A multimodal corpus for integrated language and action

Mary Swift*, George Ferguson*, Lucian Galescu†, Yi Chu*, Craig Harman*, Hyuckchul Jung†, Ian Perera*, Young Chol Song*, James Allen*†, Henry Kautz*

*Department of Computer Science, University of Rochester, Rochester, NY 14627
†Institute for Human and Machine Cognition, 40 South Alcaniz Street, Pensacola, FL 32502
{swift, ferguson, chu, charman, iperera, ysong, james, kautz@cs.rochester.edu}
{lgalescu, hjung}@ihmc.us

Abstract

We describe a corpus for research on learning everyday tasks in natural environments using the combination of natural language description and rich sensor data that we have collected for the CAET (Cognitive Assistant for Everyday Tasks) project. We have collected audio, video, Kinect RGB-Depth video and RFID object-touch data while participants demonstrate how to make a cup of tea. The raw data are augmented with gold-standard annotations for the language representation and the actions performed. We augment activity observation with natural language instruction to assist in task learning.

1. Introduction

Much progress has been made in individual areas of artificial intelligence, including natural language understanding, visual perception, and common-sense reasoning. But to advance intelligent assistants, systems that can interact naturally with humans to help them perform real-world tasks, an integrated approach is required. Our goal in the CAET (Cognitive Assistant for Everyday Tasks) project is to build a system that can automatically learn to track everyday activities from demonstration and use the models to help people perform everyday tasks. We have created an experimental environment for learning everyday tasks in natural environments using the combination of natural language description and rich sensor data. Natural language assists in task learning by providing a useful level of abstraction from the observed actions, as well as information about task parameters and hierarchical structure (Allen et al., 2007).

We focus on structured activities of daily living that lend themselves to practical experimentation in a kitchen domain. The initial corpus we describe here consists of recorded and annotated demonstrations of making tea. Subjects verbally describe what they are doing as they make a cup of tea, as they might in an instructional video. The audio and video are annotated with gold-standard representations of the language and activities performed.

In related work, the CMU Multi-Modal Activity Database (2009) is a corpus of recorded and annotated video, audio and motion capture data of subjects cooking recipes in a kitchen. However, the subjects did not verbally describe their actions. In addition, we use the Microsoft Kinect to capture 3D point-cloud data.

In the following sections we describe the corpus, the data collection and equipment used, and the gold-standard annotations for language and activity. We conclude with a discussion of future work.

2. Corpus Overview

The corpus consists of recorded and annotated sessions of people demonstrating how to make tea. The raw data comprise audio (both speech and ambient sound), video, Microsoft Kinect RGB-Depth video, RFID object-touch data and data from other environmental sensors. The raw data are augmented with gold-standard annotations for the language representation and the actions performed. Ten sessions were selected from a larger set for the gold-standard annotation. Sessions for this initial annotation subset were selected so that each of the seven participants was represented at least once. A secondary selection criterion was fluency of speech.

3. Recording Protocol

Subjects were familiarized with the experimental kitchen setup, including the electric kettle and the location of objects they might use, such as the tea, cups and cutlery. They were instructed to make tea and at the same time verbally describe what they were doing, as if teaching someone how to do it. They were asked to speak naturally. We recorded three sessions for each of seven participants (4 male, 3 female) from the research community, for a total of 21 tea-making sessions. We recorded additional sessions in the event of technical failure during recording, such as no audio from the lapel mike.

Figure 1: The experimental kitchen setup.

4. Equipment Setup

The kitchen environment is shown in Figure 1. The simple kitchen has an island workspace with a working sink. There

are cabinets and a refrigerator along the back wall. The video camera and Kinect were mounted on tripods to record the activity. Ceiling-mounted microphones record the ambient noise. Subjects wore a lapel microphone to capture speech, and an RFID sensing iBracelet on each wrist.

Audio Audio data was collected from three microphones: a lavalier microphone as an unobtrusive means of capturing speech, and two ceiling mounted microphones for recording of the environment. All equipment used was consumer-grade. The lavalier microphone was an Audio-Technica AT831cW connected to the Audio-Technica 701/L UHF Wireless System. The other two microphones were Behringer C2 condenser microphones. The interface to the computer used for data collection was via a Tascam US800 USB 2.0 audio interface. All audio was recorded at 44.1kHz in a 16-bit, linear PCM format.

