DIRECTLY COMPARING COMPUTER AND HUMAN PERFORMANCE IN LANGUAGE UNDERSTANDING AND VISUAL REASONING

CSE Technical Report 288

Eva L. Baker
Elaine L. Lindheim
Josef Skrzypek

Center for Technology Assessment
UCLA Center for the Study of Evaluation

May, 1988

The research reported herein was conducted with partial support from the Office of Naval Research, Defense Advanced Research Projects Agency, pursuant to grant number N00014-86-K-0395. However, the opinions expressed do not necessarily reflect the position or policy of this agency and no official endorsement by this agency should be inferred.

This report was presented at the 1988 Annual Meeting of the American Educational Research Association, New Orleans

Please address inquiries to: CSE Dissemination Office, UCLA Graduate School of Education, 405 Hilgard Avenue, Los Angeles, California, 90024-1521

Introduction

The focus of this report and our project is research on how to evaluate artificial intelligence (AI) implementations. Our interest came from earlier work (Baker & Atwood, 1985) evaluating instructional tutoring systems and the conflict we observed between fulfilling the promises made in AI program development and the exploration of new research ideas that popped up in the course of development. How could smart computers be assessed or even described in valid, useful ways? Because we represented a particular community (evaluation and measurement) our answers took a predictable bent. We looked to see how we could apply the rich history in measurement and evaluation to the particular technology assessment problems AI poses.

Without belaboring the numerous challenges we faced, we have forged a cooperative research program with AI researchers and are now clearly underway. Although our work includes assessment of expert systems and the overall problem of technology assessment, this paper will address two of the major areas for our first phase work: natural language understanding and vision systems. Our particular goal was to determine the extent to which AI systems can be described in terms of similar performance by groups of people. The major goal of our program was to develop appropriate evaluation models that would allow us to describe and assess AI programs in a variety of ways.

Technical Approach

Our technical approach had three major subdivisions. First, and separately for natural language and vision domains, we planned to create maps of the problems that had been addressed, with varying degrees of success, by AI programs. In natural language, a sourcebook approach was taken and will be described by my colleague. This approach assumes that a great number of problems and examples of implementations exist and can be categorized. After such examples are collected, then a conceptual framework, reviewed by psychologists, computer scientists, educators, and linguists, can result in a category system. This system will enable us to classify (probably multiply) any given implementation in terms of the problems (or goals) it undertakes. In the vision area, the numbers of examples of higher order visual processing, i.e., seeing with understanding, are relatively small, so a different approach to map making was taken. The map was concocted top-down, by looking at the major class of problems, and then seeing the few examples created.

The second phase of effort focused on identifying criteria by which any given program could be assessed, and required looking at functional specifications, the cognitive processing model underlying development, the nuts and bolts of design and implementation, and constraints.

The third, and to our view, most interesting area was the problem of describing computer implementations in terms of human performance. In a sense, we wanted to test how smart these systems were, or better still, what kind of achievement such systems were capable of. The metaphor of human achievement testing guided our work. For the remainder of this report, we will describe our efforts in benchmarking in natural language and vision.

Natural Language Benchmarking

In the simplest terms, we intend to norm a given NL system's performance on a sample of people. We have begun to create tests that measure the language functions of target programs and to benchmark systems in terms of the characteristics and abilities of the human performance. It is our intention to apply our approach to a sample of natural language implementations. A typical NL implementation used for research might consist of a discrete piece of text, perhaps a description of a common scene (Dyer, 1983). The goal of the NL developer is to demonstrate that the computer can understand both literally and inferentially what has happened. The mode of demonstration is asking questions of the system. In order to respond, complex rules are programmed describing explicitly the context needed to answer the questions. In this simplest of cases, our benchmarking approach would require the following:

1. Develop domain specifications appropriate to generate questions about the text segment.

2 Generate test items appropriate to the text.

3. Create a measure consisting of the NL developer's questions and our own and administer to "norming" or referent groups.

4. Describe NL system performance in terms of the group whose responses are most comparable to those of the system.

We are also testing the feasibility that comparisons among systems can be made. After completing the above task for each of two separate NL programs, if comparisons were desired, an additional set of steps would follow.

