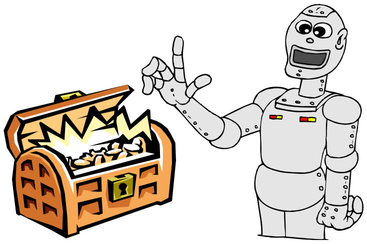
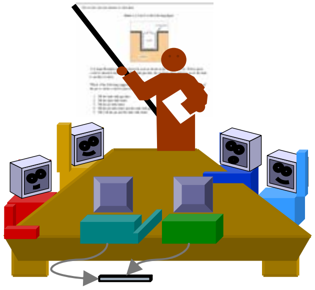
Selected Grand Challenges in Cognitive Science

Search Engine for the Real World
Test Taker
Report Generator
Reading to Learn





Prepared by the MITRE Corporation for DARPA IPTO

October 2005

Overview

This compilation of Selected Grand Challenges in Cognitive Science includes challenge specifications, briefings, supplemental materials, and FY'06 plans for 4 challenges selected by DARPA IPTO: Search Engine for the Real World, Test Taker, Report Generator, and Reading to Learn.

The overview summarizes each challenge and highlights future plans. The rest of the document is divided into individual challenge sections. Each section consists of a summary page and brief descriptions of the attached materials.

Search Engine for the Real World: *Grand Treasure Hunt Adventure*

Mobile robots use navigation and visual recognition to discover objects in a real world treasure hunt

Participants have 20 minutes to locate a treasure within a house. In order to uncover the identity and location of the treasure, the robot participants will be given hints to find 10 other required objects along the way, some of which may reveal additional clues.

In FY'06, we plan to conduct a proof-of-concept treasure hunt to assess challenge feasibility and establish baseline capability. We propose to construct a sample house, implement a communication protocol, and collaborate with UMass Lowell and MITRE robot teams to test and refine.

Test Taker: *Taking the SAT*

An autonomous system will take the SAT® and score in the 50th percentile of high school students taking the examination.

The winning system must take the SAT and score in the 50th percentile of high school students taking the examination the same year. Our investigations revealed that the Math, Reading, and Writing sections of the test require many of the capabilities IPTO seeks in a cognitive system.

For follow-on work, the next step is to meet with College Board and ETS to stimulate their support of this challenge, and then obtain sample tests and build an API and computer-readable test format.

Report Generator: *Handy Andy, the DARPA Essayist*

Automated AI Systems Compete Against Invited Human Contestants

The Handy Andy Challenge is to produce a multi-page report on any topic in response to a user request. It involves at least three subtasks: understand the request, find appropriate content, and produce an informative and well-organized write-up. The assessment will be based on both human and automated measures, maintaining two essential criteria: ranking of reports produced by such metrics will need to remain stable across different sets of judges, and reports that are in fact similar should get similar ranks.

For follow-on work, we propose to design a feasibility pilot applied to After Action Reports.

Reading to Learn: The Scholastic Grand Challenge

An autonomous system will learn from a textbook and answer the questions in the book chapter-by-chapter

This Grand Challenge focuses on having systems "learn" by reading a textbook and passing incremental, chapter-by-chapter tests. This idea arose at the January DARPA Grand Challenge Workshop in discussions of "Reading to Learn". Michael Witbrock and Lynette Hirschman put together a one-day follow-on workshop in Seattle, hosted by Bill Dolan at Microsoft.

The next step in developing the Scholastic Grand Challenge would be to write a prototype "Young Computer's First Reader." This would be a textbook written in simple English, focused on a constrained and structured subject matter (perhaps evolution or geology), consisting of at least four chapters, plus associated problems, along with an evaluation methodology. This could be used to attract participants to demonstrate the feasibility of the Scholastic Grand Challenge. Our plan is to work with Michael Witbrock and possibly other participants in the Scholastic Grand Challenge Working Group, in order to develop the reader and test performance of one or two current systems on such a sample reader.

MITRE Contributors

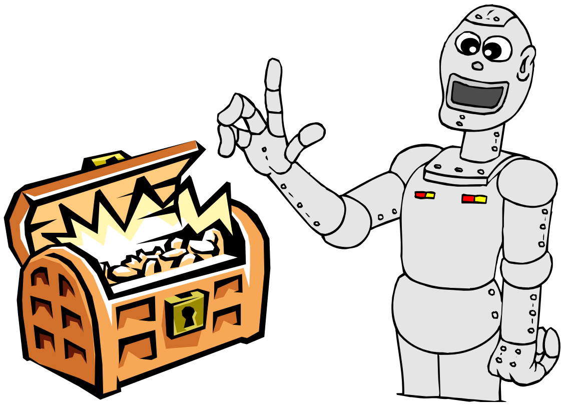
Sam Bayer
Laurie Damianos
Christine Doran

Lisa Ferro
Randy Fish
Lynette Hirschman

Inderjeet Mani
Laurel Riek
Beatrice Oshika

Search Engine for the Real World
The Grand Treasure Hunt

Mobile robots use navigation and visual recognition to discover objects in a real world treasure hunt



Participants have 20 minutes to locate a treasure within a house. In order to uncover the identity and location of the treasure, the robot participants will be given hints to find 10 other required objects along the way, some of which may reveal additional clues.

