

# Remarks on the CommAI-env

Gemma Boleda

October 9, 2016

Both the *Roadmap* and the CommAI-env are a welcome move to a broader and more ambitious vision of AI, with both a global vision and a concrete proposal on how to move towards it. However, I find a bit of a gap between the description in the *Roadmap* and its current implementation in the set of tasks that are currently available in the CommAI-env.

First, the emphasis in the *Roadmap* is in **learning through interaction using natural language**. However, the basic tasks described focus a lot on Computer Science related constructs instead: Repetition of sequences of bits, use of separators, and lots of counting, for instance.<sup>1</sup> To learn the basics of linguistic interaction (and thus enable all further learning), what the Learner should master early on is a vocabulary, a (simple) grammar, and (simple) dialogue skills, such as answering questions (both yes/no and wh-questions) or asking questions (including asking for clarification). For the latter, simple versions of tasks such as T2 and M13 seem like the way to go. For the former, task G15 is obviously necessary, but the remainder of the tasks seem far from sufficient.

Second, and relatedly, some of the tasks (G0, many of the Navigation and Proprioception tasks) seem to me to rely on a misconception about meaning. For instance, task G0 asks the Learner to do something, but the Learner will not really learn to perform anything because its action has no consequence in the real world. As it is defined, I do not see the difference between this task and the pure repetition tasks. A similar comment applies to T3: With no grounding in the environment, the Learner will not learn about object existence, but simply pattern matching between “no” vs. “a”.

My concern can be even more clearly exemplified with the Navigation and Proprioception tasks. Consider the following examples from some of the task descriptions:

- (1) **Teacher:** Look to the east.  
**Learner:** I look to the east.
- (2) **Teacher:** Turn left/right.  
**Learner:** I turn left/right.  
**Teacher:** You turned [+1].
- (3) **Teacher:** Move forward.  
**Learner:** I move forward.  
**Teacher:** You moved [+1].

These tasks are identical to the repetition tasks (modulo the substitution of “you” for “I”). If the purpose is just to learn the word strings (“east”, etc.), then it’s ok. If the purpose is to learn their meaning, such that they can be used in subsequent tasks, then they should be based on some consequence for the Learner (such as the effect of moving; seeing something, etc.). That is, the

---

<sup>1</sup>Children learn to count much later than they learn to talk.

process of learning meaning should be grounded in the environment. This aspect is included in other navigation tasks, but embedded in some more complex learning goal, such that more basic tasks that allow the Learner to grasp the meaning of actions are called for. Having the environment in place, this should be doable.

This brings me to my third concern. The *Roadmap* places a lot of emphasis on the incrementality of learning and the use of what is learned in some tasks for solving other, possibly more complex, tasks. However, clearly this is something that needs considerably more work. Two examples: (1) It is not clear to me what the relationship between K0 and the other tasks is, or what the use of learning to be silent is for the remainder of the tasks as they are currently defined (more generally, some of the tasks seem like the result of a “oh, here is a neat task that we can easily script and looks intelligently” thought process, rather than from a plan to build skills incrementally); (2) the tasks related to objects and their properties start from a very complex, 3-way *baskets x objects x properties* setup; wouldn't it make sense to start from a 2-way *objects x properties* setup, and then move to the 3-way scenario? These two examples are easy to fix, and thus my comment may seem anecdotal, but I think that it points to a huge methodological issue that is easy to underestimate, namely, how to define a set of tasks that build on one another *and* lead to an intelligent machine that can interact with humans. It reminds me of the problems encountered in symbolic approaches, when researchers started from the *bachelor = +human +male -married* case, or from taxonomies with ISA relationships, and thought it self-evident that the approach would generalize to a whole lexicon.

Finally, I have another general issue with the restriction to a particular maximum time. I totally share the underlying motivation for this (see *Roadmap* ), but, again, the implementation is very difficult, because there is no clear methodology to determine how much time a Learner should master a given task in. For instance, it takes children months to be able to utter syllables, and one year to utter their first word. What is the equivalent in machine time, and how do we know? As an exercise, I encourage the CommAI-env team to separately decide the maximum time for each task, and compare the results.