Video Video was captured using a tripod-mounted Flip Mino HD camera and stored as 720p HD using H.264 (60 fps) and AAC codecs in an MPEG-4 container. The video is used to support the annotation and to present the results of recognition in context. We have not yet explored using it for computer vision.

Kinect RGB-Depth cameras allow for the easy but accurate collection of synchronized depth information in addition to the RGB image. In particular, the Microsoft Kinect has been the sensor of choice for many machine vision tasks over the past year as it provides a low cost and robust alternative to video-only cameras. The Kinect is particularly suitable for indoor activity data collection due to its high depth accuracy and framerate despite its constrained field of view (of around 60 degrees). Problems of human and object segmentation and localization that are difficult for ordinary video have the potential to be solved in a more straightforward manner using RGB-Depth data.
Video was collected using the two cameras on the Kinect, RGB and depth. Using the OpenNI drivers [1] for the Kinect, the two images were aligned so that each pixel in the depth image is aligned with the RGB image. The cameras were centered and placed approximately one meter away from the kitchen table with the entire kitchen lab inside the field of view. Each collected frame was indexed, timestamped and stored in two separate image streams both 640x480 pixels in resolution, at an average of 6.5 frames per second. Figure 2 shows both the RGB (left) and depth (right) streams for a given frame.

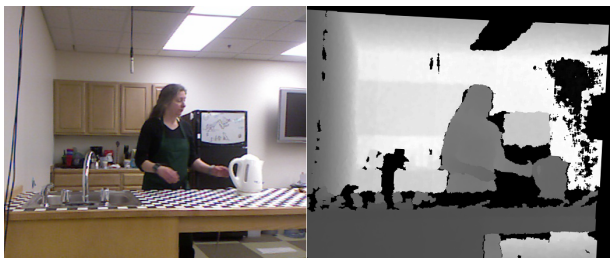

Figure 3: Kinect data frame showing RGB and depth streams.

Environmental Sensors RFID tags were attached to objects in the scene that were interacted with. The subject wore an RFID sensing iBracelet on each wrist, which records the RFID tag closest to the subject's wrist. Only one RFID tag was detected at any given time, and the current tag being detected was sampled every .2 seconds. The RFID detection was somewhat unreliable as nearby tags can interfere with each other, and often tags went undetected for periods of time. In an attempt to overcome these limitations, we attached multiple tags to each object to improve detection rates.
Other sensors were attached to kitchen appliances and cabinets to provide additional data as the subject interacted with the environment. The electronic tea kettle was attached to a Watts up PRO ES Electricity meter which recorded power consumption in real time (at the rate of about 1 Hz). Insteon contact sensors were placed on the kitchen cabinet doors and drawers as well as the refrigerator door to detect when they were opened and closed.

5. Annotation

We use the ANVIL annotation tool (Kipp, 2001) to create a multi-layered representation of the session data. Our data annotations consist of the speech transcript, a logical form for each utterance, an event description extracted from the logical form, gesture annotations for actions, objects, paths and locations, as well as object and agent tracking data extracted from the Kinect. Each layer is described below.

Speech The speech was transcribed based on the lapel microphone recording and the transcription was segmented into utterances. Breaking points were typically at the end of sentences. However, since the speech was spontaneous, we had many utterances that were not complete sentences (*e.g.*, missing predicates); in such cases, we considered long pauses to mark utterance boundaries. There were a few instances when several sentences were uttered in a continuous sequence, with very short or no discernible pause between them; for these we considered the whole segment to be a single utterance, rather than breaking it up into sentences.

Language The speech transcriptions were parsed with the TRIPS parser (Allen et al., 2008) to create a deep semantic representation of the language, the logical form (LF). The parser uses a semantic lexicon and ontology to create an LF that includes thematic roles, semantic types and semantic features, yielding richer representations than "sequence of words" models. Figure 3 shows a graphical representation of the logical form for the utterance *Today I'm going to make a cup of tea.* In triples of the form (:* CREATE MAKE), the second term is the semantic type in the ontology for the word, which is the third term. Nodes are connected by edges indicating dependencies, with labels for the relation, such as the thematic role of the argument. Tense, aspect and modality information is used as input for temporal reasoning, described below as future work.
The parser-generated LFs were hand-corrected for the gold standard annotation. This was particularly necessary for utterances where multiple sentences were transcribed as a single utterance.
To facilitate using language representation as features in activity recognition models, we added new semantic types in the ontology to correspond to objects and actions in the domain, such as *tea, cup, steep*. The new labels were usually specific subtypes of already existing semantic types. For example, the word *tea*