1 One or more constructs would be posited.

2. Anchor items would be developed and administered to the same norming groups.

3. Analyses to assess the equating options would be conducted.

Clearly, this approach appears to gloss over some important differences between systems and people. For example, we have not decided (nor is it really feasible) to measure explicitly important other language performances the comparison group can accomplish in addition to those targeted by the system. People are obviously infinitely more creative and proficient in language than any system yet or to be devised. Yet, a quick reading of our project might imply that we will infer that system performance equals human performance. We will not. Conversely, there will be aspects of system performance clearly superior to what people can do -- perfect reliability for one example not familiar to psychometricians. Our approach is exploratory and its utility will depend upon how sensible and understandable our comparisons will be.

The Natural Language System: IRUS

Unlike the simple text based example above, the actual system to which we applied our methods provides us with a greater challenge. The program under study is a natural language query system. Essentially such a program permits the user to ask questions in regular English prose to another computer program, perhaps a database or an expert system. The natural language system, IRUS (Bates, Stallard, & Moser, 1985.), is an interface between the user and the set of information desired for access, and provides a rapid, natural and convenient method for obtaining information. The particular interface we are assessing has been designed, at least so far, to serve as a general purpose interface to a broad range of databases and expert systems. It is a basic syntactic shell that needs to be filled with specifics in order to work. To use IRUS, it must be specifically adapted to a designated database. The particular semantics (content) of the database or expert system must be translated into rules used by IRUS. Table 1 shows a sample of queries that IRUS can deal with, demonstrated in two different domains.

Table 1
IRUS in Two Domains

IRUS in a library science domain:

> "Of the books on Artificial Intelligence, how many have been classified as textbooks?"
>
> "Have there been as many requests for books about medicine this year as we planned for in our budget?"
>
> "Which organizations that we receive reports from have responded to either of our recent questionnaires?"

IRUS in the domain of Navy ships:

> "List the number of ships that are deployed in the Indian Ocean."
>
> "What's the name of the commander of Frederick?"i
>
> "What is Vinson's current course?"

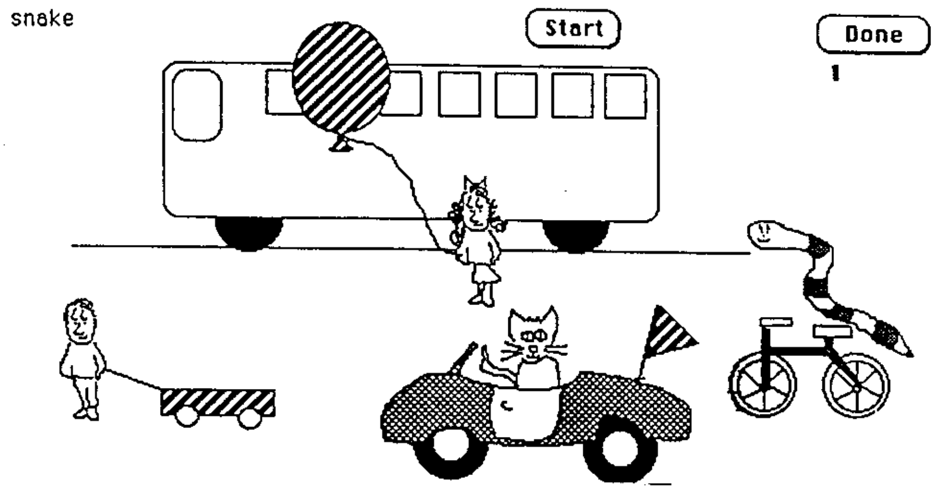
Our knowledge of what IRUS can do has come from a system test of IRUS, where the system successfully answered a series of 165 questions. We have taken these questions, classified them into semantic and syntactic categories, and developed a set of test specifications designed to measure human ability to understand these questions.

Because IRUS is always embedded in a specific domain, e.g., ships status, it is clear that our measurement approach needed to separate out the understanding of the question from the ability to provide the answer. Clearly, making comparisons based on the correctness of answers about the location of Navy ships makes no sense. Because we believed that many of the IRUS query types could be answered by very young children, we decided to develop a measure that would provide for children a very simple database - one consisting of animals, people, houses, their attributes, and positions.

The measure includes a pretest that determines whether students understand the elements in the database. The pretest is shown in Figure 1. By screening out examinees who cannot identify the database elements, we are able to infer that students' selection of the correct answer is based upon their understanding of the question. In our study, unlike the real IRUS applications, the databases function only to permit us to assess language function.

