Modeling an Unstructured Driving Domain: A Comparison of Two Cognitive Frameworks

Bradley J. Best (bjbest@adcogsys.com)
Adaptive Cognitive Systems, 1942 Broadway St. #305
Boulder, CO 80302 USA

Kevin R. Dixon (krdixon@sandia.gov)
Sandia National Laboratories, PO Box 5800 MS1188
Albuquerque, NM 87185-1188 USA

Ann Speed (aespeed@sandia.gov)
Sandia National Laboratories, PO Box 5800 MS1188
Albuquerque, NM 87185-1188

Michael D. Fleetwood (mdfleetwood@gmail.com)

ABSTRACT: *This paper outlines a comparison between two cognitive modeling frameworks: Atomic Components of Thought – Rational (ACT-R; Anderson & Lebiere, 1998) and a framework under development at Sandia National Laboratories. Both frameworks are based on the cognitive psychological literature, although they represent different theoretical perspectives on cognition, with ACT-R being a production-rule-based system and the Sandia framework being a dynamical-systems or connectionist-type approach. This comparison involved a complex driving domain in which both the car being driven and the driver were equipped with sensors that provided information to each framework. The output of each framework was a classification of the real-world situation that the driver was in, e.g., being overtaken on the autobahn. Comparisons between the two frameworks included validation against human ratings of the driving situations via videotapes of driving sessions, along with twelve creation and performance metrics regarding the method and ease of framework population, processor requirements, and maximum real-time data sampling rate.*

1. Introduction

Despite many years of computational modeling in cognitive psychology, the vast majority of models are dedicated to understanding and replicating human behavior on laboratory tasks. However, there are a few modeling frameworks that are being used in real-world domains such as augmenting cognition in various high-consequence environments (Schmorrow, 2005; also ACT-R, SOAR, e.g., Anderson and Lebiere, 1998; Newell, 1990;; Zachary & Le Mentic, 1999), and for emulating human behavior in high-fidelity modeling and simulation environments (e.g., TAC-AIR-SOAR, e.g., Jones, Laird, Nielsen, Coulter, Kenny, and Koss, 1999). Interestingly, there have been relatively few studies that have explicitly set out to compare the relative strengths and weaknesses of these frameworks (e.g., AMBR, Gluck and Pew, 2001, 2005; NASA AvSP SWAP HPM, Leiden and Best, 2007) despite the importance for the understanding afforded by such comparisons – especially in high-consequence domains.

This paper reports a comparison between ACT-R, a production-rule-based system, with a cognitive framework under development at Sandia National Laboratories, a system similar to a recurrent neural network, in a complex driving domain in which both the driver and the car were equipped with various sensors. This comparison evolved out of work that had previously been done by Sandia in the driving domain (Dixon et al., 2005) for the DARPA Augmented Cognition program. The goal of this previous research was to build a model that could interpret the driving context encountered by the driver in real time based on sensor inputs from both the car and driver. Because of the difficulty and complexity of the domain, we were interested in comparing and contrasting radically different cognitive architectures in this real-world domain. The comparison included model agreement with human raters' judgments of context, along with performance metrics such as the relative ease and methods by which models are built, training-data requirements, maximum real-time data sampling rate, and processor needs. The goal of this multidimensional comparison was to evaluate the relative strengths and weaknesses of each architecture for the driving domain, with a longer-term goal of conducting similar multidimensional comparisons in other domains and using additional cognitive frameworks. The remainder of the paper details the two cognitive modeling frameworks considered, the domain and method of comparison, and finally the results and discussion.

2. The Cognitive Frameworks

The two cognitive frameworks considered in this comparison are the widely used ACT-R system and a framework that has been under development at Sandia National Laboratories since 1999.

2.1 ACT-R

ACT-R is a unified architecture of cognition developed over the last 30 years at Carnegie Mellon University. At a fine-grained scale it has accounted for hundreds of phenomena from the cognitive psychology and human factors literature. The version employed here, ACT-R 5.0, is a modular architecture composed of interacting modules for declarative memory, perceptual systems such as vision and audition modules, motor systems such as a manual module, all synchronized through a central production system. This modular view of cognition is a reflection both of functional constraints and of recent advances in neuroscience concerning the localization of brain functions. ACT-R is also a hybrid system that combines a tractable symbolic level that enables the easy specification of complex cognitive functions, with a subsymbolic level that tunes itself to the statistical structure of the environment to provide the graded characteristics of cognition such as adaptivity, robustness and stochasticity.

