The Human Speechome Project

Deb Roy[1], Rupal Patel[2], Philip DeCamp[1], Rony Kubat[1], Michael Fleischman[1], Brandon Roy[1], Nikolaos Mavridis[1], Stefanie Tellex[1], Alexia Salata[1], Jethran Guinness[1], Michael Levit[1], and Peter Gorniak[1]

[1] Cognitive Machines Group, MIT Media Laboratory
[2] Communication Analysis and Design Laboratory, Northeastern University
dkroy@media.mit.edu

Abstract. The Human Speechome Project is an effort to observe and computationally model the longitudinal course of language development for a single child at an unprecedented scale. We are collecting audio and video recordings for the first three years of one child's life, in its near entirety, as it unfolds in the child's home. A network of ceiling-mounted video cameras and microphones are generating approximately 300 gigabytes of observational data each day from the home. One of the worlds largest single-volume disk arrays is under construction to house approximately 400,000 hours of audio and video recordings that will accumulate over the three year study. To analyze the massive data set, we are developing new data mining technologies to help human analysts rapidly annotate and transcribe recordings using semi-automatic methods, and to detect and visualize salient patterns of behavior and interaction. To make sense of large-scale patterns that span across months or even years of observations, we are developing computational models of language acquisition that are able to learn from the childs experiential record. By creating and evaluating machine learning systems that step into the shoes of the child and sequentially process long stretches of perceptual experience, we will investigate possible language learning strategies used by children with an emphasis on early word learning.

1 The Need for Better Observational Data

To date, the primary means of studying language acquisition has been through observational recordings made in laboratory settings or made at periodic intervals in children's homes. While laboratory studies provide many useful insights, it has often been argued that the ideal way to observe early child development is in the home where the routines and context of everyday life are minimally disturbed.

Unfortunately, the quality and quantity of home observation data available is surprisingly poor. Observations made in homes are sparse (typically 1-2 hours per week), and often introduce strong observer effects due to the physical presence of researchers in the home. The fine-grained effects of experience on language acquisition are poorly understood in large part due to this lack of dense longitudinal data [1].

P. Vogt et al. (Eds.): EELC 2006, LNAI 4211, pp. 192–196, 2006.

In general, many hypotheses regarding the fine-grained interactions between what a child observes and what the child learns to say cannot be investigated due to a lack of data. How are a child's first words related to the order and frequency of words that the child heard? How does the specific context (who was present, where was the language used, what was the child doing at the time, etc.) affect acquisition dynamics? What specific sequence of grammatical constructions did a child hear that led her to revise her internal model of verb inflection? These questions are impossible to answer without far denser data recordings than those currently available.

2 Pilot Study

The Human Speechome Project (HSP) attempts to address these shortcomings by creating the most comprehensive record of a single child's development to date, coupled with novel data mining and modeling tools to make sense of the resulting massive corpus. The recent surge in availability of digital sensing and recording technologies enables ultra-dense observation: the capacity to record virtually *everything* a child sees and hears in his/her home, 24 hours per day for several years of continuous observation. We have designed an ultra-dense observational system based on a digital network of video cameras, microphones, and data capture hardware. The system has been carefully designed to respect infant and caregiver privacy and to avoid participant involvement in the recording process in order to minimize observer effects.

The recording system has been deployed and at the time of this writing (June 2006), the data capture phase is ten months into operation. Two of the authors (DR, RP) and their first-born child (male, now six months of age, raised with English as the primary language) are the participants. Their home has been instrumented with video cameras and microphones.

Our ultimate goal is to build computational models of language acquisition that can "step into the shoes" of a child and learn directly from the child's experience. The design and implementation details of any computational model will of course differ dramatically from the mental architecture and processes of a child. Yet, the success of a model in learning from the same input as a child provides evidence that the child may employ similar learning strategies.

3 Ultra-Dense Observation for Three Years

Eleven omni-directional mega-pixel resolution color digital video cameras have been embedded in the ceilings of each room of the participants' house (kitchen, dining room, living room, playroom, entrance, exercise room, three bedrooms, hallway, and bathroom). Video is recorded continuously from all cameras since the child may be in any of the 11 locations at any given time. In post processing, only the relevant video channel will be analyzed for modeling purposes. Video is captured at 14 images per second whenever motion is detected, and one image

per second in the absence of motion. The result is continuous and complete full-motion video coverage of all activity throughout the house.

Boundary layer microphones (BLM) are used to record the home's acoustic environment. These microphones use the extended surface in which they are embedded as sound pickup surfaces. BLMs produce high quality speech recordings in which background noise is greatly attenuated. We have embedded 14 microphones throughout the ceilings of the house placed for optimal coverage of speech in all rooms. Audio is sampled from all 14 channels at greater than CD-quality (16-bit, 48KHz). When there is no competing noise source, even whispered speech is clearly captured.

Concealed wires deliver power and control signals to the cameras and microphones, and transmit analog audio and networked digital video data to a cluster of 10 computers and audio samplers located in the basement of the house. The computers perform real-time video compression and generate time-stamped digital audio and video files on a local 5-terabyte disk array. With video compression, approximately 300 gigabytes of raw data are accumulated each day. A petabyte (i.e., 1 million gigabyte) disk array is under construction at MIT to house the complete three-year data set and derivative metadata. Data is transferred periodically from the house to MIT using tape storage.

