February 2009

Human XP: Using Virtual Worlds to Capture Common Sense

Jerry Weltman
Louisiana State University, jweltm2@tigers.lsu.edu

Human XP: Using Virtual Worlds to Capture Common Sense

Jerry Weltman
Louisiana State University
jweltm2@tigers.lsu.edu

ABSTRACT

A long-standing goal of Artificial Intelligence is to program computers with common sense. Virtual worlds could be especially valuable in this endeavor because they represent people and objects in 3-D space interacting with a world similar in many ways to our own. I propose the Human Experience Project, a new type of annotated corpus for collecting common sense, based on scenes in a virtual world. The proposal calls for volunteer contributors over the internet to create 3-D scenes about everyday life. The contributors' annotations will describe the objects in the scenes, the actions, and the hidden mental states of the actors. The goal is to amass a collection of scenes and annotations that researchers can use to model common sense. I present a sample scene and outline the broad requirements of this ambitious project.

Keywords

knowledge acquisition, common sense, virtual worlds, large-scale collaboration.

INTRODUCTION

Providing computers with common sense is a vast and intractable problem. Even after fifty years of Artificial Intelligence research, we still do not know how much effort it will take to teach computers simple facts that every child knows, such as sand tastes bad, it is fun to receive a gift, and it is sad when a pet dies. No one knows how many commonsense facts exist or which ones are the most important. Although there are ongoing projects to collect commonsense data (Lieberman, Smith, and Teeters, 2007; Lenat and Guha, 2000; Speer, 2007), none of these makes use of the rich representations of reality possible in virtual worlds.

I am proposing the Human Experience Project. The idea is to build a tool, based on virtual worlds, that allows thousands of contributors on the internet to create scenes of everyday life. The goal is to collect a large corpus of annotated 3-D scenes, which I call the Human XP corpus. This corpus would provide researchers with valuable raw data for designing neural networks, semantic networks, statistically-based rules, or other structures for AI and natural language processing.

VIRTUAL WORLDS AND COMMON SENSE

Virtual worlds can teach a computer a lot about common sense. If the data in virtual worlds could be mined, each virtual building, tree, and person could provide a digital mapping from the virtual world to items in the real world, telling a computer program where an item is commonly found (a tree is typically outside), its relative size (trees are larger than people), and how it moves (trees sway in the wind and fall down during hurricanes). Each built-in avatar gesture, movement, and facial expression could provide commonsense data about what a body can do. Each complex component, made up of groups of simpler component could provide data about the composition of objects. Scenes with 3-D animation could provide data about actions, sub actions, and concurrent events.

In short, many aspects of common sense that are so difficult to express in text-based structures arise naturally in a virtual world setting. What is missing are the annotations that would allow an AI tool to analyze the virtual world and understand the meaning behind the actions of a scene. Understanding a scene means to be able to answer basic questions. What objects are in the scene? Where are they located? Who are the actors? What are they doing? Why are they doing it? A large challenge of Human XP is to provide an integrated environment that makes it easy for amateurs to answer these questions through various types of annotations.

Virtual worlds also offer a different type of reward over game-playing interfaces currently being used to extract commonsense data (see the games listed in Lieberman, Smith, and Teeters, 2007). Intellectually labor-intensive projects such as Wikipedia show that thousands of people are willing and ready to share their knowledge for the common good of

humanity. And the fabulously detailed exhibits in Second Life show how much effort people are willing to expend to satisfy their need to create and share their ideas. The enormous task of collecting human experiences needs this pool of talent, and the Human XP project needs to attract them through a multi-modal collaborative environment where they not only view each other's work but also have fun working together.

WHO'S LIFE IS IT, ANYWAY?

What does it mean to say that an experience is part of everyday life? The life of a farmer in Costa Rica will be vastly different from that of a stockbroker in Kuala Lumpur. It would be preposterous to presume that there is a common life experience for all humans. Focusing on life experiences from a single cultural point of view could alienate contributors. There is no easy answer to the question about whose life to model except to say that all human experiences are equally valid for gathering common sense, and that Human XP needs all of them.

