# Annotating situated actions in dialogue

*Christopher Tam, Richard Brutti, Kenneth Lai, James Pustejovsky*
Brandeis University

DMR @IWCS 2023

June 20, 2023

Brandeis

Chris Tam

Richard Brutti

Kenneth Lai

James Pustejovsky

Brandeis

# Talk Outline

- Introduction
  - Motivation and discussion
- Background
  - Developments in action annotation
  - AMR and extensions
- Approach
  - Example datasets and annotations
- VoxML specification
  - Subevents and aspect
  - Common ground simulation
- Future Work
  - Challenges and applications

Brandeis

# Talk Outline

- **Introduction**
  - **Motivation and discussion**
- Background
  - Developments in action annotation
  - AMR and extensions
- Approach
  - Example datasets and annotations
- VoxML specification
  - Subevents and aspect
  - Common ground simulation
- Future Work
  - Challenges and applications

Brandeis

# Introduction

- Why are actions important for dialogue?
- Actions contribute to context
  - Anaphora
    - [*lifts pencil*] “I used this for the sketch.”
    - “My brother said ‘thumbs up’!”
  - Bridging
    - “I went to the store today.” [*takes fruit out of a grocery bag*]

Brandeis

# Introduction

- Why are actions important for dialogue?
- Actions change the state of the world
  - Actions can add, modify, or delete items in the common ground
  - Tracking objects and actions in the environment is necessary for situated grounding

Brandeis

# Introduction

- What do we need to represent?
  - Multimodal interaction - how speech, gesture, and action build on each other
  - Common ground tracking - how actions update object locations and cause physical transformations
  - Lexical aspect - how actions progress over time, and how to represent them temporally

Brandeis

# Talk Outline

- Introduction
  - Motivation and discussion
- **Background**
  - **Developments in action annotation**
  - **AMR and extensions**
- Approach
  - Example datasets and annotations
- VoxML specification
  - Subevents and aspect
  - Common ground simulation
- Future Work
  - Challenges and applications

Brandeis

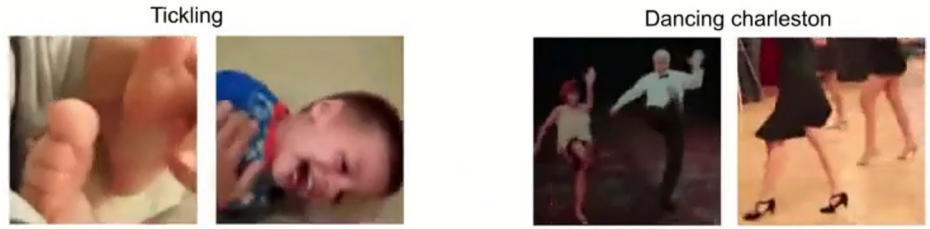
# Background - evolving datasets of action annotations

**Kinetics**

Charades

VidSitu

Kinetics - example classes




Only one label per video clip

-ignores simultaneity of actions
-no semantic roles / intentions
-no dialogue

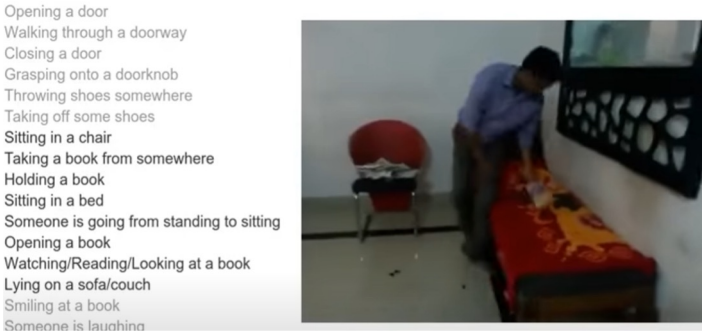
Brandeis

# Background - evolving datasets of action annotations

Kinetics

**Charades**

VidSitu




Multiple time-stamped labels per video clip

-simultaneous actions
-semantic roles, but not explicitly labeled
-no dialogue

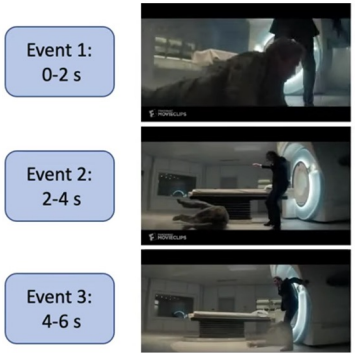
Brandeis

# Background - evolving datasets of action annotations

Kinetics

Charades

**VidSitu**




| crouch (to bend forward) | |
|---|---|
| Arg0 (entity crouching) | man in black suit and tie |
| ArgM (direction) | forward |
| ArgM (location) | hospital ward |

| drag ((try to) cause motion) | |
|---|---|
| Arg0 (dragger) | mri machine |
| Arg1 (thing dragged) | man in black suit and tie |
| ArgM (direction) | backward |
| Scene | hospital ward |

| explode (go boom) | |
|---|---|
| Arg1 (thing exploded) | mri machine |
| ArgM (location) | hospital ward |