The central part of the architecture is the production module. A production can match the contents of any combination of buffers, including the goal, which holds the current context and intentions, the retrieval buffer which holds the most recent chunk retrieved from declarative memory, visual and auditory buffers that hold the current sensory information, the manual buffer which holds the current state of the motor module, e.g., typing, as well as buffers defined specially for the task. The highest-rated matching production is selected to effect a change in one or more buffers, which in turn trigger an action in the corresponding modules. This can be an external action or an internal action (e.g., requesting information from memory).

Although the production matching mechanism is rule-based, it also has many desirable "soft" properties such as adaptivity to changing circumstances, generalization to similar situations, and variability. The key mechanism that supports this capability is the utility learning mechanism, which allows the probabilistic selection of productions based on their learned cost and likelihood of success.

During the matching phase, production rules whose conditions match perfectly to the current state of various information buffers qualify to enter the conflict set. Since ACT-R specifies that only one production can fire at a time, the rule with the highest expected utility is selected from the conflict set as the one to fire. The utility of a production rule is learned by a Bayesian mechanism as a function of its past history to reflect the probability and cost of achieving its goal. This is reflected in the formula for expected gain of a production:

$$E = PG - C$$

Expected gain, then, is the product of the probability of a production's success (P) and the value of the goal (G), minus the expected cost (C) of executing the production. The probability of success is given by:

$$P = Successes / (Successes + Failures)$$

The probability of success, according to this formula, is the ratio of successes to the total number of experiences. Successes and failures are both defined in terms of a set of priors and the subsequent learning:

$$Successes = \alpha + m$$
$$Failures = \beta + n$$

These quantities all combine to determine the likelihood of a production's selection in the conflict resolution phase. That likelihood is given by:

$$P\ (production) = e^{Ei/t} / \Sigma j\ e^{Ej/t}$$

In this equation, t represents the value of the noise used. Noise is a central component to allow selection of productions that are currently less likely to succeed based on past history. This is an essential feature of systems that operate in environments characterized by changing base rates, and has been used to explain human probability matching behavior (Holland, 1975).

Two versions of the ACT-R model were initially built – rule-based and instance-based. Despite these monikers, both types incorporate rules in the procedural memory module and instances in the declarative memory module. A primary difference between the two is in the method of population – the rule-based version requires an external entity (e.g., human or machine-learning algorithms) to generate the rules that populate its procedural memory, whereas the instance-based model generates a knowledge base automatically as it encounters incoming sensory data.

2.2 Sandia's Cognitive Framework

Sandia's Cognitive Framework (SCF), has been in development at Sandia National Laboratories since 1999 (e.g., Abbott et al., 2006; Forsythe & Xavier, 2006; Jordan et al., 2002) and is written in C++. The SCF comprises a sparsely interconnected semantic network that feeds into a context recognition memory, which is also a sparsely connected network (see Dixon et al., 2005, for a more detailed dynamical-systems formulation of the SCF). Nodes in the semantic network represent domain-specific concepts, which in this case included Brake Pedal Force, Front Wheel Speeds, Turn Signal Status, and so forth. Contexts, on the other hand, represent unique patterns of activation across the semantic network that indicate different physical situations. For the domain considered in this paper, the contexts include "being overtaken," "approaching a slow-moving vehicle," or "coming to an intersection."

The perceptual interface provides sensory information to the semantic network. In this case, the perceptual interface maps the sensors from the vehicle into a format that the framework can understand. This information activates nodes in the semantic network, which spreads to other semantically related nodes in the network. A semantic relationship is determined by how related a particular person

considers two concepts. For example, most people tend to identify a semantic relationship between the concepts of "black" and "white." The strength of the spreading activation is determined in part by the level of activation for the sending node and in part by the strength of the semantic link between the sending and receiving nodes. This spreading activation results in a pattern of activation emerging across the semantic network. Context recognition is the similarity between this temporal pattern of activation and context templates that the framework has stored. While multiple candidate templates may be partially activated, ultimately only one context is recognized and acted on.