On the other hand, for simplicity of representation, I believe it is important to start with the experiences of children in order to limit the amount of detail that scenes represent. I choose a six-year-old child because many children are able to read simple narratives by that age, and understanding simple narratives is an important application of commonsense modeling.

I will choose scenes from the fictional life of a boy named Max who lives a middle-class life in the USA. Clearly, equally valuable data can be obtained from life experiences of a child in other cultures and socio-economic strata. Indeed, a fundamental requirement of Human XP is to make it possible for participants from many points of views to contribute. Yet a middle-class American boy will resonate with the commonsense intuitions of many computer users.

The next section shows an example of a simple annotated scene. Then I discuss the broad requirements for this very ambitious project.

A SIMPLE ANNOTATED SCENE

The following scene is called "Max breaks a vase." The scene begins with Max bored with nothing to do, sitting in a living room. He notices a lovely vase, walks over to it, picks it up, and drops it, smashing it into several pieces. It is great fun! The scene shows Max's mental state before breaking the vase and what happens as a result of his action.

A scene consists of one or more frames. Each frame has two areas. On the right side is a 3-D depiction of a time slice of Max's life. On the left side is the "maxometer," a representation of feelings of positive, negative, and neutral sensations experienced by any part of Max's body, including Max's brain. The maxometer uses three colors to categorize feelings: negative is shown by the color gray, positive is pink, Teaching machines about everyday life and neutral has no color. The relative intensity of each feeling is represented by a corresponding slider bar. All of the frame data, including the actions on the right and the maxometer data on the left are completely specified by a Human XP volunteer contributor.

When the scene begins, shown in Figure 1, Max is feeling somewhat bored. Boredom is clearly an unpleasant feeling, so the brain is colored gray.

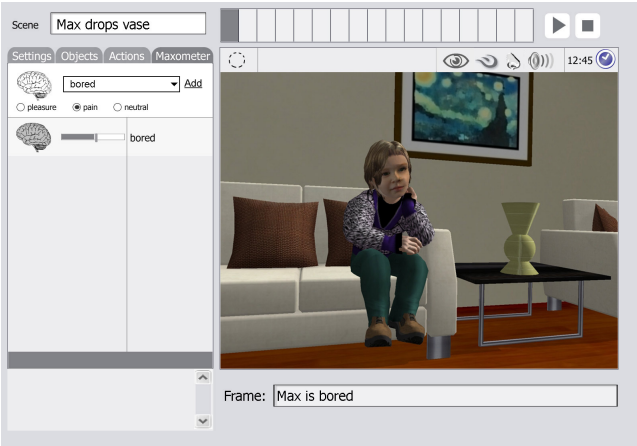

Figure 1: Max is Bored

In Figure 2, Max notices the vase and feels curious. Some of us might judge curiosity to be a negative feeling, and some of us might judge it to be a positive feeling, if it is in small doses. For now, let's judge it to be somewhat pleasant, so the brain is colored pink. The slider bar shows the level of curiosity is slightly past the "medium" level. The focus of Max's vision is represented with a circle on the vase. When the focus is present, it indicates that Max is attending to the image of the vase.

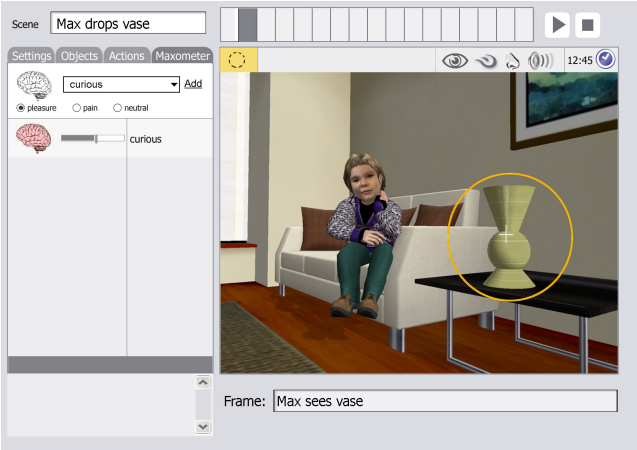

Figure 2: Max Sees the Vase

In Figure 3, Max grasps the vase. The Maxometer shows that his hands feel the pleasant sensation of the expensive ceramic that he is about the smash. His mental state has moved from curiosity to excitement

