

Generating a Novel Dataset of Multimodal Referring Expressions

Nikhil Krishnaswamy and James Pustejovsky

nkrishna@brandeis.edu • jamesp@brandeis.edu

<http://github.com/VoxML/public-data/tree/master/EMRE>



Introduction

- In peer-to-peer communication, gesture can directly ground spatial information
- Language affords abstract strategies to distinguish similar objects
- As environmental complexity grows, so does the language used to single out specific items
 - object indicated by deixis is usually topic of discussion
 - deixis may be ambiguous based on, e.g., distance from agent to target object, other objects close to target, etc.
 - supplemental language can create more useful definite descriptions
- Speech/gesture “ensemble” may involve deixis to ground location, language to specify further
 - as task’s language requirements grow more complex, subjects rely on other modalities to carry semantic load
 - humans intelligently mix modalities in real time
- We present a novel dataset of Embodied Multimodal Referring Expressions (EMRE) – data generation, annotation, evaluation, preliminary analysis, and expected uses



Video and Quantitative Data

- Data gathered using VoxSim semantic event simulator, based on VoxML semantic modeling language
- Object reference may ground to gesture, language, or both, subject to constraints
 - Where do these constraints occur? Where do humans prefer one referring modality to another?
- 6 possible targets: **non-uniquely-colored blocks**
- 3 landmarks: **cup, knife, plate** (not used as targets)
- Captured videos show 3D avatar referring to each possible target object with gesture and/or English
- 50 object configurations x 6 targets x 5 referring strategies
- Gesture only (deixis), language only (x2), or ensemble (x2)
- Linguistic descriptions use **relative** or **absolute** distance
 - Relative: *This* is closer to me than *that* similar object
 - Absolute: *This* is in the closer half of the table to me; *that* is in the farther half
- ≤3 randomly-generated relational descriptors of target relative to other objects

```
point
TYPE = [
  HEAD = assignment
  ARGS = [
    A1 = x:agent
    A2 = y:finger
    A3 = z:location
    A4 = w:physobj•location
  ]
  BODY = [
    E1 = extend(x, y)
    E2 = def(vec(x → y × z), as(w))
  ]
]
```



Annotation

- Parameters stored: referring modality, distance distinction/type, descriptive phrase, relational descriptors, object coordinates, relation set, agent-target Euclidean distance