The approach of the SCF has been to model specific individuals, as opposed to modeling generic human behavior. As such, the validation efforts have focused on agreement between SCF context recognition and the judgment of the specific human being modeled. Validation efforts have been applied in domains such as a military air-traffic control simulator, early detection of insider crimes such as embezzlement, and determination of the target of attack in a physical security environment. Across domains, modeling efforts achieve 85% to 95% agreement between the model and its modeled human (Abbott et al., 2006; Dixon et al., 2005; Jordan et al., 2002).

3. The Domain and Method of Comparison

For the comparison, we used an SCF-based model that had already been built for the DARPA Augmented Cognition project, and two ACT-R models created specifically for the purposes of this comparison. The SCF model was designed to assist drivers by identifying potentially dangerous driving situations and taking measures to mitigate this load to improve performance.

The data for these experiments were collected in August of 2004 on the autobahns and village roads south and east of Stuttgart, Germany. The test vehicle was an instrumented S-class Mercedes sedan, and the test subjects were five DaimlerChrysler employees (not professional drivers) who consented to be recorded for the experiment. Each of the subjects was recorded driving three loops on a pre-selected set of roads. Each loop comprised about 200 kilometers of both autobahn and village driving. The roads were unaltered for the experiment, in that no attempt was made to modify the traffic patterns, traffic signals, or pedestrian activity. Data collected from the car and driver included:

- Brake Pedal Force
- Front Left Wheel Speed
- Front Right Wheel Speed
- Rear Left Wheel Speed
- Rear Right Wheel Speed
- Accelerator Pedal Deflection
- Left Blinker On
- Right Blinker On
- Steering Wheel Angle
- Distance to Nearest Moving Object
- Object Speed
- Lateral Acceleration
- Blind Spot Detector
- Driver Head Yaw
- Driver Seating Position Yaw
- Seat RMS Pressure Derivative
- Seat Pressure Backward Correlated RMS

A typical signal in the vehicle was sampled at 50Hz. In addition to these raw sensors, we created higher-order features, such as a "Collision Danger" signal and "Average Front Wheel Velocity" from the constituent raw signals. More details on the experimental setup of the driving conditions can be found in Bruns et al. (2005).

For the DARPA work, it was decided that it would be most useful to identify a pre-defined set of potentially difficult driving situations. These situations were selected based on their utility in identifying driving situations and predictability given the sensor suite. For example, getting on an onramp for the autobahn is a potentially difficult situation, there is a reasonable chance of experiencing this during the experimental conditions, and there is the potential that the available sensors may be used to infer this situation. Being in a construction zone is a potentially difficult situation, but it may never be encountered during the experiment. Consequently, no data may exist to support or reject the hypothesis. The driving situations used were:

- Approaching or waiting at intersection
- Leaving intersection
- Entering onramp or high-speed roadway
- Being overtaken
- High acceleration or dynamic state of vehicle
- Approaching slow-moving vehicle
- Preparing or changing lanes

In order to generate human ratings of specific driving contexts, we asked several users to give their perception of when the above situations were occurring. Scorers watched videos of a camera pointed out the windshield of the car during the experiments. Scorers used a graphical user interface in which they could indicate their contextual interpretation at any given time. Scorers were told they could choose more than one context at a time and they could rewind the video to correct any errors in judgment, such as spurious clicks or mistaken interpretations. The scorers were two researchers from Sandia National Laboratories.

Training and testing data constituted data sampled from the vehicle and driver-based physiological sensors at 4Hz. In total for the five subjects, we had almost 24 hours of driving data (343,946 samples) that were labeled by the scorers. As mentioned previously, each driver drove three loops (A, B, C) on a pre-defined circuit of roads during the experiment, such that each subject drove on the same roads as the other subjects during the corresponding loop. We randomly selected two loops (loops A and B) to train the classifier (18.3 hours or 263,851 samples) and one loop (loop C) for cross validation (5.6 hours or 80,595 samples). The same loops were put into each set for each driver. Data used to train and test each framework to recognize the seven

chosen situations was identical and was a large subset of the data collected from the car.