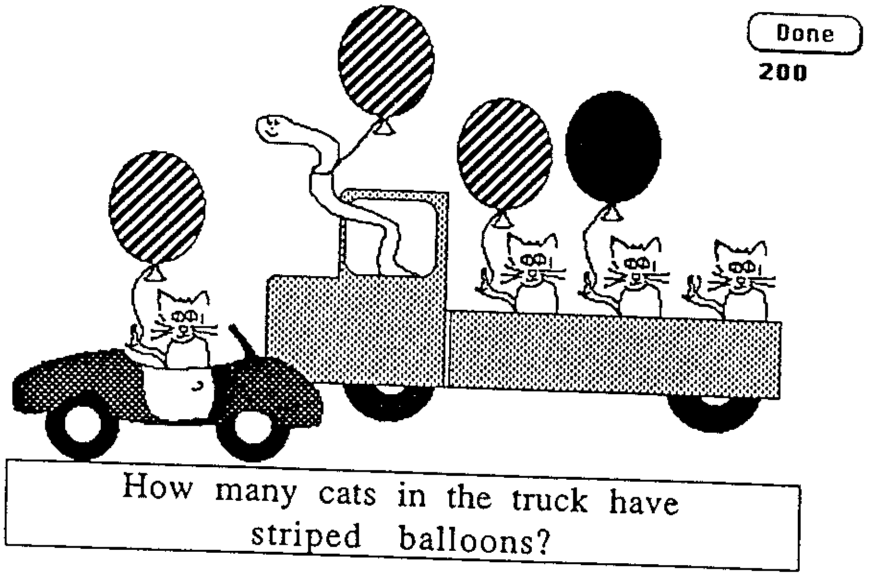
We have included copies of some of the test items presented to students for our prototype test (See Figures 2 and 3). The test was implemented in Hypercard and administered on Macintosh SE computers.

Figure 1
IRUS Pretest



IRUS Vocabulary Pretest. Students are shown a prompt (e.g., "Point to the snake") and asked to respond.

Figure 2
Sample Test Item

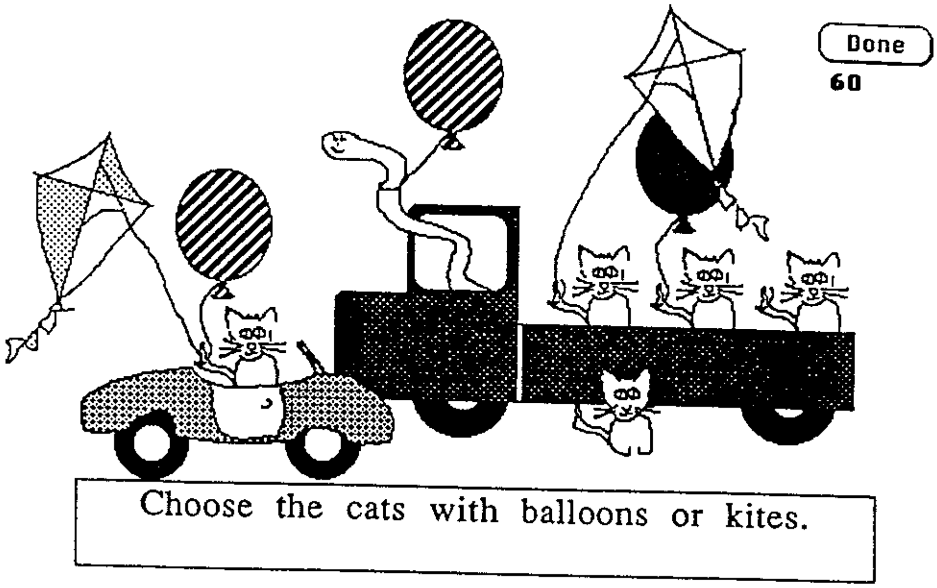


How many cats in the truck have striped balloons?

Comparable IRUS queries:

How many ships in the Third Fleet are C-3?
How many of the ships in Indian Ocian are C-5?

Figure 3
Sample Test Item



Choose the cats with balloons or kites.

Comparable IRUS query:

List the ships that are C4 or that are C5.

Procedures

In order to determine the appropriate language understanding level at which to administer the IRUS test, we have piloted the test with early elementary school and preschool students. Those students who are reading at a second grade level or higher read the questions themselves and the test administrator reads aloud the queries for younger students.

Depending upon the type of query, examinees answer with either an oral response or by pointing to the answer on the computer screen. When the examinee response orally, the administrator types the answer on the screen so that it is entered in the computer transcript of the test. For example, a student might answer the query, "How many cars have striped flags", by saying "four." The administrator types "four" into the the transcript. When the examinee points to an answer on the screen, the administrator uses the mouse to highlight the student's choice. (For example, a student might answer the query "Choose the cats with striped balloons" by pointing to any cats on the screen that fulfilled the requirement. The administrator would click on each animal identified by the student.)