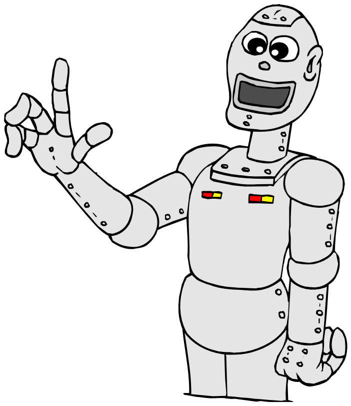
In FY'06, we plan to conduct a proof-of-concept treasure hunt to assess challenge feasibility and establish baseline capability. We propose to construct a sample house, implement a communication protocol, and collaborate with UMass Lowell and MITRE robot teams to test and refine.

Attached documentation:

Challenge Description
The $1 Million Grand Treasure Hunt
Detailed description of the challenge, rules, and specifications

Briefing (modified from version used in AAAI Presidential Address)
Single-slide overview of challenge with supporting slides in more details

FY'06 Proposal
Plans for follow-on work [6 SM]



The $1 Million
Grand Treasure Hunt

Mobile robots use navigation and visual recognition to discover objects in a real world treasure hunt

Table of Contents

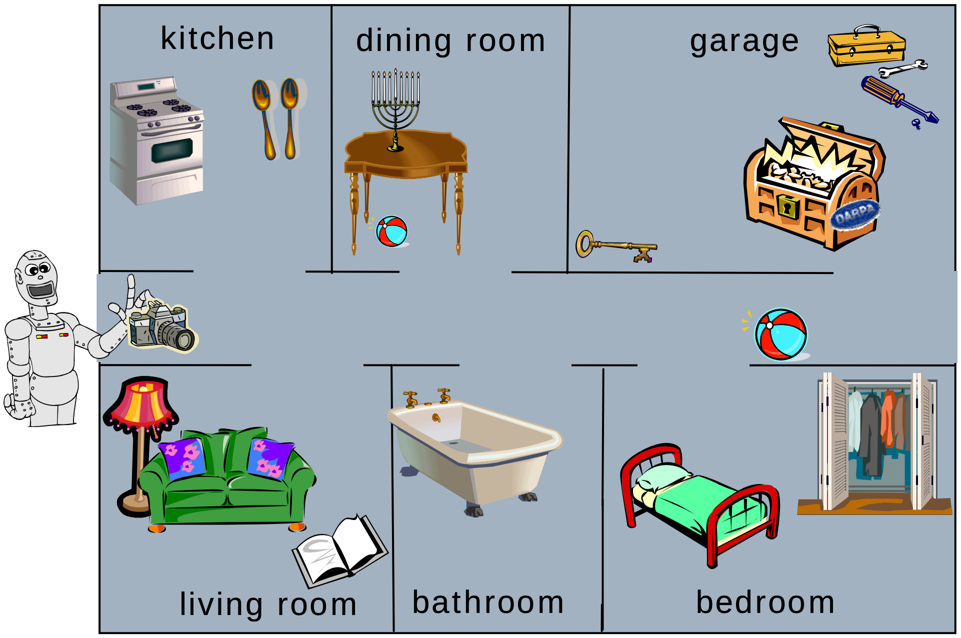


Figure 1 Sample Layout of Rooms and Objects

1 Overview

1.1 The $1 Million Grand Treasure Hunt

Find the $1 million treasure in less than 20 minutes: Participants will have 20 minutes to locate a treasure somewhere within a house. In order to uncover the identity and location of the treasure, the robot participants will be given hints to find 10 other required objects along the way, some of which may reveal additional clues.

1.2 Challenge Goal

The goal of the Grand Challenge Treasure Hunt is to foster the development of robots that are capable of demonstrating a cognitive understanding of the real world. Intended research foci include recognition of physical objects, comprehension of object function and physics, and cognitive localization. Robots will be expected to understand a heterogeneous set of descriptors about the world, which will be visual, textual, and audile. Demonstrating such versatility will be a great boon to the robotics and AI communities.

1.3 The Game Explained

1.3.1 The Goal

The object of the competition is to find a treasure inside the house. During the competition, the judges will give hints to the participants for finding additional objects that contain clues for identifying or locating the treasure. A participant's game ends when 1) all objects have been found and the robot identifies and locates the treasure, or 2) the allotted game time has been exceeded. Figure 2 provides an overview of the game as a flow diagram.

1.3.2 Prizes

The participant to have successfully identified and found the treasure, in the shortest amount of time within the game time constraints, will receive $1 million dollars in prize money. If there is no winner, a $10,000 prize will be awarded to the participant who has found the most objects and clues leading to the treasure.