The initial rule-based ACT-R model was built by first assessing the relative utility of each of the sensors and higher-order features for classifying the 7 situations using a multiple linear regression analysis for each situation and an analysis of the hierarchical structure of the various sensors. Some of the driving situations (e.g., approaching an intersection), however, did not have features that could be straightforwardly encoded by hand. For these cases, the C4.5 rule induction algorithm (Quinlan, 1993) was used to build a classifier decision tree which was then translated to ACT-R production rules.

The instance-based ACT-R model was built by first developing a decision skeleton of rules that relied on retrieving similar instances from declarative memory. This set of rules would attempt to retrieve a previous situation similar to the currently encountered situation and then apply the previous (successful) decision. Initially, due to lack of experiences, this model would be forced to guess, but over time after encountering more situations its declarative memory would become populated with many prior instances allowing it to effectively analogize to the current situation.

A model using the SCF was created using a supervised learning algorithm based on the training data from loops A and B. This learning algorithm estimates the optimal parameters for the dynamical-systems equations in the SCF for the given input-output pairs in the training data (Dixon et al., 2005). Using the 18.3 hours of training data from loops A and B, the SCF learning algorithm converged in about 1 hour of running time on a laptop computer and achieved about 95% agreement with the human raters on the 5.6 hours of cross-validation data from loop C.

Once the models were built and tested, they were compared on 12 quantitative and qualitative characteristics, which are presented in Tables 1-3.

4. Results

Tables 1-3 present results of these comparisons for the ACT-R Rule-based model, the ACT-R Instance-based model, and the SCF respectively.

Table 1: Results for the ACT-R Rule-based model

Characteristic	ACT-R: Rule Based
Agreement with Human Raters	96% (using 5.2% of training set – when 14% training set used, drops to 47%)
Max. real-time sampling rate	500Hz
Ease and Method of Building	Can be done easily with machine learning although manually-created rules sometimes do best
Data Needs	Substantial risk of overfitting
Processor Needs	Less intensive than instance-based – can easily run on typical laptop
Storage of Experience	Bayesian probabilities
Method for Recognizing Contexts	Multiple rules per context
Knowledge Representation	Decision trees / production rules
Primary Intended Use	Replicate human behavior in laboratory tasks
Types of Plausibility	Based on J. Anderson's Adaptive Control of Thought – Rational psychology theory – neurophysiology validation underway
Learning /adaptability / ability to change over time	Symbolic and subsymbolic learning can take place

Table 2: Results for the ACT-R Instance-based model

Characteristic	ACT-R: Instance Based
Agreement with Human Raters	89%
Max. real-time sampling rate	40-50Hz
Ease and Method of Building	Cannot build manually – instances automatically created in declarative memory based on training set
Data Needs	Risk of overfitting not as substantial as rule-based
Processor Needs	Very processor intensive – can run on conventional laptop, but takes substantial time to train
Storage of Experience	Many instances per category
Method for Recognizing Contexts	Nearest neighbor comparison to individual instances based
Knowledge Representation	Instances
Primary Intended Use	Replicate human behavior in laboratory tasks
Types of Plausibility	Based on J. Anderson's Adaptive Control of Thought – Rational psychology theory – neurophysiology validation underway
Learning /adaptability / ability to change over time	All novel instances are recorded – once recorded they are not lost

Table 3: Results for the SCF model

Characteristic	SCF
Agreement with Human Raters	95%
Max. real-time sampling rate	Several hundred thousand Hz
Ease and Method of	Depending on application can be very easy. Some apps still require manual

Building	knowledge elicitation (KE). Automatic KE can use text, machine transactions
Data Needs	Low risk of overfitting
Processor Needs	Not intensive – can easily run on typical laptop
Storage of Experience	Functionally none
Method for Recognizing Contexts	Highest weighted sum of activation across semantic network that surpasses recognition threshold – multiple contexts active at same time
Knowledge Representation	Symbolic / connectionist hybrid
Primary Intended Use	Augmenting human cognition in data intensive, real-world environments
Types of Plausibility	Psychology much more well-developed than neurophysiology but neither is as well-developed as ACT-R
Learning /adaptability / ability to change over time	Currently none inside framework – once built the model is static

