ISA-17 VoxML Track Annotation Guidelines

1 Data and Interface

The dataset consists of a set of approximately 100 images. These images are taken from the Visual Genome dataset (Krishna et al., 2017) and show objects in common configurations or activities.

The annotation interface runs locally in your web browser. Please follow the installation and setup instructions at:

`https://github.com/csu-signal/VoxML-Track-Annotation-2021`.

You will be asked to put in your email when you launch the tool. This is solely to mark which images task participants have already completed so you can leave and resume the task without being served the same images again (so please put in the same email each time!). Your email is being used **only** as a unique ID and will be used for no other purposes. We will remove the email IDs from the records when the annotation is complete and the dataset is compiled.

Hit "SUBMIT & GO TO NEXT" to proceed to the next image. At this point, your annotation of the current image will be saved as a record in Firebase associated to your email ID. We will download all recorded annotations from Firebase at the conclusion of the shared task period. All records are stored in a dedicated backend linked to a secure email address: `signallab [dot] ai [at] gmail [dot] com`. You may contact us at this email if you have questions.

2 Guideline

For each image, we ask annotators to answer the following questions using the provided annotation interface:

1. Provide a brief caption of the image.
2. What activity (if any) is the focus of the scene?
3. Identify the objects/entities in the scene (including people and animals). Use "ADD OBJECT" to add a new object/entity. If more than one object of the same type is in the image (e.g., multiple cups) you only have to annotate one.

4. For each entity, annotate activities associated with that object (e.g., that are shown being performed, or could be performed with that object). Use passive voice ("be *X*ed") where appropriate. Use "Add Activity" to add a new activity associated with the same object.

   - *(Recommended)* List up to 4 where applicable.
   - If an activity is *depicted in the image*, check the "Shown In Scene" box.
   - If an activity is *not* depicted in the image, annotate the circumstances that would need to be changed for this activity to be performed. Complete the prompt: *To <activity>, <object> must...* Use passive voice where appropriate.

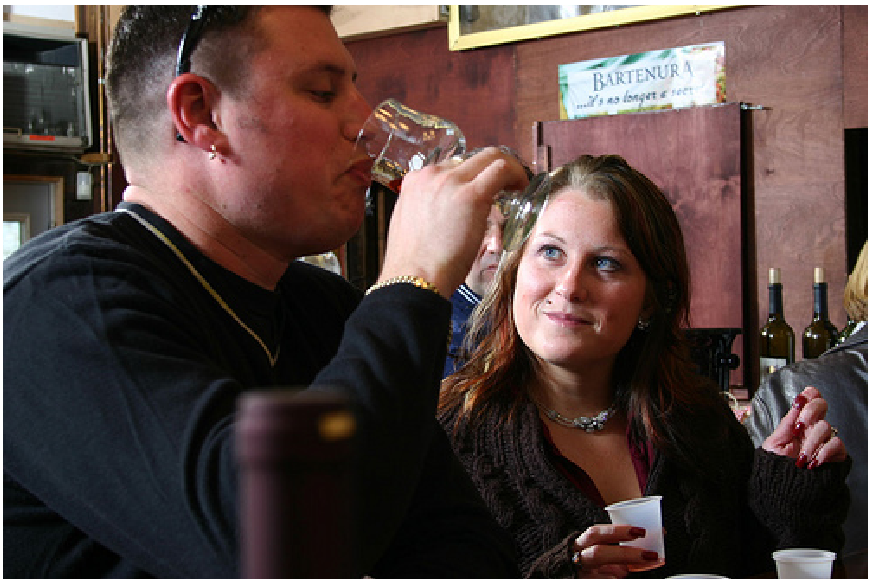
5. Identify the major spatial and configurational relations between the objects/entities in the scene. Below are some sample relations:

   - `left`, `right`, `in front`, `behind`, `in`, `on`, `above`, `below`, `over`, `under`, `near`, `beside`, `touching`, `not touching`, `holding`, `containing`, `supporting`

   Annotations need not be restricted to this relation set but we recommend adhering as closely to it as possible.

3 Example

Below is a sample image and possible annotations:



1. Provide a brief caption of the image.
   A man and woman drinking together at a restaurant

2. What activity (if any) is the focus of the scene?
   man drinking from glass

3. Identify the objects/entities in the scene (including people and animals).
   man, woman, glass, cup, bottle, sunglasses

4. For each entity, annotate activities associated with that object. If the activity is not shown in the scene, annotate the circumstances that would need to be changed for this activity to be performed.

   - glass:
     be drunk from (✔shown in scene), *be held* (✔shown in scene), *be set down* (To *be set down*, the *glass* must be put on a surface and released)
   - cup:
     be held (✔shown in scene), *be drunk from* (To *be drunk from*, the *cup* must be lifted to the woman's mouth)
   - bottle:
     be drunk from (To *be drunk from*, the *bottle* must be opened and lifted to someone's mouth), *roll* (To *roll*, the *bottle* must be placed horizontally on a surface)

5. Identify the major spatial and configurational relations between the objects/entities in the scene.
   man ***holding*** *glass*, *woman* ***holding*** *cup*, *woman* ***beside*** *man*, *bottle* ***in front of*** *man*, *bottles* ***behind*** *woman*, *cup* ***in front of*** *woman*, *man* ***touching*** *glass*, *glass* ***containing*** *liquid*, *man* ***wearing*** *sunglasses*, *bottles* ***on*** *surface*

4 About the Shared Task

This annotation task is part of an effort to develop the VoxML modeling language (Pustejovsky and Krishnaswamy, 2016) into an ISO Standard for the semantic representation of visual information. We ask participants in the shared task to annotate as many of the images as they can using the provided tool, and to submit a short project note (up to 4 pages) covering the adequacy of the annotation guideline, facility of the annotation tool, properties they observed in the raw image data while annotating, and optional deeper investigations into other things that would be interesting to annotate (other data, other properties we should have included, problems pertaining to standardization, etc.). These project notes will be reviewed for presentation (oral or poster) at the 17th Joint ACL-ISO Workshop on Interoperable Semantic Annotation, space permitting.

Annotations and project notes are due April 21, 2021. Thank you for participating!

References

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.-J., Shamma, D. A., et al. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Pustejovsky, J. and Krishnaswamy, N. (2016). Voxml: A visualization modeling language. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4606–4613.