1.3.3 Challenge Participation

Participants will be given a single attempt to find the treasure. Participants may be single robots or teams of robots. Treasure hunters (either single robots or teams) will participate one at a time.

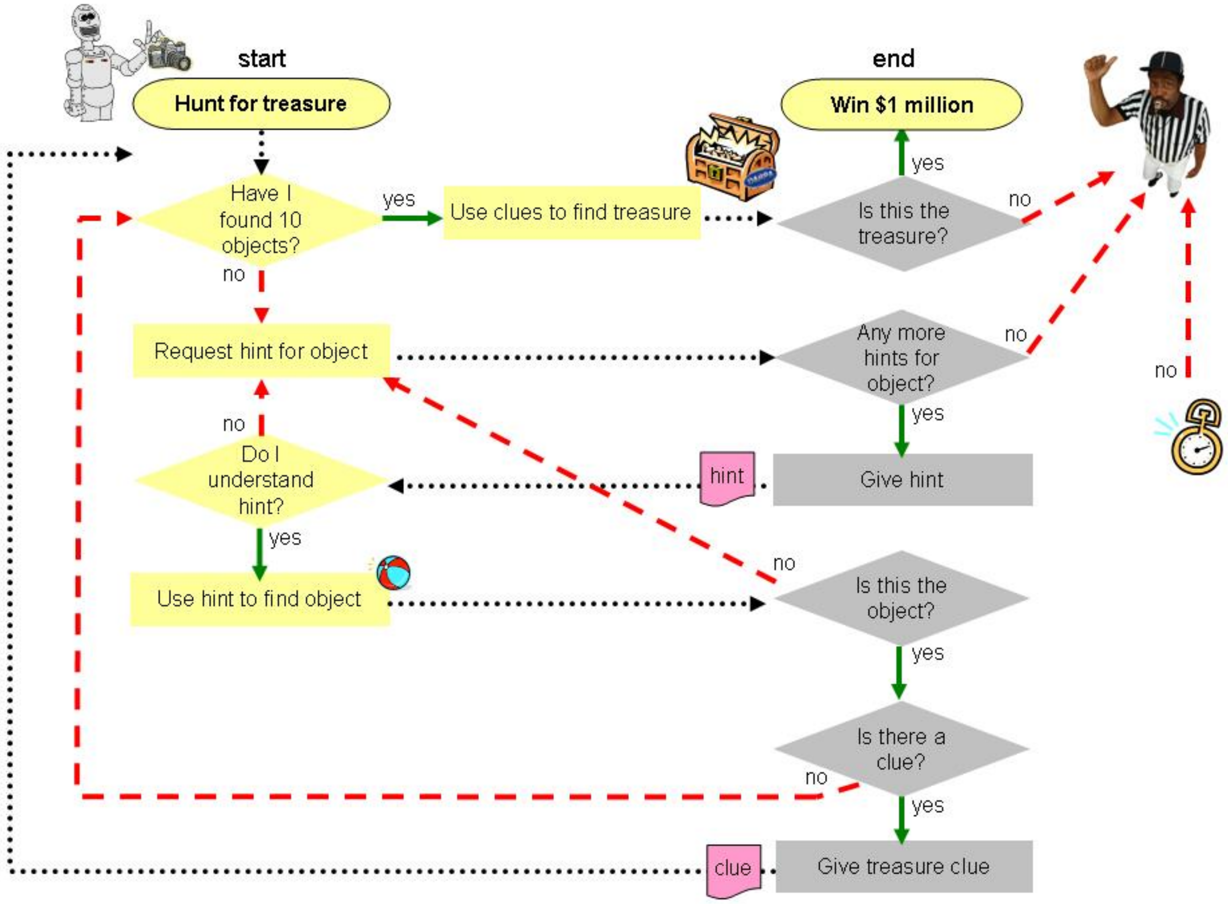


Figure 2 Overview of the Grand Treasure Hunt. Participant actions & decisions are shown in yellow. Judge action & decisions are shown in grey. Time (represented by the clock) is a pervasive decision throughout the challenge. If no participant finds the treasure, a smaller prize will be awarded to the participant that finds the most required objects.