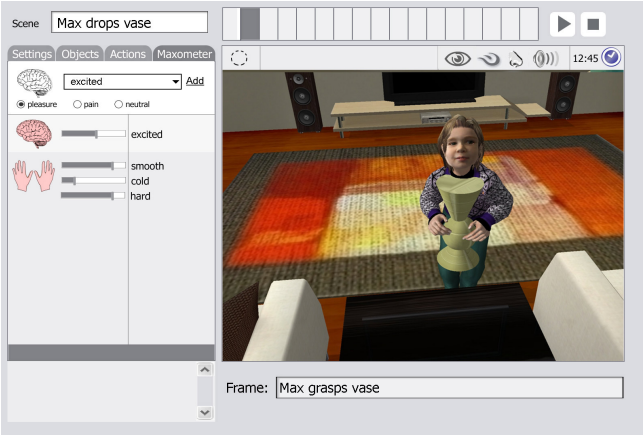

Figure 3: Max Grasps the Vase

Over the next few frames (not shown), Max lifts up the vase and lets go, causing the vase to fall to the floor. Finally, in Figure 4, the vase breaks apart. The sound of the vase crashing is depicted as a little megaphone centered on the site of the crash.

The above scene is valuable for commonsense modeling because it shows a typical environment where a boy might break a vase, how long the action takes, why someone might smash a vase, and what happens as a result of the action. Also implicit in each scene are the relative sizes of objects such as rooms, doors, furniture, etc, typical positions of those objects, and how humans interact with them. All of this information is part of the human experience. The 3-D presentation is a convenient way to represent this experience because it can simultaneously depict the interaction of many types of information. It explicitly states relationships between actions, the world, the body, and the mind. One can imagine that as the number scenes grows, so too grows the amount of information we have about what it means to have a body, experience sensations, and cope with mental states such as desire and fear.

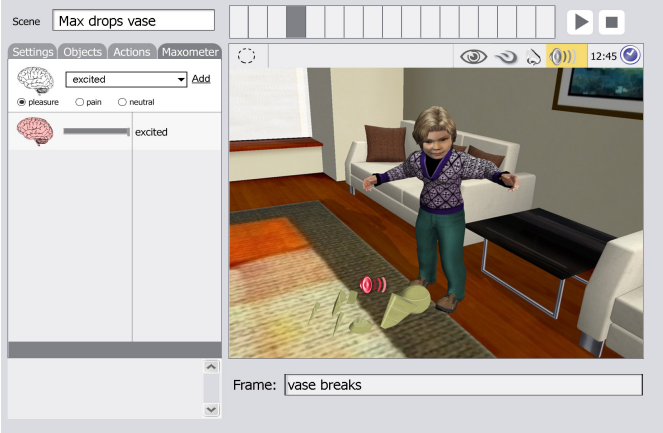

Figure 4: The Vase Breaks

BROAD REQUIREMENTS FOR THE PROJECT

Creating a scene will require a software tool which I call Scene Builder. In many respects, Scene Builder will be similar to current 3-D animation software and 3-D modeling packages that allow users to put objects into a scene to create animated movies. However unlike movies, Human XP scenes allow the audience to manipulate the viewing angle. Also, the annotations (which consist of scene labels, object labels, and maxometer readings) are an integral part of the scene.

Viewpoint

In Human XP there is no single audience viewpoint. The action and objects exist in the virtual world and a viewer can enter this world from any angle to view it. Thus, typical movie techniques such as the use of camera angles, close-ups, wide pans, etc. do not determine what the audience sees. While the events in the scene are fixed, the audience can rotate or zoom to any part of the world. In fact, for collaboration purposes, a group of avatars may enter a scene to watch it and discuss it while it unfolds. This type of live interaction requires a virtual world environment where the objects in a scene and the "actor" avatars are treated separately from live collaborator avatars.

Scene Annotations

As participants create a scene, they are expected to label it, along with its sub scenes, down to each individual frame. For example, the previous section had scene labels such as "Max grasps the vase" and "The vase breaks". Labels are critical to Human XP because they show the relationship between language and action, so Scene Builder must provide an easy labeling interface.