PropBank descriptions of actions

-ignores simultaneity, inaccurate timestamps
-explicit semantic roles, can convert to AMR
-”speaking” actions, but no dialogue content

Brandeis

# Background - AMR

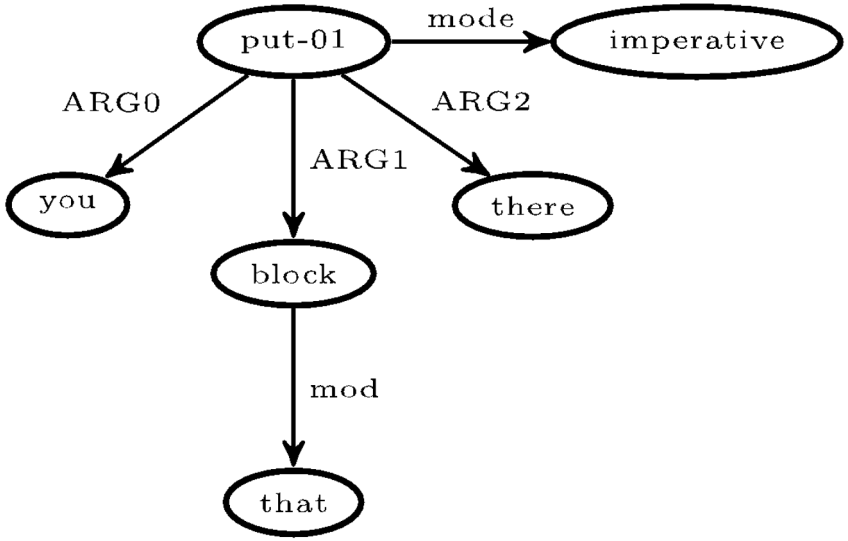
- Abstract Meaning Representation (AMR) is a graph-based meaning representation that expresses the meaning of a sentence in terms of its predicate-argument structure
  - Relatively easy to annotate
  - Readable by both humans and machines
  - Existing community of researchers

Brandeis

# Background - AMR

- Put that block there.

```
(p / put-01
    :mode imperative
    :ARG0 (y / you)
    :ARG1 (b / block
        :mod (t / that))
    :ARG2 (t2 / there))
```




Brandeis

# Background - AMR

We can use AMR to capture roles across multiple modalities with **MS-AMR**

- Move that tower about one block over.

```
(m / move-01 :mode imperative
    :ARG0 (y / you)
    :ARG1 (t / tower
        :mod (t2 / that))
    :ARG2 (a / about
        :op1 (q / distance-quantity
            :unit (b / block)
            :quant 1))
    :direction (o / over))
```

- (left arm: move, left; left hand: front, claw;)

```
(i / icon-GA
  :ARG0 (s / signaler)
  :ARG1 (s2 / slide-01
          :ARG1 (t / tower)
          :direction (l / left))
  :ARG2 (a / actor))
```