[1] http://www.openni.org/

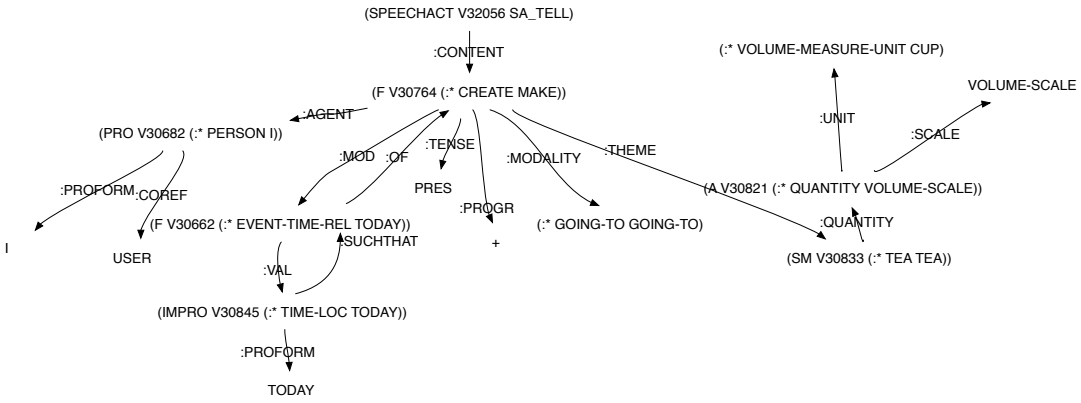

Figure 2: Logical form for *Today I'm going to make a cup of tea*.

was in the ontology under the general type TEAS-COCKTAILS-BLENDS, so we created the specific subtype TEA. This extension gives us greater transparency in the surface representation, but we retain the richness of the hierarchical structure and semantic features of our language ontology.

The LFs are input to the TRIPS Interpretation Manager (IM), which computes crucial information for reasoning in the domain, including reference resolution. The IM extracts a concise event description from each clause, derived from each main verb and its arguments. The event descriptions are formulated in terms of the more abstract semantic types in the LF, resulting in short phrases such as CREATE TEA, CLOSE LID, and POUR WATER INTO CUP. These phrases will be used as language features in our activity recognition models. Figure 4 shows an example of an extraction from the LF for *Place tea bag in the cup*. The objects *bag* and *cup* are identified as REFERENTIAL by the IM and it also includes the coreferential index for the first mention of the term.

```
(EXTRACTION-RESULT
 :VALUE ((EVENT V38801)
  (PUT V38801) (:THEME V38801 V38818)
  (:SHORT-DESCRIPTION V38801 (PUT (:* BAG BAG) INTO CUP))
  (:INTO V38801 V38887)
  (:TEMPORAL-RELATION V38801 >NOW) (:TENSE V38801 PRES)
  (REFERENTIAL V38818) (BAG V38818)
  (:ASSOC-WITH V38818 V38814) (:COREF V38818 V38185)
  (REFERENTIAL V38887) (CUP V38887)
  (:COREF V38887 V33594)
  (NONREFERENTIAL V38814) (TEA V38814))
  :WORDS (PUT THE TEA BAG INTO THE CUP))
```

Figure 4: Extraction for the utterance *Place tea bag in the cup*.

Activity For activity annotation, ground truth was created manually by observing recorded videos and annotating actions performed by test subjects. Domain actions, their attributes and admissible attribute values were pre-defined and stored in the ANVIL specification, allowing annotators to easily select them with the GUI.

The duration of each activity was annotated with start and end time relative to the video with centisecond accuracy. Each video was observed several times and some segments were observed more than ten times to produce accurate annotation. On average, it took about 25 minutes to annotate a video. Nevertheless, for actions not clearly visible in the video, the timing information can have a certain degree of error (mostly less than a second). Simultaneous actions (*e.g.*, opening a kettle while moving it) were also annotated with overlapping time duration.