In a talk-aloud procedure designed to validate students' understanding of the questions they have answered, students are asked to explain their responses to the more complex queries. After their response has been entered in the computer, the administrator asks, "Why did you say '. . .'" or "How did you know that ' . . .' was the answer?" These responses are tape recorded for analysis in conjunction with the test transcript.

Results

Results to date indicate the following:

1. Students reading at or above a second grade level generally can recognize all of the elements in the database when those elements are presented in the pretest.

2. Students reading at or above a second grade level have difficulty with those queries that are more than 9 words long and that contain more than one delimiter, when the relationship between delimiters is expressed by either the conjunction "and" or the conjunction "or."

3. Students reading at or above a second grade level often answer a query with a response that is literally incorrect but pragmatically valid. For example, when asked, "How many cars have striped flags?", they may respond by pointing to all the cars that meet the stipulation rather than answering with a specific number.

Vision Benchmarking

Our approach in the vision area focused on looking at common measures of visual tasks already in literature and analyzing these in terms of their appropriateness to vision system evaluation. The approach inverts that undertaken in natural language, where we started with the tasks that computer systems can solve and created measures appropriate for people.

Our theoretical approach posits that the human visual system provides us with a means of comparing machine vision systems with human performance, and it is possible that visual tasks can be constructed to evaluate the performance of human and machine vision systems. In addition, the human visual system provides us with a better understanding of what are the critical underlying set of visual tasks that must be performed by a general purpose system. Along this line, the analysis of visual deficits in the human visual system may shed some light on the functional organization and the neural mechanisms underlying the performance of particular visual tasks. The discovery of a core set of visual tasks along with their possible neural substrates will provide invaluable information on how parts of a *general* propose machine vision system should be constructed.

Visual Tasks

We assume that the human visual system is a working example of a general purpose vision system. Consequently, it should be possible to enumerate most of the visual tasks that such a system must perform. Some of these tasks might be very difficult for machine vision. However, there probably exists a kernel of visual tasks, a subset of which underlies most of the visually guided behavior in an unconstrained environment. Hence, one of the goals of this research effort is to develop a list of these visual tasks.

Visual tasks are a component of many testing batteries. The underlying mechanisms which permit human vision to succeed in such tasks are not well known and are investigated in neuropsychology and cognitive psychology. Understanding these mechanisms would perhaps allow us to define some of the underlying primitive visual functions and transfer these to machine vision. The problem is that we need to select tasks that allow meaningfully comparison of human and machine vision.

The following are a few sample visual tasks taken from published tests. One of these is the Hidden Figures Test CF-1 (Ekstrom, French, Harman, & Dermen, 1976) which requires the participant to find a given shape in some complex picture (the shape undergoes no rotation or size changes when place in the picture). (Figure 4.) This task is difficult for people because of the cluttered background, although we can solve the visual task after we spend a good deal of time on it. a computer program, on the other hand, could accomplish a template match on a run-length encoded image in milliseconds because of size and rotation invariance. Therefore a comparison between human and machine performance based on this task would not be meaningful, i.e., the machine would always win.

Another example test is called the Surface Development Test (Ekstrom, et. al., 1976). In this case, the task is to fit a given surface together into a line drawing representation of the three dimensional object (Figure 5). Some tasks of this type are easy while others are very difficult and might involve very different problems of selecting and representing models, manipulating internal models of complex three dimensional objects, and finally matching the model against the data. Some of these problems, such as geometric reasoning, pose a great challenge to AI in general and in this case the human would most likely always beat the computer.