Brandeis

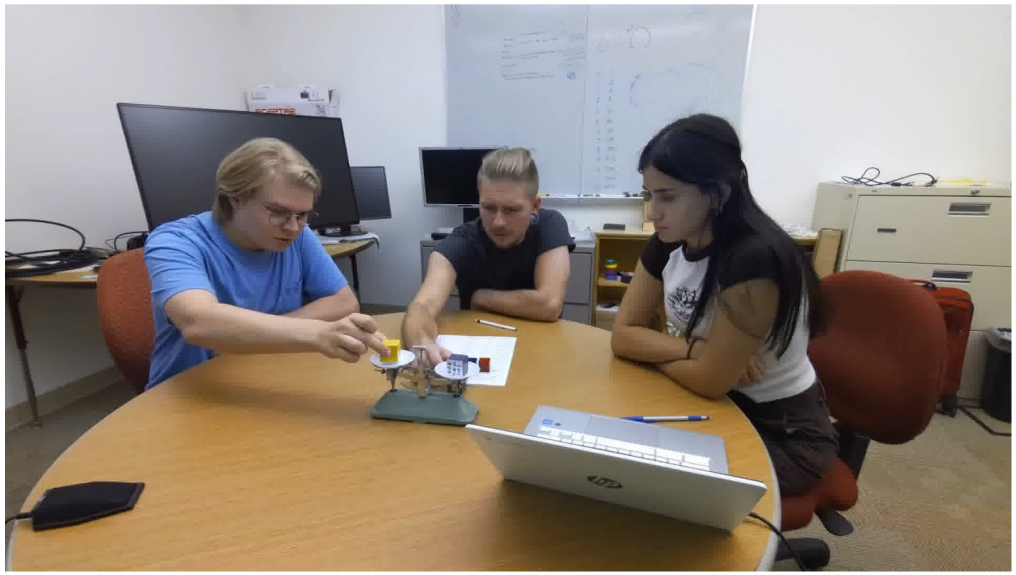
# Talk Outline

- Introduction
  - Motivation and discussion
- Background
  - Developments in action annotation
  - AMR and extensions
- **Approach**
  - **Example datasets and annotations**
- VoxML specification
  - Subevents and aspect
  - Common ground simulation
- Future Work
  - Challenges and applications

Brandeis

# Annotation Approach - Weights Task Corpus

- Designed to elicit teamwork between participants (based in collaboration frameworks)
- 2-3 people must determine the weights of the wooden blocks
- Participants have access to blocks, a scale, a worksheet, and a computer with a survey



Brandeis

# Annotation Approach - Weights Task

Participants speak, gesture, and interact with each other and the available objects to determine the weight of the blocks.

Brandeis

# Annotation Approach - Weights Task

Multimodal dataset:

speakers can refer to past events (cataphor - what does “did” mean?)

Brandeis

# Annotation Approach - Weights Task

Multimodal dataset:

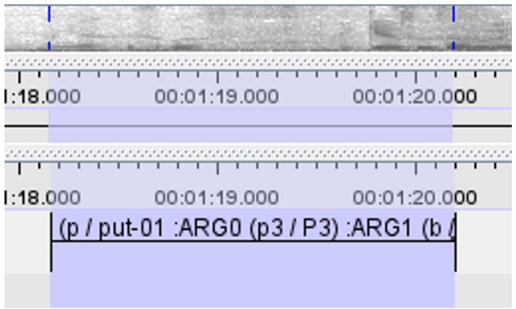
actions are suggested and confirmed by dialogue acts

they also function as deictic gestures

Brandeis

# Annotation Approach - Weights Task

```
[P3 puts the red block on the scale]
(p / put-01
    :ARG0 (p1 / P3)
    :ARG1 (b / RedBlock)
    :ARG2  (s / Scale))
```




Brandeis

# Annotation Approach - Weights Task

```
[P1 puts the green block on the scale]
(p / put-01
    :ARG0 (p1 / P1)
    :ARG1 (g / GreenBlock)
    :ARG2  (s / Scale))

"We did it (already)."
(d / do-02
    :ARG0 (w / we)
    :ARG1 (i / it))
```

*coreference and part-whole

# Annotation Approach - Epic Kitchens


- Spontaneous first-person recordings of individuals in kitchens
- Typically a single participant with little speech

Brandeis

# Annotation Approach - Epic Kitchens

Participants carry out daily actions, using a wide variety of tools and objects that continuously update the common ground

Brandeis

# Annotation Approach - Epic Kitchens

Actions in this domain are more complex:

Can be annotated in many different ways

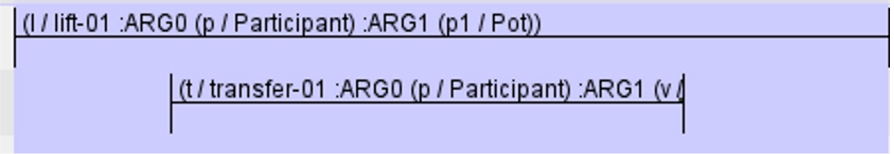
Have meaningful physical results to the common ground (soaped, cut, boiled)

Brandeis

# Annotation Approach - Epic Kitchens

```
(l / lift-01
        :ARG0 (p / Participant)
        :ARG1 (p1 / Pot))


(t / transfer-01
    :ARG0 (p / Participant)
    :ARG1 (v / Vegetables)
    :ARG2 (b / Bowl)
    :ARG3 (p1 / Pot)
    :instrument
    (c / Chopsticks))
```




Brandeis

# Talk Outline

- Introduction
  - Motivation and discussion
- Background
  - Developments in action annotation
  - AMR and extensions
- Approach
  - Example datasets and annotations
- **VoxML specification**
  - **Subevents and aspect**
  - **Common ground simulation**
- Future Work
  - Challenges and applications