Actions (*move, put, fill, ...*) and their attributes, such as objects (*cup, kettle, spoon ...*) and paths (*to, from, into*) are annotated as separate tracks so that they can be accessed programmatically for a compositional analysis. The argument structure of the activity annotation follows linguistic structure as closely as possible. The actions take a theme, possibly followed by a relation with its own associated entity (*e.g.*, ACTION:*pour* THEME:*the water* RELATION:*into* ASSOCIATED-ENTITY:*the cup*). We annotated the composite actions as a separate track for ease of viewing the overall activity annotation.

Figure 5 shows an example of the language and activity annotation for the action of putting the tea bag into the cup. Here the extraction (EX) tier for the language annotation is highlighted. The concise activity description (PUT (:* BAG BAG) INTO CUP) represents the full extraction information, which appears the attributes window (see also Figure 4). Kinect data on object tracking is also shown.

Kinect, RFID and other sensors We have integrated annotation tiers for information extracted from the Kinect data showing agent location, objects, and objects held. The RFID sensor data logs included the ID of the active tag and a timestamp when the iBracelet sensed that tag. We also have an indication of which of the two iBracelet devices sensed that particular tag, but we ignored this information in our annotation. During the handling of an RFID-tagged object, there would be multiple readings of the attached tag(s); typically 3-4 sensing events per second. We annotated intervals of use of an object, between the first and the last continuous readings for that object; pauses of at least 1s were considered positive evidence that the object was not longer held/touched.

The sessions included in the corpus only included contact sensor data for one kitchen cabinet (where the tea and the cups were stored), and the refrigerator (where there was milk) For these, we marked up the intervals of use. Contact sensor data is recorded when an opening or closing event happens; typically such an event will trigger the sensor a few times, depending on how fast the opening/closing was. We took the first record of an opening action as indicating the start of "using" the cabinet/appliance; conversely, the last record of a closing action was taken to indicate the end of "using" the cabinet/appliance.

The recorded power sensor data provides wattage information, at the rate of about 1 Hz, for the electric kettle. We did not use the actual values of the power. Rather, we simply marked up the intervals with positive power consumption. In our case the electric kettle was the only appliance connected to the sensor, which makes the power-on intervals excellent indicators of when the kettle was turned on and off. The time it took for the kettle to boil varied (it depends on how much water was used) from 27s to 67s, with an average of 46.3s.

6. Discussion and Future Work

We have described our gold-standard corpus for research on integrated learning by demonstration from natural language, activity recognition and reasoning. The ten annotated sessions of our corpus comprise 184 transcribed utterances with logical form and