Audio and video recordings can be controlled by the participants in the house using miniature wall-mounted touch displays. Cameras are clustered into eight visual zones (cameras that view overlapping physical spaces are grouped into zones). Eight touch displays are installed next to light switches around the house, each enabling on/off control over video recording in each zone by touching the camera icon. Audio recording can also be turned on and off by touching the microphone icon. To provide physical feedback on the status of video recording, motorized shutters rotate to conceal cameras when they are not recording. The "oops" button at the bottom of the display (marked with an exclamation mark) opens a dialog box that allows the user to specify any number of minutes of audio and/or video to retroactively and permanently delete from the disk array.

4 Data Management

The network of cameras and microphones are generating an immense flow of data: an average of 300 gigabytes of data per day representing about 132 hours of motion-compressed video per day (12 hours x 11 cameras) and 182 hours of audio (13 hours x 14 microphones). In just the first six months we have collected approximately 24,000 hours of video and 33,000 hours of audio. At this rate, the data set is projected to grow to 142,000 hours of video and 196,000 hours of audio by the end of the three year period. Clearly, new data mining tools must be designed to aid in analysis of such an extensive corpus.

We are developing a multichannel data visualization and and annotation system that will enable human analysts to quickly navigate, search, transcribe salient regions of data. Our long term plan is to adapt and apply computer vision techniques to the video corpus in order to detect, identify, and track people and

salient objects. Since the visual environment is cluttered and undergoes constant lighting changes (from direct sunlight to dimmed lamps), automatic methods are inherently unreliable. Thus, similar to our approach with speech transcription, we plan to design semi-automatic tools with which humans can efficiently perform error correction on automatically generated meta-data. The combination of automatic motion tracking with human-generated identity labels will yield complete spatiotemporal trajectories of each person over the entire three year observation period. The relative locations, orientations, and movements of people provide a basis for analyzing the social dynamics of caregiver-child interactions.

5 Modeling *In Vivo* Word Learning

In previous related work, we developed a model of early word learning called CELL (Cross-Channel Early Lexical Learning) which learned to segment and associate spoken words with acquired visual shape categories based on untranscribed speech and video input [2]. CELL was evaluated on speech recordings of six mothers as they played with their pre-verbal infants using toys. This model demonstrated that a single mechanism could be used to resolve three problems of word learning: spoken unit discovery, visual category formation, and cross-situational mappings from speech units to visual categories. The model operated under cognitively plausible constraints on working memory, and provided a means for analyzing regularities in infant-directed observational recordings.

Three simplifications made in CELL may be contrasted with our new modeling effort using the HSP corpus. First, CELL was evaluated on a relatively small set of observations. Caregiver-infant pairs were only observed for two one-hour play sessions, held about a week apart. The data was thus a snapshot in time and could not be used to study developmental trajectories. Second, observations were conducted in an infant lab leading to behaviors that may not be representative of natural caregiver-infant interactions in the home. It is unclear whether CELL's learning strategy would work with a more realistic distribution of input. Third, visual input was oversimplified and social context was ignored. The only context available to CELL was video of single objects placed against controlled backdrops. As a consequence, the model of conceptual grounding in CELL was limited to visual categories of shapes and colors underlying words such as *ball* and *red*. It could not learn verbs (since it did not model actions), nor could it learn social terms such as *hi* and *thank you*.

The HSP corpus overcomes the limitations inherent in collecting small corpora within laboratory settings as was done with CELL. To move beyond the simple speech-to-image semantics of CELL, we will apply new semantic representations including sensory-motor grounded "semiotic schemas" [3] and "perceived affordances" [4,5]. In the latter, stochastic grammars are used to model the hierarchical and ambiguous nature of intentional actions. In [5], sequences of observed movements are parsed by behavior grammars yielding lattices of inferred higher level intentions. Verb and noun learning is modeled as acquiring cross-situational mappings from constituents of utterances to constituents of intention lattices. We

plan to use a similar approach with the HSP data, but with a semi-automatic procedure for learning behavior grammars from video data. Words related to routines (baths, meals, etc.) and names of locations (crib, highchair, etc.) might be modeled on this basis.

6 Conclusions

The Human Speechome Project provides a natural, contextually rich, longitudinal corpus that serves as a basis for studying language acquisition. An embedded sensor network and data capture system have been designed, implemented, and deployed to gather an ultra-dense corpus of a child's audio-visual experiences from birth to age three. We have described preliminary stages of data mining and modeling tools that have been developed to make sense of 400,000 hours of observations. These efforts make significant progress towards the ultimate goal of modeling and evaluating computationally precise learning strategies that children may use to acquire language.

References

1. Tomasello, M., Stahl, D.: Sampling children's spontaneous speech: How much is enough? Journal of Child Language **31** (2004) 101–121.
2. Roy, D., Pentland, A.: Learning words from sights and sounds: A computational model. Cognitive Science **26(1)** (2002) 113–146.
3. Roy, D.: Semiotic schemas: A framework for grounding language in action and perception. Artificial Intelligence **167(1-2)** (2005) 170–205.
4. Gorniak, P.: The Affordance-Based Concept. PhD thesis, Massachusetts Institute of Technology (2005).
5. Fleischman, M., Roy, D.: Why are verbs harder to learn than nouns? Initial insights from a computational model of situated word learning. In: Proceedings of the 27th Annual Meeting of the Cognitive Science Society (2005)
6. Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guinness, Michael Levit, Peter Gorniak. (2006). The Human Speechome Project. Proceedings of the 28th Annual Cognitive Science Conference.