1.4 General Rules

1.4.1 A $1 million prize will be awarded to the first participant to find the treasure within 20 minutes

1.4.2 Participants must find all 10 objects before attempting to identify the treasure.

1.4.3 Participants are given a single opportunity to identify the treasure.

1.4.4 Participants may be a single entity or a team/swarm

1.4.5 Participants (or participant teams) must operate autonomously

1.4.6 Participants may interact with human judges at any time but not with their human teams

1.4.7 None of the participants may be modified or reprogrammed once the first participant begins the competition.

1.4.8 Participants must not leave any trace of their presence (i.e., must return all objects to their original position and cause no damage)

1.4.9 Participants may use GPS

1.4.10 Participants may not access the internet during a competition

1.4.11 All objects will be placed on the floor and be no more than 90cm in height

1.4.12 Objects will be not be placed in inappropriate rooms (e.g., a stove will not be placed in a bedroom)

1.4.13 Participants must identify and find an object by submitting a photograph of the object to the judges

1.4.14 Participants may ask for additional hints to find objects

1.4.15 The only clues provided to help identify the treasure will be those earned by finding the specified objects.

1.5 Provisions

1.5.1 A list of 50 possible objects in the house will be provided in advance (e.g., cup, ball, chair, lamp, hammer)

1.5.2 No map of the environment will be provided

1.5.3 General specifications will be available (e.g., number and types of rooms, width of hallways)

2 Physical Objects

Physical objects, including the treasure, will be placed throughout the house. In each room, there will be at least one room-identifying object (e.g., a stove would identify the room as a kitchen). While no room-identifying object will be placed in an inappropriate room (e.g., a stove will not be found in the bathroom), not every object will identify a room (e.g., a ball could be found in any room, including the hallway).

All objects in the house will be selected from a list of generic objects provided in advance of the competition. A sample list is shown below.

Non Room Identifying Objects	***Kitchen* Objects**	***Living Room* Objects**	***Dining Room* Objects**	***Bathroom* Objects**	***Bedroom* Objects**	***Garage* Objects**
Mirror	Cup	Sofa	Dining table	Toothbrush	Bed	Toolbox
Chair	Spoon	Television	chandelier	Bath tub	Pillow	Rake
Ball	Knife	Bookcase		Toilet	Wardrobe	Hammer
Lamp	Plate	Books		Towel		Bicycle
Plant	Stove	Magazines		Hair dryer		Screwdriver
Telephone	Refrigerator	Coffee table		Toilet paper		Hose
						Pliers

Table 1 Sample list of objects that may be found in house. Some objects are room-identifying objects.

2.1 Location & Size

All objects will be placed on the floor. Each object position will be marked by a white "X" (5 cm in diameter). An object may be placed behind another object or barrier or partially occluded by another object or barrier, but an object will not be placed inside anything such as a drawer or a cabinet.

The height of an object is limited to 90cm.

2.2 Finding Objects & Receiving Clues

During the competition, participants prompt the judges for hints which identify objects that the robots are required to find. Any object in the house that meets all of the requirements specified by the hint will be considered correct. (There may be more than one object that matches hint criteria.)

The robot may request additional hints to help it identify the object. When the robot has located the specified object, it must transmit the appropriate message to the judge along with a photograph of the object. The object must occupy more than 50% of the photograph.

The judge will confirm or refute identification of the object. If an object has been incorrectly identified, the participant may request another hint (in a different format and/or type but same level of specificity) and continue the search.

If an object has been confirmed correct, the judge may also provide a clue leading to the treasure. There may not be a clue associated with every object. When the robot is ready to find the next object, it must prompt the judge for another hint.

2.3 Interacting with Objects

A robot may touch, move, or photograph any object as long as it is returned to its original position (within 10 cm of the center of the white “X” and rotated no more than 45 degrees from its original position) before attempting to identify the treasure.

Objects may not be altered in any way with the exception of moving or photographing. Damaging or otherwise irreparably altering an object may result in disqualification

3 Objects & Hints

Judges will instruct robots to find objects during the game by giving hints. All hints will be transmitted with associated metadata indicating format and level of specificity. A participant may request that the same hint be sent in a different format up to three times. Each time an additional hint is provided, it will be presented in a different format. The order in which hint formats are presented will vary from participant to participant for each object, but all participants will be exposed to the same hint formats by the end of the competition.

3.1 Level of Specificity

A hint can be generic (e.g., find any object like this one) or specific (e.g., find the object exactly like this one).

3.2 Formats

Hints may be presented in any one of the following formats:

Text: ASCII code containing no more than 50 words of American English.

Audio: WAV File sampled at 8khz using 8bit quantization. The maximum length of an audio hint will be 15sec. Hints will be recorded by native male speakers of American English.

Photograph: JPG file of the object against a contrasting neutral background. The object will be photographed so that it takes up as much of the photograph as possible and oriented in the way which best illustrates important or identifying characteristics of the object.

Replica: An example of the object to be found will be placed on the floor in front of the participant. (Prior to the start of the competition, the teams may specify a particular sensor the judge should use when presenting a replica to the participant robot.) The object will be placed in front of the robot for approximately ten (10) seconds but the robot may leave at any time to begin searching.

3.3 Types

Hints may be one of eight (8) types as shown in the table below. For each object in the treasure hunt, there will be only four (4) possible hint types available – either a set of generic hints or a set of specific hints.

The table below presents all allowable hints types. Each hint type has an associated identification code used when transmitting the hint to the robot. For purposes of illustration, the specified object is a red baseball located under a table in the kitchen.