5. Discussion

With regards to agreement with human raters, for this particular task, the models performed comparably well. The major differences between them, such as real-time sampling rate and the presence of learning, did not have a significant impact on the effectiveness of any of the models. By and large, it seems that in terms of agreement with humans, the performance of the model is determined more by the team doing the modeling than by the architecture of the framework being used (cf. AMBR comparisons; Gluck and Pew, 2001, 2005). Capable modeling teams will produce models that successfully capture the target behavior, sometimes in spite of the architecture used, rather than due to it. As a result, more 'practical' measures of the architectures (e.g., processor needs, ease of model building) may be more important in selecting an architecture than the underlying techniques used for learning and performance.

However, there are some other critical differences between the models that have implications for their applicability to different situations. For example, the difference in the ability for each framework to learn from its past experiences can have significant impact on the way a framework is implemented in a field situation – if users must take a model offline, update it, then re-engage it, the requisite skills are significantly different than if the user can engage the model for the first time and have it running transparently in the background. In this way, the SCF is currently limited with regards to its potential user base in that there is no learning. By way of comparison, the instance-based version of ACT-R could be much more widely used by people with less computational modeling experience, or could be more effectively deployed in situations where there is no time for humans to be constantly updating the model.

A second critical difference between frameworks has to do with the ability of each to handle large amounts of data in real time. Many real-world problems involve the deployment of sensors that produce huge amounts of data, sometimes the rate of change in this data is critical in order to develop an understanding of the dynamics of the object being sensed. The ability of a framework to handle high sampling rates in real time becomes critical to it's ability to make sense of those data – especially if critical patterns emerge over very short timespans. The SCF was specifically designed to handle these high-input kinds of situations. The rule-based version of ACT-R is likely sufficient for many such situations, but the instance-based version is limited in comparison to the other two. In particular, as the instance-based ACT-R model runs, instances accumulate in declarative memory and the calculations required to track changes in activation over time and select chunks based on memory retrieval patterns can become prohibitively expensive from a practical perspective (i.e., the instance-based model slows down as it gains experience).

A third critical difference between the technical capabilities of the three frameworks concerns the data needed in order to build/train the model. One very interesting finding in this study was the drastic decline in the performance of the rule-based version of ACT-R when 14% of the available training data were used. This is a result of over-fitting that occurs when building the decision tree classifier used to produce the rule-based model. Decision tree classifiers are known to be vulnerable to over-fitting (see Quinlan, 1993, for a discussion of this issue), especially in domains that have noisy data, and thus care must be taken to safeguard against that possibility. The method used in the current modeling effort to prevent over-fitting was to selectively bias data sampling towards cases that were positive instances of one of the categories. Since the source data were sparsely populated with these instances (approximately 5% of the data represented positive cases), they were more 'informative' to the decision tree classifier. As the proportion of instances increased to values approaching 10% of the available data, however, an excess of negative training cases came to overwhelm the classifier because all of the positive instances had already been used.

Thus, though the decision tree method learned from the data, given the data characteristics, learning had to be biased by selecting a non-random subset of the available data. The instance-based method, on the other hand, produced its best results using the complete data. These two methods, though implemented in the same architecture, produce widely divergent results in terms of scalability, effort needed to adapt to data sets, and computational efficiency.

Finally, there are some task idiosyncrasies that are important to note. First, while the task for each model was to identify the contexts the driver was in, the comparison between the models actually involved a confound: the human raters used different data than the models did – the human raters used video taken during the driving and the models used data from the sensors used to instrument the car and the driver. Ultimately, one would think that there

would be some correlation between the behaviors of the driver due to context recognition and the data coming from the car (e.g., brake pedal force is likely correlated to the driver recognizing a need to reduce the speed of the car), but those correlations were not made explicit, and our analyses demonstrated that the relationships between the sensors and driving situations are by no means simple. This lack of explicit correlation between sensors and human-interpreted context is potentially a problem as the people rating the videos were not the same people who drove the car and the ratings were done offline. Similarly, the data provided to each model were data gathered from multiple people – videos were of several people driving and the video ratings were done by several, rather than one person. Finally, the categories eventually used by the human raters and by both models were determined *a priori* – it is unclear if all of the humans involved in the task would have provided those seven categories on their own.