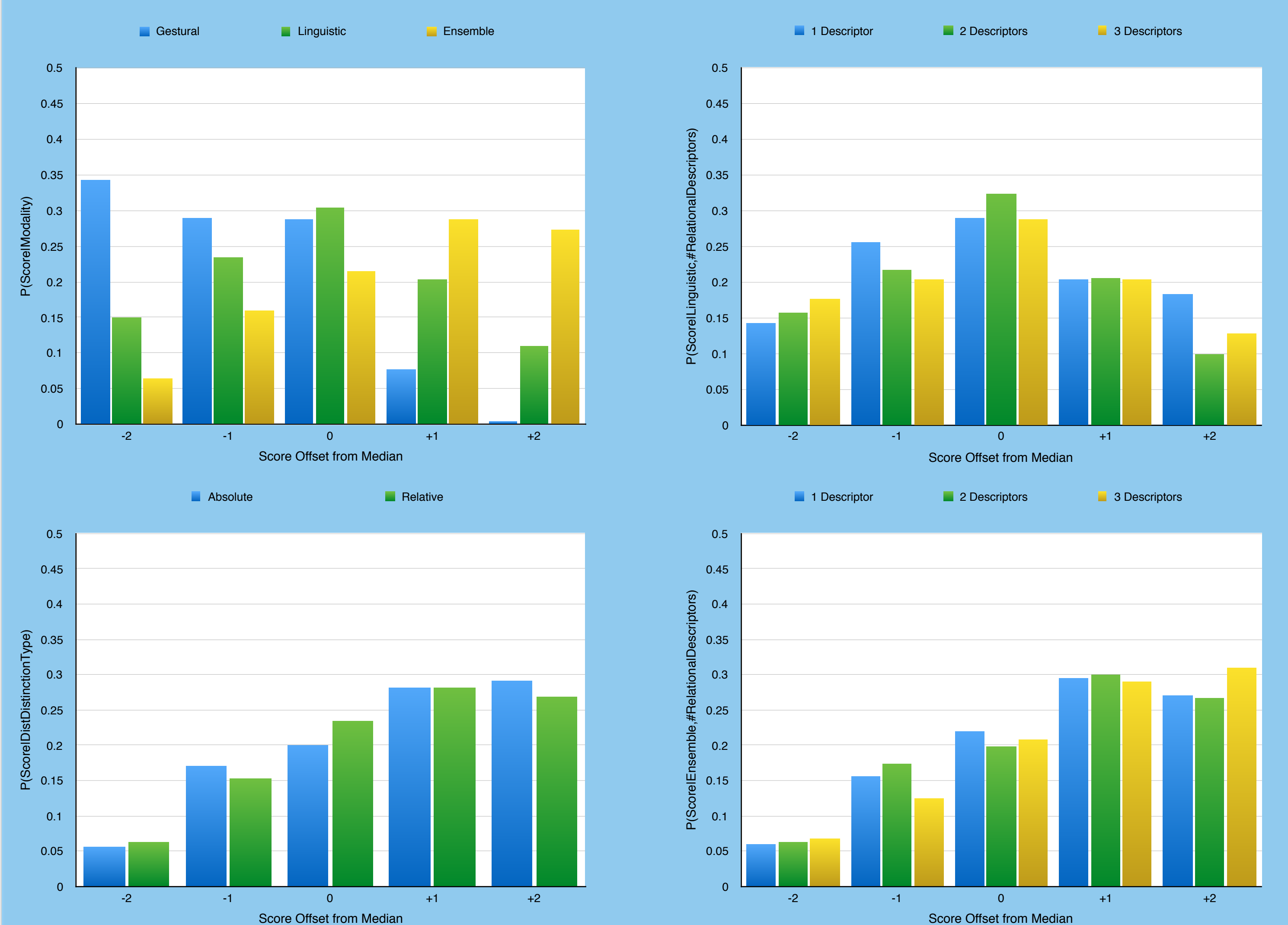
```
Id          ...
FilePath    ...
TargetObj   "block7"
ObjCoords   block6:<0.50482; 0.81595; 1.07951>
            purple_block3:<-0.01634; 0.81598; -0.15706>
            block7:<0.53786; 0.81595; 0.22346>
RefModality Ensemble
DescriptionStr "This purple block in front of the green block and in front of the red block."
RelationSet   touching cup square_table
            behind knife red_block1
            left red_block1 cup
ObjDistToAgent 1.13457977771759
DistanceDistinction true
DistDistinctionType Absolute
RelationalDescriptors in front of the green block
```

Sample database entry

- Videos grouped by configuration, posted to MTurk
- Likert-type ranking (1-5): **how natural is the reference method in the video? (≤3 ties allowed)**
- Fee: USD 0.10/HIT; Time: 30 minutes
- Workers optionally describe how they would refer to target object

- Result: 1,500 videos depicting referring methods for objects in various configurations with quantitative values, annotated by 8 workers each

Analysis



Discussion

- Analyzed probability distributions of high- and low-ranked referring expressions relative to conditions in video containing them
 - probability of score 1-5
 - probability of score compared to task’s median score (±2)
- Ensemble modality most natural, gesture-only insufficient, language-only sufficient but sub-optimal
 - more descriptors ~ better score
- Absolute distance distinction somewhat preferred to relative

Future Work

- Deploying a model:
 - Must capture strong predictors and more subtle dependencies
 - If dependencies from a particular configuration require choosing modality at runtime: CNN over relations in scene, weighted by information gain over descriptor
 - If avatar cannot use hands: need an intelligent model of linguistic-only reference
 - If prior actions construct context: sequential model over EMRE relation sets, ANN classifier over live configuration

Conclusions

- EMRE blends gesture and English text-to-speech
- Used by avatar in HCI scenario to generate REs that are *appropriate*, *salient*, and *natural* in context
- Strong human preference for “ensemble” modality
- Convincing case for computer to incorporate gestural output for fluent HCI
- We seek to build models for generating/recognizing/classifying referring expressions that are *natural* and *useful* to human users of computational dialogue systems