Figure 4
Sample Question from the Hidden Figure Test

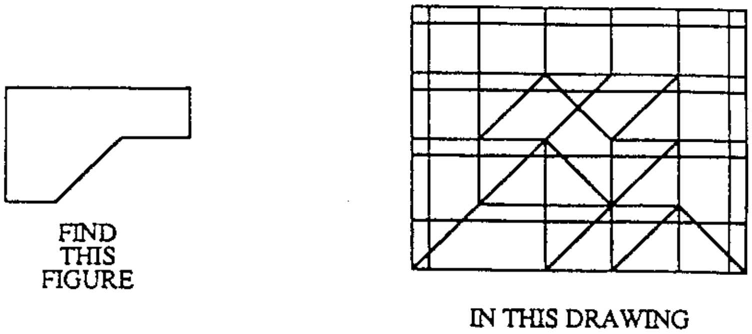
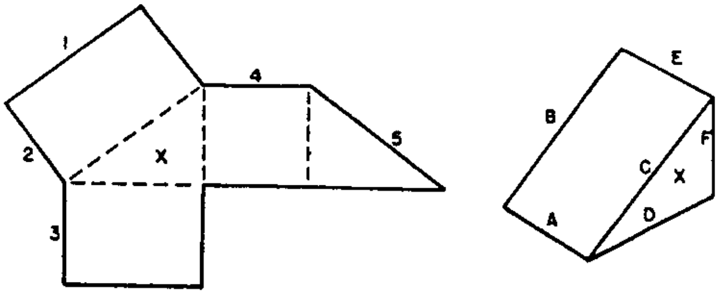
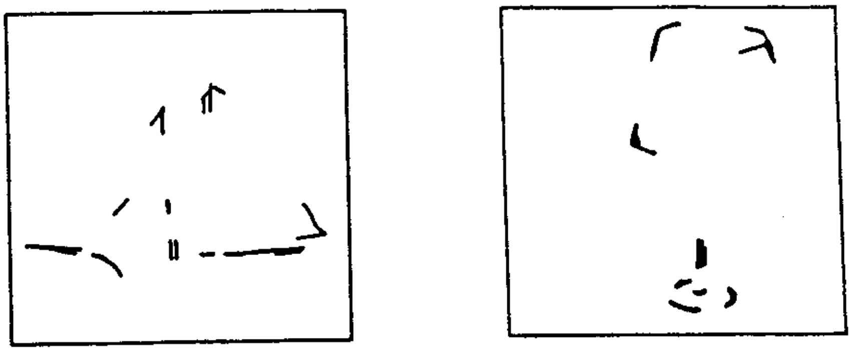


Figure 5
Sample Question from the Surface Development Test



Another task to consider is the Gestalt Completion Test CS-1, which might be closely related to segmentation. In this task, a subject is given incomplete data such as that presented in Figure 6. The subject must determine what the data represents. Problems posed by this test include how to match the incomplete data against a set of models that one has acquired over time. There are many difficult questions to address. What is a model? How are models represented? How much data is needed? How are models matched against incomplete data? How do we implement something like illusory contours that would help in the performance of this type of task? Do we first complete the data by filling in the illusory contours?

Figure 6
Sample Question from the Gestalt Completion Test



We have begun developing a prototype of a visual task sourcebook. The selected visual tasks range in difficulty and in their association with low-level or high-level visual processing. Although the sourcebook is far from complete, it does provide a more extensive set of visual tasks which should conceivably be handled by a general purpose machine vision system.

After we identify a set of tasks, the idea is to demonstrate the extent to which alternative vision programs can solve such tasks and at what performance levels. The tension is, of course, between special purpose systems and specially adapted to solve problems we provide.

We have also convened one meeting of experts in cognitive and physiological psychology, neuroscience, and computer science to consider our approach. This meeting did not result in the creation of a framework for vision system evaluation or for our sourcebook reference system.

Next Steps

We are pursuing the performance testing of IRUS. We recognize that we must build a case for construct validity of our measures. We plan to do so by administering parallel tests to students related to language competency and reading comprehension. The result of these efforts will be to develop more detailed profiles of the characteristics of humans that succeed at natural language tasks. Our second set of tasks will be generalizing our benchmarking approach to other NL programs. We then will have the task of attempting to employ equating techniques to assess the comparability of our measures.

In the visual processing area, we are retrenching and plan to initiate a review by the vision community of the approaches they would use to compare human to machine vision. Based on these recommendations, we will determine whether searching for a set of consistent benchmarks for inclusion in vision programs is a feasible approach.

References

Baker, E.L., & Atwood, N. (1985). *Intelligent computer-assisted instruction (ICAI) study* (Report to Jet Propulsion Laboratory). Los Angeles: UCLA Center for the Study of Evaluation.

Bates, M., Stallard, D., & Moser, M. (1985). The IRUS transportable natural language database interface. In *Expert Database Systems*. Menlo Park, CA: Comming Publishing Company.

Dyer, M.C. (1983). *In-depth understanding*. Cambridge, MA: MIT Press.

Ekstrom, R.B., French, J.W., Harman, H.H., & Dermen, D. (1976). *Kit of factor-referenced cognitive tests*. Princeton, NJ: Educational Testing Service.