A label is most valuable when its meaning is clear: no double meanings, ambiguous phrasing, or unresolved pronoun references. Thus, ideally, the labels should be parsed into a logical form such as first-order predicate logic. Unfortunately, there exists no natural language parser that can reliably extract the meaning of natural language discourse. In fact, the goal of this project is to provide research data for this very purpose! Fortunately, the nature of this project assumes that participants are willing to work with the system to provide extra information, and thus a template-based approach, as described by (Singh et al., 2002) will be appropriate. Examples of templates are:

1) ?N1 is ?ADJ

Bob is hungry

2) ?N1 ?V [a] ?N2

Bob eats a sandwich

Object Labels and Parts

An object in Human XP is much more than a simulated 3-D surface. First, a label is mandatory. So a simple straight-backed chair object would be labeled as such. But more importantly, the chair object should be composed of sub objects: four legs, a seat, and back, each labeled appropriately. The decomposition of objects is part of common sense. That is, most six-year-olds

know that a chair has legs, and that there is a part you sit on a part that you can lean against, and thus this knowledge is part of the child's understanding of a scene involving a chair.

The material of the components should be labeled as wood, and wood itself should be labeled with some properties such as hard, opaque, and smooth. Furthermore, the movement of the folding chair should be part of its definition. That is, there should be a "fold" and "unfold" motion associated with the chair with animated frames of the chair seat moving up and down. When a participant puts the folding chair into a scene, one of the movements available for animation will be folding the chair.

How much detail should participants represent? My general rule of thumb is to label parts that a typical first-grader might know. A child would know that a chair is made of metal or wood and that it slides open. On the other hand, the hinges that allow the sliding action are probably not part of the child's awareness and need not be represented.

In some respects, Human XP could be viewed as a child's folk ontology of world objects. However unlike a formal ontology, Human XP does not attempt to categorize and organize objects into cohesive structures. The information about objects is based on the experiences that come from scenes. Human XP should be able to represent some generalizations that a child already knows such as a folding chair is a type of chair and a chair is a type of furniture. But most of these generalizations should arise because of situations in scenes. More research is needed to define how general relationships are formed, and how they are represented in Human XP.

Creating a warehouse of Human XP objects is no small task and will take an army of dedicated participants. While there already exist thousands of 3-D models of objects, free and non-free, only freely available models may be part of Human XP. Scene Builder must make it possible to import these models and add the tags necessary to make them available in a Human XP scene.

More than any other task, creating the "sets" for scenes will require large-scale cooperation. Human XP should provide a sort of virtual workshop where avatars of participants can share components and ideas as they create items and decide attributes and movements.

As a final note on objects, there is a significant software architecture issue when a virtual world must work with complex objects rather than surfaces. Think of a brick wall that is not simply a textured surface but rather a collection of thousands of individual bricks. The Human XP virtual world server must be able to work with objects at various levels of detail so that it renders the wall as single object but also is able to decompose it when necessary (e.g. Max cracks one of the bricks with a hammer).

The Maxometer

Each frame has a maxometer panel that shows the important perceptions and mental states of Max (or any actor). As a frame is created, the participant will imagine what the actors must be feeling and annotate the maxometer. However, only the most salient feelings should be annotated, that is, the ones at Max's attention. Otherwise, the participants would be overwhelmed in trying to annotate every sensation.

Representing and reasoning about mental states is an important aspect of AI (Bacon, 1995). For example, the BORIS system (Dyer, 1983) created a small number of mental states which could be combined to form more states. But Human XP does not seek to create a model of mental states or reason about their structure. For example, the perception of HARD and SOFT could be combined so that SOFT simply means a low value of HARD. But the goal of Human XP is not efficiency but rather expressiveness. When possible, we want to allow people to use the words that come naturally and not force alternative expressions. We want to *collect* information, not to order and categorize.

The maxometer falls under the field of folk psychology. The Belief, Desire, Intention (BDI) agent framework also draws upon folk psychology, but can still be rather complex (for example, see the LORA model in Wooldridge, 2000). Human XP should provide an intuitive visual representation of mental states that is appealing to lay people.