Brandeis

# Using VoxML - introducing subevents and aspect

```
(p / put-01
      :ARG0 (p / participant)
      :ARG1 (b / RedBlock)
      :ARG2 (t / table))
```

$$
\begin{bmatrix}
\textbf{put} & \\
\text{LEX} = & \begin{bmatrix} \text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event} \end{bmatrix} \\
\text{TYP} = & \begin{bmatrix} \text{HEAD} = \textbf{transition} \\ \text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{bmatrix} \\ \text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while((\neg at(y,z) \wedge hold(x,y)), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)] \end{bmatrix} \end{bmatrix}
\end{bmatrix}
$$

lexical aspect

semantic roles

subevent structure

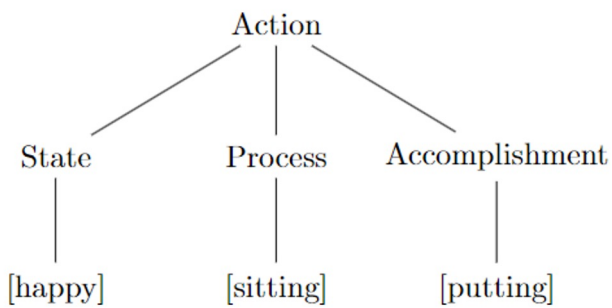
Figure 3: An example VoxML program corresponding to the PropBank predicate *put-01*.

Brandeis

# Using VoxML - the aspect type taxonomy

An action taxonomy gives general descriptions of how annotations should encode temporal information

States, processes, achievements, accomplishments - different ways of representing actions through time




Interval

*processes*
e.g. move

Point

achievement
e.g. arrive

Interval + Point

accomplishment
e.g. put

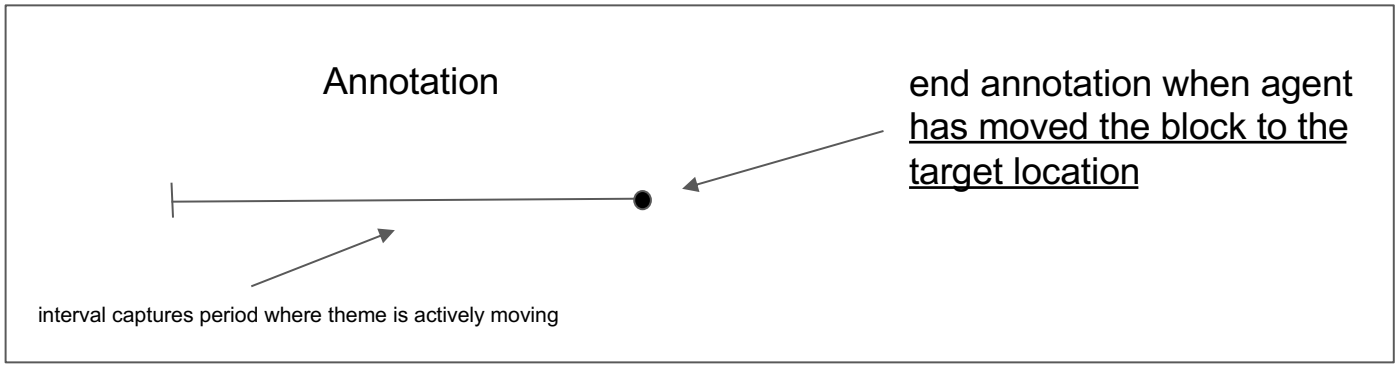
Brandeis

# Using VoxML - the aspect type taxonomy

$$
\begin{bmatrix}
\textbf{put} \\
\text{LEX} = \begin{bmatrix} \text{PRED} = \textbf{put} \\ \text{TYPE} = \textbf{transition\_event} \end{bmatrix} \\
\text{TYP} = \begin{bmatrix}
\text{HEAD} = \textbf{transition} \\
\text{ARGS} = \begin{bmatrix} \text{A}_1 = \textbf{x:agent} \\ \text{A}_2 = \textbf{y:physobj} \\ \text{A}_3 = \textbf{z:location} \end{bmatrix} \\
\text{BODY} = \begin{bmatrix} \text{E}_1 = grasp(x,y) \\ \text{E}_2 = [while((\neg at(y,z) \wedge hold(x,y)), move(x,y)] \\ \text{E}_3 = [at(y,z) \rightarrow ungrasp(x,y)] \end{bmatrix}
\end{bmatrix}
\end{bmatrix}
$$

Figure 3: An example VoxML program corresponding to the PropBank predicate *put-01*.