ID	Specificity	Format	Explanation	Example
GT	Generic	Text	Name of the generic object sent in text form	Ball
GW		WAV file	WAV file of the spoken generic object name	"*Ball*"
GJ		JPEG	Color photograph of an example of the object against a blank background	Picture of an orange basketball
GR		Replica	Physical example of a similar object	An orange basketball
ST	Specific	Text	Name of the object followed by a variable number of descriptors (specifying location, relational position, color, size, etc.)	Ball – Red - Kitchen
SW		WAV file	Spoken name of the object followed by a variable number of descriptors	"*Ball – under table*"
SJ		JPEG	Color photograph of the specific object to be found	Picture of a red baseball (kitchen and table are not visible)
SR		Replica	Physical example of an object which is as similar to the actual object as possible	A red baseball replica

Table 2 Allowable Clue Types and their Associated IDs, Levels of Specificity, and Format

There may be multiple examples of objects in the house. For example, there may be multiple balls of different colors. If the hint is a generic type, any ball is considered acceptable. If the hint is a specific type, only those meeting all of the criteria will be considered correct. For example if the hint is (ST) "ball - red", a green ball is not correct.

Participants must be able to filter out irrelevant information in hints. For example, a generic hint like a picture of a ball (GJ) may include some color information, e.g., *a red ball*. The assumption is that the participant must find the object (*ball*) and not the color (*red*).

3.4 Descriptors

There may be one or more descriptors associated with a specific hint. Examples of descriptors are color, location, relational position, and size. Location may take forms such as "the ball in the kitchen" or "the ball under the table". The hint "ball – red - white – in -

kitchen – under – table" defines the object to be found as a red and white ball under the table in the kitchen.

The selection of acceptable descriptor types and syntax is deferred to the first challenge steering committee.

It is important to realize that descriptors such as "under the table" contain the need to understand "under" and know what a "table" is. Finding the red ball under the table is actually finding the *specific* red ball in variable (under, next to, to the right of …) relation to a generic table.

Descriptors are only relevant for Specific-Text (ST) and Specific-WAV (SW). When a picture of the object is transmitted (SJ) or when an example is shown (SR), there is no attempt to define which characteristics of the object shown are important in finding a match.

3.5 Ordering/Randomness

In an attempt to avoid a bias in favor of participants capable of 'understanding' some hint formats better than others, the order in which hint types will be presented for each object will vary from object to object for each participant.

Prior to the start of the competition, ten (10) of the forty-eight (48) possible orderings of hints will be selected. The judges will randomly choose, without replacement, from these ten hint sequences each time a new object is to be described.

For example if the hint order for object number three is {GW, GR, GJ, GT} then the first hint sent to the robot for object number three will be a WAV file with the word "ball" spoken. If a new hint is requested, the judge will present a ball to the participant. If required, the next hint would be a JPG of a ball and the last hint would be a text file with the word "ball". The next robot would have access to the same hints but they may be presented in the order {GR, GT, GJ, GW}. The {GW, GR, GJ, GT} order would be encountered by this next robot when receiving hints for a different object.

4 The Treasure & Clues

The treasure can be any object within the house and will be included in the object list provided in advance. For example, the treasure may be a rare vase, a box full of old coins, a valuable painting, etc. The treasure will not be hidden or placed inside another object, but its identity and location will only be revealed over time as required objects are found. When an object has been successfully located, the judge may reveal a clue leading to the treasure. Not all objects will have clues. Each clue will implicitly reveal just one dimension of the treasure, .e.g., color (C), shape (S), location (L), or object type (T), by referring to one of the required objects that may or may not yet have been found. For example, instead of revealing that the color of the treasure is red, the clue might refer to the color of object #4.

Clues will be transmitted to the participant via the same mechanism as hints. Clues will always be in text format.

The table below illustrates an example of hints, objects, and clues leading up to the identity and location of a treasure. In this case, the treasure is a rare, red book in the living room.

Object #	Object Hint	Object	Treasure Clue
1	cup	Any Cup	**L-2**: Same room as object #2
2	Sofa – green	Sofa in Living Room	*Sorry – no clue provided*
3	Ball	Any ball	**C-4**:includes color of #4
4	Toothbrush	Red Toothbrush in bathroom	*Sorry – no clue provided*
5	Bed	Any bed	*Sorry – no clue provided*
6	Bicycle – picture	Brown bicycle in garage	*Sorry – no clue provided*
7	Hose	Any hose	*Sorry – no clue provided*
8	Lamp	Green Lamp	*Sorry – no clue provided*
9	Television	Television	**T-10**: Same type of object as #10
10	Book - orange	Orange book	*Sorry – no clue provided*

Table 3 Example of Hints, Objects, Clues Used to Find a Treasure (red book in living room)

5 Communication

All participating robots must have the ability to receive and respond to commands from the judges. Participants must also be able to transmit a message indicating they have found an object.

The goal of this competition is not to be able to deal with harsh communication scenarios; participants may assume they will have perfect communication at all times.