QUESTIONS ABOUT HUMAN XP

The simple example of Max breaking the vase leaves open many questions about scene creation and Human XP in general. They are discussed in this section.

Who owns and manages the intellectual property of Human XP?

Human XP should follow the Wikipedia model of ownership: All scenes, objects, and annotations are available under the GNU Free Documentation License. Data from the Human Experience Project must be freely available.

Also as per Wikipedia, there is no central management. The scenes and objects are freely editable through a wiki environment with a virtual world component to help ease collaboration.

How does a participant decide what scene to work on?

In the beginning of the project, there will most likely be a general theme of scenes so that the same actors and sets can be re-used. We can capture a lot of valuable commonsense data with scenes with Max and his Mommy at home, eating, playing inside and out, perhaps going to the park, etc.

What if a participant puts in inappropriate information?

A participant could put in inappropriate information such as bad scene labels, unreasonable maxometer readings, or unrealistic actions. A participant could easily have chosen a name like "Max jumps off a building" when Max was bored or even "blah blah blah." Similarly, the participant could show Max's feet as feeling warm and fuzzy even though they were not touching anything. However, just as with Wikipedia, when other people see this scene, they will be able to go through the frames and see all of the information. Over time, inappropriate frames will be found and corrected.

Max's feet touch the floor, so why do they not appear on the maxometer?

The maxometer shows only feelings that command Max's attention and thus explain the scene. We do not lose any data about the scene by making this simplification as long as Scene Builder can be designed with general principles about the human body. For example,

- Max usually breathes regularly.
- Max's heart beat's regularly.
- If a part a Max's body comes into contact with something, the body part will feel the exterior properties of the object, but Max may not attend to these feelings, etc.

There could be a number of feelings at the sub-attention level that Scene Builder keeps track of but will not be represented unless they become salient to the scene. If Max tries to run five miles, a participant should hopefully register Max's breathing on the maxometer to show that he is in pain, and thus it explains why he stops running; if he jumps on the couch, and delights in the bounciness, he should register the bouncy feeling and the excitement it causes; if it has been a long time since Max went to the bathroom, his bladder should register as feeling pressure so we know why Max runs to the bathroom. A general rule of thumb is to show perceptions and mental states that affect Max's actions so that the data in Human XP can teach a computer *why* a person does something.

How would you depict Max feeling happy about a possible future action?

So far, we have only seen mental states caused by the action in the current scene. Scene Builder should also allow mental states that refer to scenes of the past, counterfactual scenes, future unrealized events, and imaginary events. Remembering and imagining, along with anticipating, considering, realizing, choosing, deciding not to do something, etc, are mental processes that require this type of reference.

How do you depict other people?

Scene Builder should allow multiple actors in a scene, similar to other 3-D animation software. Furthermore, when more actors are added, it should be possible to represent each actor's point of view. That is, each actor should have their own maxometer. Multiple points of views helps explains the motivations behind each actor as they interact for a common purpose or perhaps at cross purposes.

How do you depict people talking?

Scene Builder will use dialog bubbles, similar to comic strips. Participants will be able to attach prosodic and paralinguistic features to these bubbles such as tone of voice, volume, and speech rate. Note the maxometer should gather valuable data about speech acts because it can represent the true motivations and feelings behind a person's utterance.

How will the system allow lay people to create scene annotations at a useful level of detail?

To reduce the need for formal training, I intend to build the framework so that it prompts contributors to explain the actions of the scene. (I am still researching how to do this.) Also, I believe that the wiki format, which encourages all participants to read and modify each other's work, will result in a common understanding of a useful level of annotation.

RELATED WORK

This proposal is inspired in part by work at the MIT Media Lab to collect commonsense data (Singh et al., 2002; Singh, Barry, and Liu, 2004). Stories about real life can provide valuable commonsense data. However, the challenge is to get ordinary people to write text that can be programmatically analyzed. One of the projects, ComicKit (Williams et al, 2005), offers an interactive comic strip interface for a contributor to construct a story graphically from a set of fixed actors, props, actions, settings and captions. These fixed elements make it more likely that the contributor will create a sequence of actions from the given elements, resulting in a story that can be easily analyzed.