Ultimately, the relative usefulness of any framework will be determined largely by the needs of the current problem space. However, this comparison points out some specific directions for future technical improvements in each of the frameworks in order to make them more widely useable. This kind of comparison can also inform current psychological and neurophysiological theories about the fundamental characteristics of human cognition, which can in turn lead to development of novel cognitive architectures.

6. Acknowledgements

Portions of this work were funded by the DARPA Augmented Cognition Program and portions were funded by Sandia National Laboratories' Lab Directed Research and Development Program. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

7. References

Abbott, R., Bauer, T., Berry, N., Forsythe, C., Larson, K., McClain, J., Shaneyfelt, W., Small, D., Speed, A., Xavier, P., Wagner, J., Skocypec, R. (2006). Next generation intelligent systems (Augmented Cognition) Grand Challenge LDRD Final Report. Sandia National Laboratories SAND2006-2506.

Anderson, J. R. & Lebiere, C. (1998). The Atomic Components of Thought. Mahwah, NJ: Erlbaum.

Bruns, A., Hagemann, K., Schrauf, M., Kohmorgen, J., Braun, M., Dornhege, G., Muller, K., Forsythe, C., Dixon, K., Lippitt, C.E., Balaban, C.D., & Kincses, W.E. (2005). EEG- and context-based cognitive-state classifications lead to improved cognitive performance while driving. Proceedings of the 1st International Conference on Augmented Cognition, Las Vegas, Nevada.

Dixon, K.R., Lippitt, C.E., & Forsythe, J.C. (2005). Supervised machine learning for modeling human recognition of vehicle-driving situations. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, Alberta, Canada.

Forsythe, C., & Xavier, P.G. (2006). Cognitive models to cognitive systems. In C.Forsythe, M.L.Bernard, & T.E. Goldsmith (Eds.) Cognitive Systems: Human Cognitive Models in Systems Design., Mahwah, NJ: Lawrence Erlbaum Associates.

Gluck, K. A., & Pew, R. W. (2001). Overview of the agent-based modeling and behavior representation (AMBR) model comparison project. In Proceedings of the 10th Computer Generated Forces and Behavioral Representation Conference. 10TH-CGF-066. 3-6. Orlando, FL: Division of Continuing Education, University of Central Florida.

Gluck, K.A., & Pew, R.W. (2005). (Eds). Modeling Human Behavior with Integrated Cognitive Architectures: Comparison, Evaluation, and Validation. Mahwah, NJ: Erlbaum.

Holland, J. H. (1975). Adaptation in Natural and Artificial Systems. Ann Arbor: University of Michigan Press.

Jones, R. M., Laird, J. E., Nielsen P. E., Coulter, K., Kenny, P., and Koss, F. (1999). Automated intelligent pilots for combat flight simulation, AI Magazine, 20, 27-42.

Jordan, S., Forsythe, C., Speed, A., Wenner, C. & Goldsmith, T. E. (2002). Extensibility of knowledge-based human agent simulation. Sandia National Laboratories SAND 2002-3714, Albuquerque NM.

Leiden, K., & Best, B. J. (2007). A cross-model comparison of human performance modeling tools applied to aviation safety. In D. C. Foyle & B. L. Hooey (Eds.) Human Performance Modeling in Aviation: Surface Operations and Synthetic Vision Systems. Mahwah, NJ: Lawrence Erlbaum.

Newell, A. (1990). Unified Theories of Cognition. Boston, MA: Harvard Press.

Quinlan, R. (1993): C4.5: Programs for Machine Learning. Morgan Kaufmann, San Diego.

Schmorrow, D.D. (Ed.) (2005). Foundations of Augmented Cognition. Mahwah, NJ: Lawrence Earlbaum Associates.

Zachary, W. & Le Mentec, J. (1999). A framework for developing intelligent agents based on human information processing architecture. In Proceedings of the IASTED International Conference, Honolulu, Hawaii.