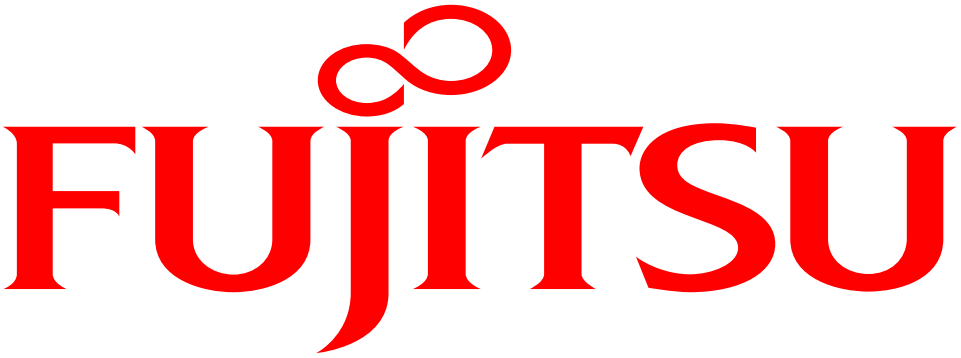
24

Best Prize

FUJITSU

- Prediction of actions and places by the time series recognition from images with Multimodal LLM
  - Tomohiro Ogawa, Kango Yoshioka, Ken Fukuda, and Takeshi Morita

- 彼らの研究は、タスク1とタスク2の両方に取り組み、タスク2では複数の画像を組み合わせて動的要素を識別するなど、さまざまな方法でLLMを使用することに優れていた。



shaping tomorrow with you