<u>resultant subevent</u> implies a change to the common ground




Brandeis

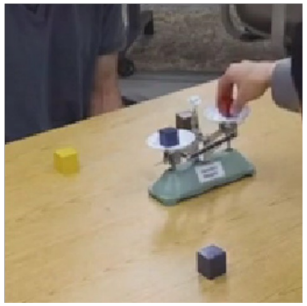
# Using VoxML - recreating the common ground

How can we give machines persistent understanding of dialogue?

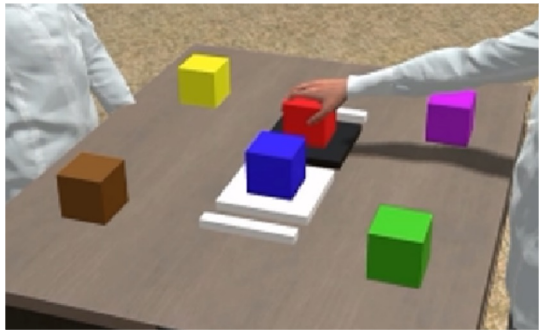
We can model common ground as a set of propositions, changing over time:

{Red block is on the scale
Blue block is on the scale
Green block is on the table
…
P2 is grabbing the red block}

Simulation reflects reality:



Weights Task Video



VoxSim simulation

Commands are run through executable VoxML structures:

P1: put(RedBlock, on(Scale))

Brandeis

# Talk Outline

- Introduction
  - Motivation and discussion
- Background
  - Developments in action annotation
  - AMR and extensions
- Approach
  - Example datasets and annotations
- VoxML specification
  - Subevents and aspect
  - Common ground simulation
- **Future Work**
  - **Challenges and applications**

Brandeis

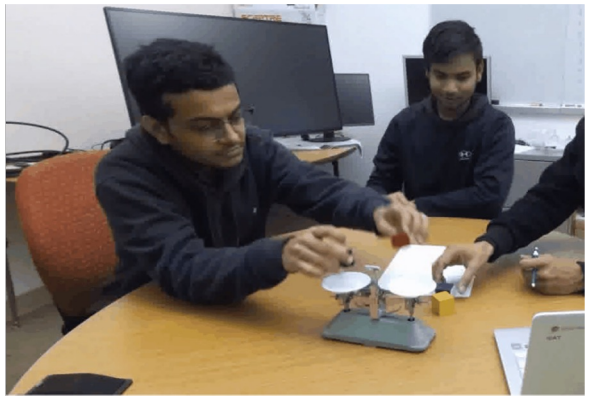
# Future work: Challenges - Simultaneity

*how many actions are occurring in this frame?*

Brandeis

# Future work: Challenges - Simultaneity

*lifting pan while adjusting heat*



*weighing two blocks at once*

Brandeis

# Future work: Challenges - Granularity/Vocabulary

**Nouns can be coreferenced, but many verbs have similar senses (e.g., scrub / wash / wipe)**

Choice when deciding the granularity of actions:

**- annotate “sit” as an interval (state), or annotate “sit_down” / “stand_up” (two transitions)?**

Annotation guidelines must have **a strict set of predicates and entities under the domain**

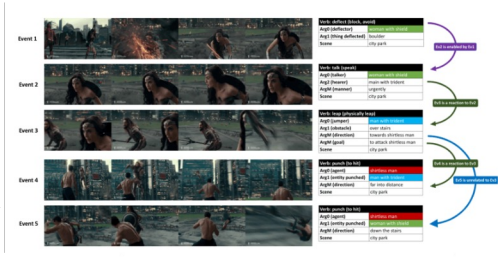
Brandeis

# Future work: Automation

End-to-end AMRs

-Action recognition models trained on VidSitu

-Direct conversion from PropBank SRL to AMR



AMR from captions

-Caption models trained on MSR-VTT type datasets

-Parse caption into AMRs

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

Note: models trained on the above datasets cannot capture simultaneous actions or multimodal links

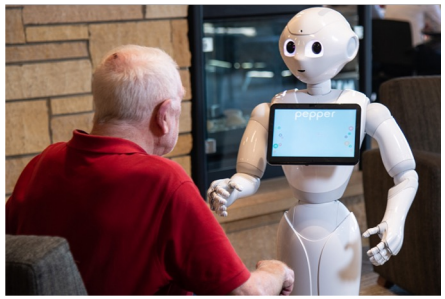
Brandeis

# Future work: Applications

What data lends itself to multimodal AMR and action annotation?

What fields could it potentially help?

*problem-based learning*



*home assistance*

*digital support*

Brandeis

# Acknowledgements

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805.

Brandeis