Human XP uses similar elements to tell a story, but the elements contain more information, and thus more commonsense data. For example, ComicKit has flat settings and simple, abstract objects; Human XP is in 3-D with a realism that contains a wealth of commonsense information. ComicKit has a list of word actions like "walk" or "sit" that move the story forward; Human XP has 3-D animations of these verbs, showing a computer what they mean in a world that is similar to our own.

Human XP frames have a continuous clock and show how long the action takes. This continuity of time constrains the pace of the Human XP scenes and makes them easier for a program to analyze. For example, the frames in the "Max breaks the vase" scene must all be connected by sub frames that show finer details of how Max walks to the vase, how he reaches to grasp it, etc. In contrast, a comic strip shows separate snapshots of actions, leaving the audience the job of filling in the details from their own commonsense[1]. But these details are precisely what we would like Human XP to supply.

Finally, the use of the maxometer in Human XP allows a far richer representation of perceptions and mental states than comic strip thought bubbles. The maxometer not only shows simultaneous feelings, but also makes it easier to depict desires, goals, memories, and the hidden meanings behind dialog utterances.

CONCLUSION

The Human Experience Project proposed here advocates a single integrated environment to capture the full range of the human experience: objects in 3-D space, time ordered events, mental states, body perceptions, goals, hidden meanings, and different points of view. The initial focus of Human XP is on everyday scenes of childhood. The focus on child experiences provides a natural limit on the type of everyday knowledge that we want to collect but still represents a valid subset of commonsense that we wish to model.

One obvious disadvantage of Human XP over other internet-based commonsense collection projects is its cost to implement and complexity to use. More detailed requirements are needed before we can evaluate the cost. However, it is clear that virtual worlds offer a huge potential to collect new types of data that reflect the human experience.

ACKNOWLEDGMENTS

Many thanks go to Margit Link-Rodrigue, who did the graphic illustrations and offered valuable suggestions.

REFERENCES

1. Bacon, William F. (1995) What everyone knows about attention; in Michael T. Cox and Michael Freed (Eds.) *Representing Mental States & Mechanisms: Papers from the AAAI Symposium. Technical Report SS-95-05*, March 27-29, Stanford, California.
2. Dyer, M.G. (1983) *In-Depth Understanding: A Computer Model of Integrated Processing for Narrative Comprehension*, Cambridge, MA, MIT Press.
3. Lieberman, H., Smith D., and Teeters, A. (2007) Common Consensus: A Web-based game for collecting commonsense goals. Presented at the Commonsense Workshop at the ACM International Conference on Intelligent User Interfaces (IUI-07), January, Honolulu.

[1] The expectation that the audience will fill in the details between comic frames, referred to as *closure,* is an important aspect of comic strip story-telling (McCloud, 1994).

4. Lenat, D. and Guha, R. (2000) *Building Large Knowledge Based Systems: Representation and Inference in the Cyc Project*; Reading, Massachusetts, Addison-Wesley.
5. McCloud, S. (1994) *Understanding Comics: The Invisible Art*, New York, Harper Perennial.
6. Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002) Open Mind Common Sense: Knowledge acquisition from the general public, *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems. Lecture Notes in Computer Science (Volume 2519)*, Springer-Verlag, Heidelberg.
7. Singh, P., Barry, B., and Liu, H. (2004) Teaching machines about everyday life, *BT Technology Journal*, 22(4):201-210.
8. Speer, R. (2007) Open Mind Commons: An inquisitive approach to learning common sense, *Proceedings of the Workshop on Common Sense and Intelligent User Interfaces*, January 28, Honolulu, Hawaii.
9. Williams, R., Barry, B., Singh, P. (2005) ComicKit: Acquiring story scripts using common sense feedback, *Proceedings of the 10th international conference on Intelligent user interfaces, (IUI-05)*, January, San Diego, California.
10. Wooldridge, M. J. (2000) *Reasoning about Rational Agents*, Intelligent Robots and Autonomous Agents Series, MIT Press, Cambridge, MA.