Prior to the competition, a more detailed communication spec will be provided to participants, including detailed examples, sample code, etc.

5.1 Hardware Logistics

Robots must be able to communicate wirelessly to the judges via 802.11g. Base stations will be set up inside the competition space so that all areas within the space will provide sufficient signal strength. A unique channel is guaranteed to be available to participants while running the competition to avoid packet collisions with other networks.

5.2 Communications Protocol

All communication is text, unless otherwise specified.

From	To	Message	Description
Judge	Robot	BEGIN	Instruction to start game. The robot may not begin moving before it receives this instruction.
Judge	Robot	END	Instruction to end game. The robot must stop moving and sensing immediately.
Judge	Robot	STOP	Command to stop moving. Note, this is different from END; it simply requires the robot to cease movement.
Judge	Robot	NEED HELP	Request for robot status. Robot should respond promptly.
Robot	Judge	FINE/STUCK	Response to status request. Indicates either that the robot is OK or is stuck and needs judge intervention.
Robot	Judge	GIVE_HINT *<object #>*	Request for object hint. This command can be used to request the first hint for an object as well as additional hints for the same object.
Judge	Robot	HINT *<object #>* *<hint ID>* *<hint-start>* … *<hint-end>*	Each hint contains the following information. 1. Associated object # 2. Hint Format ID 3. A *hint-start* indicator 4. The hint 5. A *hint-end* indicator
Robot	Judge	FOUND_OBJECT *<object #>* *<object-start>* … *<object-end>*	The robot sends this message when it has found a specified object. Message contents are: 1. Object # 2. An *object-start* indicator 3. Jpeg of the object 4. An object-*end* indicator
Judge	Robot	OBJECT_CONFIRMATION *<object #>* *<confirmation>*	Confirmation of correct identification of object: 1. Object # 2. Confirmation {Yes, No}

Judge	Robot	CLUE *<clue ID>* *<dimension>* *<object #>*	Each clue contains the following information. 1. A unique clue ID 2. Dimension {C, S, L, T} (one of color, shape, location, type) 3. Referenced Object #
Robot	Judge	FOUND_TREASURE *<object-start>* … *<object-end>*	The robot may send this message only once during the game, when it has found the treasure. Message contents are: 1. An *object-start* indicator 2. Jpeg of the object 3. An object-*end* indicator
		TREASURE_CONFIRMATION *<confirmation>*	1. Confirmation {Yes, No}

Table 4 Communications Protocol for Robot / Human Judge Dialogue

6 Disqualifications

A participant team will be disqualified, at judge discretion, for any of the following:

6.1.1 Persistent infringement of the rules of the game

6.1.2 Damage to the house, irreparable damage to any object in the house, or harm to another robot

6.1.3 Harm to judge

6.1.4 Interaction with human team mates during competition

6.1.5 Looking over walls

7 Robot Constraints

7.1 Autonomy

The robotic equipment is to be fully autonomous. Human operators are not permitted to enter any information into the equipment during a search.

7.2 Construction

7.2.1 Restrictions

There are no physical size restrictions beyond those imposed by the minimum dimensions of the competition space..

Robot wheels (or other surfaces which contact the floor or walls of the house) must be made of a material which does not harm or mark the contacted surface.

A robot must not have, in its construction, anything that is dangerous to itself, a human, another robot, or the competition environment. At the judges' discretion, a robot may be removed from the course and prevented from participation.

7.2.2 Sensors

There is no limit on the type or number of sensors that a robot may employ, but all sensing and processing must be performed onboard.

All sensors must be mounted less than 90cm from the floor. Robots will be disqualified for "looking" over the tops of walls.

7.3 Communication

Robots must be capable of two-way wireless communication as described in the section on Communications.

8 Competition Space & Environment Requirements

The competition space will be a single-story house consisting of 6 or more rooms, a central hallway, no stairs and no doors. Information about the actual number and type of rooms, layout, and placement of entry ways will not be provided in advance. No map of the structure will be provided.