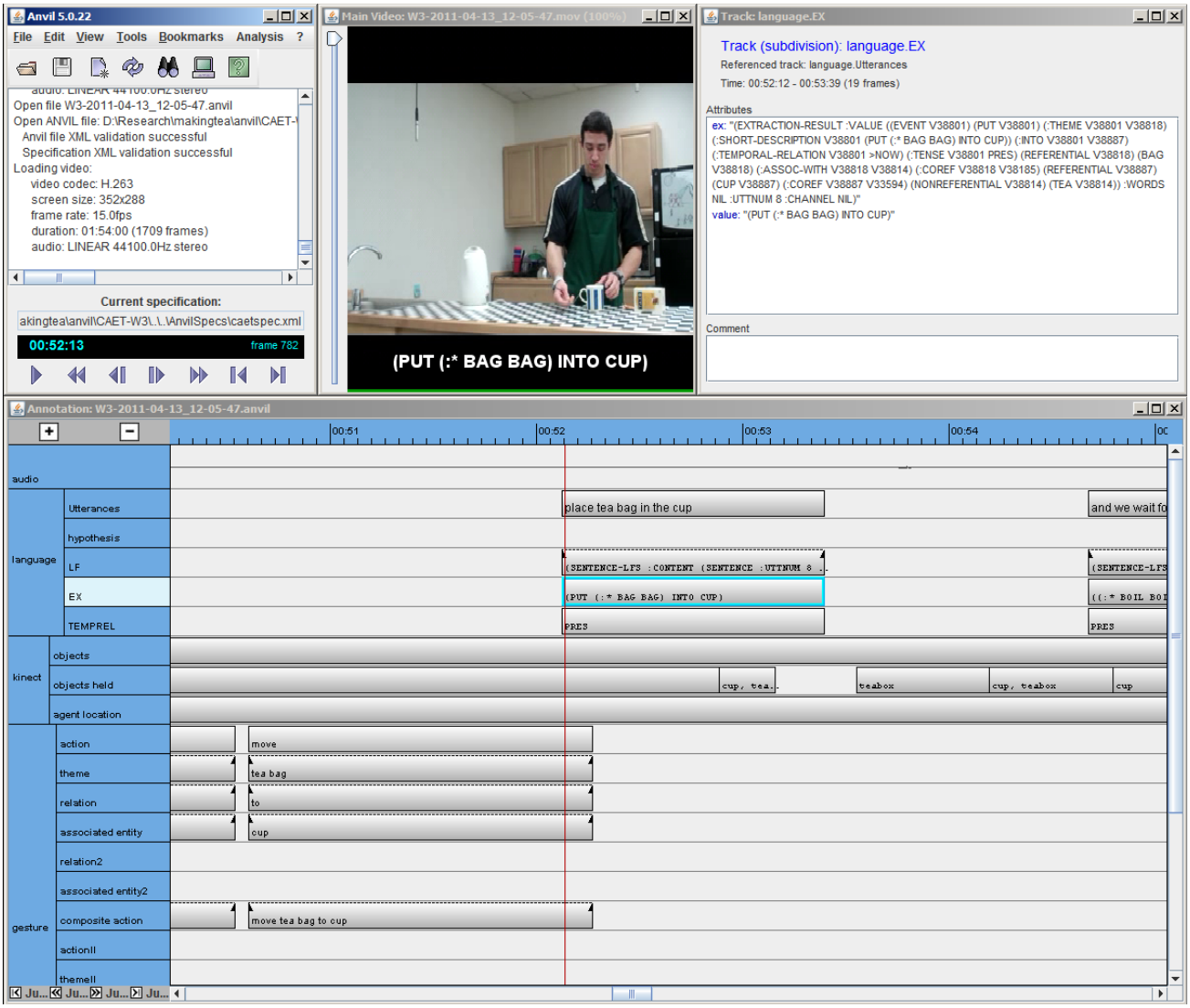

Figure 5: Language and activity annotation for *Place tea bag in the cup*.

extraction annotations, 345 annotated activities and 1.2GB (approximately 7200 frames) of Kinect data.

The use of RGB with depth data makes it possible to more easily segment and recognize human and object interactions. From the dataset we extract generally useful and "meaningful" features, such as semantic visual features, (*e.g.*,"left hand is directly above cup"). Using features at this level of abstraction, we plan on experimenting with more complex models of recognition and learning.

As noted above, we do not have the ground-truth data for all of the tag-object assignments, so we use an algorithm to assign the most probable object name for each RFID tag detected in the scene using the subject's description of the task. To learn the name of a given ID, we gather the nouns mentioned by the subject while each object is being interacted with, convert the nouns to ontological concepts using the gold-standard parse data, and determine the concept with the highest probability of being mentioned when that ID is detected. While we only have a small amount of data, the labels generated by this algorithm agreed with a human annotator, who used the video to determine the mappings, for six out of the eight tags. In the future, we hope to extend this algorithm to the Kinect data, making use of language to provide names for detected objects without the need for hand annotation.

Other research with our corpus involves computing temporal relations between the natural language descriptions and the actions performed using Markov Logic Networks and information in the event extractions. We are adding hand-annotated gold-standard temporal relations for each event extraction for evaluation.

7. Acknowledgements

This work was supported in part by NSF grants IIS-1012205 and IIS-1012017, Activity Recognition and Learning for a Cognitive Assistant. We are grateful for comments from 3 anonymous reviewers.

8. References

J. Allen, N. Chambers, G. Ferguson, L. Galescu, H. Jung, M. Swift, and W. Taysom. 2007. PLOW: A collaborative task learning agent. In *Proc. AAAI*.

J. Allen, M. Swift, and W. de Beaumont. 2008. Deep semantic analysis of text. In *Proc. Semantics in Text Processing*, STEP '08, Venice, Italy.

Michael Kipp. 2001. ANVIL - a generic annotation tool for multimodal dialogue.

F. De la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. 2009. Guide to the Carnegie Mellon University Multimodal Activity (CMU-MMAC) database. Technical Report CMU-RI-TR-08-22.