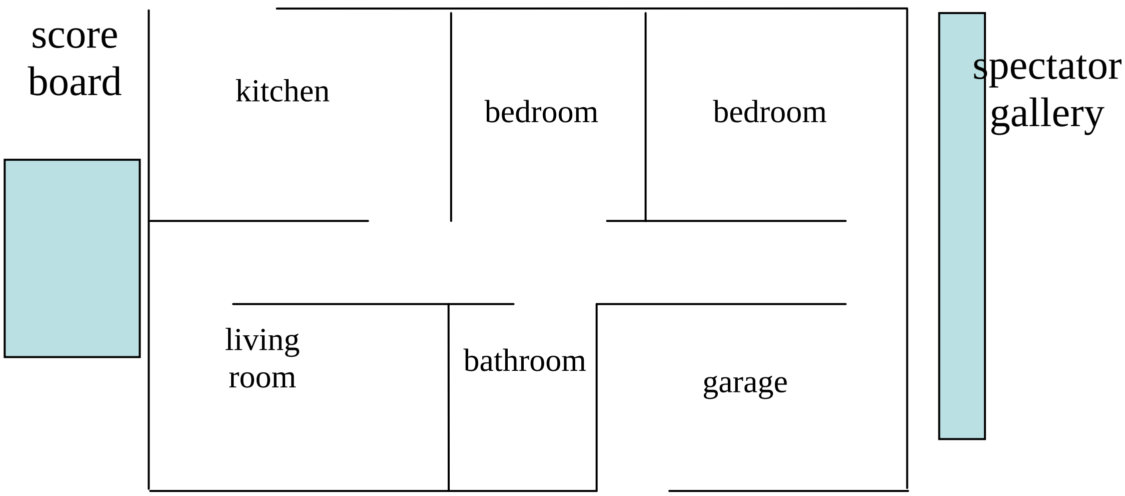


Figure 1 Sample House Layout

8.1 Rooms

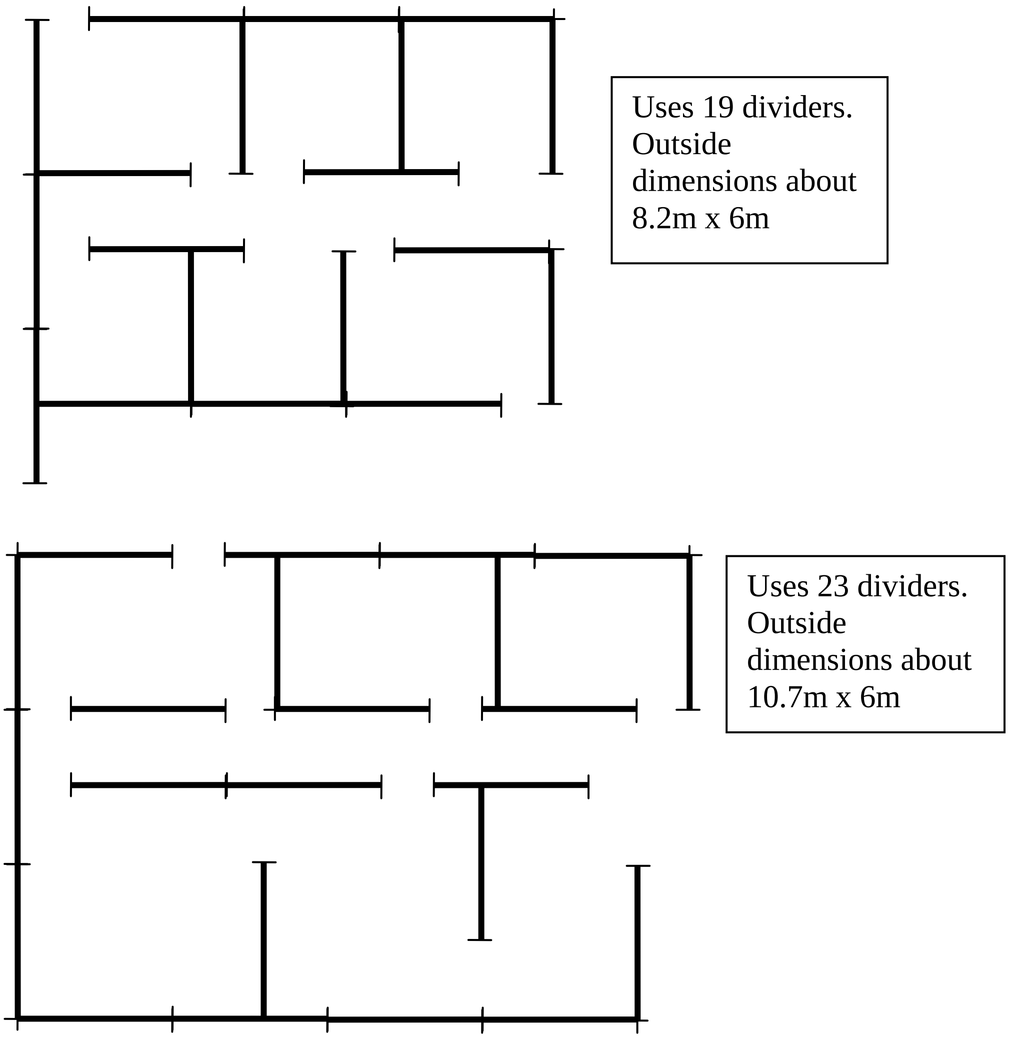
The competition space will consist of six (6) or more separate rooms connected by a central hallway. Rooms will include one kitchen, bathroom, bedroom, living room, garage, and optional dining room. There may be multiple bathrooms or bedrooms, but no more than one of the other room types.

Each room will have some identifying characteristic to distinguish it from another room. (E.g., a stove in a kitchen). When the presence of an object in a room is considered sufficient to uniquely define the type of room, this information will be indicated on the list of possible objects distributed prior to the competition.

Each room will be a minimum of 240cm x 240cm. Larger room sizes are permitted.

A single, central hallway will connect all rooms.

The reference diagrams below show two examples of structures which comply with these requirements built with dividers (├──┤) made from standard 4’ x 8’ sheets of plywood cut in half.



Uses 19 dividers. Outside dimensions about 8.2m x 6m

Uses 23 dividers. Outside dimensions about 10.7m x 6m

Figure 2 Two Sample House Constructions

8.2 Access

Access to the house or rooms will consist of openings in walls, which are not blocked in any way, and have no coverings such as doors. Each room must be accessible from the central hallway. Additional access between rooms is permitted but not required. The placement of the room openings is not fixed and is at the discretion of the competition host.

All openings will be at least 80cm wide.

The opening at the front of the house must provide access to the central hallway in the structure. Additional external access points are allowed but not required.

8.3 Hallway

There will be a single hallway, centrally located from the front of the house to the back of the house

The hallway will be at least 100cm wide.

8.4 Walls

Walls will be at least 120cm tall.

There is no constraint on the color or material of walls. For example options, at the discretion of the competition host, include walls which are opaque, transparent or semi-transparent, reflective or sound-absorbing.

Items (including reflective mirrors) may be hung on walls but their presence must not cause a violation in minimum clearance rules.

8.5 Floor Coverings

Floors will be smooth.

The floor material is not specified except that carpet may not be used.

There is no constraint on the patterns or colors of the floor covering.

8.6 Obstructions & Complicating Factors

Within the house, robots can expect to encounter physical objects in their path as well as complications such as humans, bright lights, intentionally shaded areas and sound sources.

Physical objects placed in rooms or in the hallway must leave a path around the obstruction which is at least 100cm wide.

Judges are allowed to enter and move about the competition space when rescuing a stuck robot or when placing a replica type hint. Robots may not touch, move, collide with, or in any way harm a human judge.

At the discretion of the competition host, various levels of lighting brightness and positioning relative to objects and rooms are permitted. Sound sources may also be positioned throughout the competition space.

8.7 Scoreboard & Monitor

A scoreboard and monitor will provide spectators with a live video display of the house, the communication between the robot and the judges and the current scores in the competition. The scoreboard may optionally display a live video feed from the robot during the competition.

9 Judges

Two judges will oversee the competition: the coordinator and the referee. The coordinating judge is responsible for communicating with the robot. The referee will be physically observing the robot, assuring compliance with the rules, and monitoring the clock.

The coordinating judge will transmit clues and evaluate the participants' responses. If the robot becomes "stuck" and requires intervention, the referee will assist in returning the robot to the starting point.

Both judges will be responsible for returning all objects to their original location and orientation prior to each treasure hunt.

Judges are allowed to enter and move about the competition space when rescuing a stuck robot or when placing a replica type hint. However, in general, the judges should not interfere with the robot's sensors.

Judges are not liable for injury suffered by spectators, robot participants, or their human teams. They are also not liable for any property damage.

10 Targeted Technology

10.1 Perception

10.1.1 Generic object recognition (e.g., identify a real ball in a real environment)

10.1.2 Image understanding (e.g., identify a ball in a photo)

10.2 Reasoning & Inference

10.2.1 Associating real objects to generic names (e.g., a physical instance of a ball is what is meant by the term 'ball')

10.2.2 Understanding context (knowing where certain objects can be found, e.g., a stove would be found in the kitchen)

10.3 Learning

10.3.1 Creating internal map of physical environment (e.g., house floor plan)

10.3.2 Associating objects with their locations (e.g., 'remembering' that a tool box was in the garage)

10.3.3 Learning from experience (e.g., learning not to repeat a mistake twice, or remembering a shortcut into the bathroom from the bedroom)

10.4 Knowledge Representation

10.4.1 Associating objects to their attributes and functionality

10.5 Innovation

10.5.1 Discovering multiple ways to solve problems (example left to future robot participants)

11 Acknowledgments

This challenge was created in support of DARPA IPTO's effort to produce Grand Challenges in Cognitive Science.

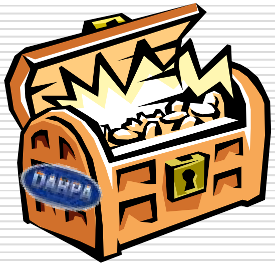
Contributor	**Affiliation**	**Email**
Ronald Brachman	DARPA IPTO	
David Gunning	DARPA IPTO	dgunning@darpa.mil
Rodney Brooks	CSAIL MIT	brooks@csail.mit.edu
Laurie Damianos	The MITRE Corporation	laurie@mitre.org
Randall Fish	The MITRE Corporation	fishr@mitre.org
Michael Littman	Rutgers University	mlittman@cs.rutgers.edu
Laurel Riek	The MITRE Corporation	laurel@mitre.org
Manuela Veloso	Carnegie Mellon University	veloso@cs.cmu.edu

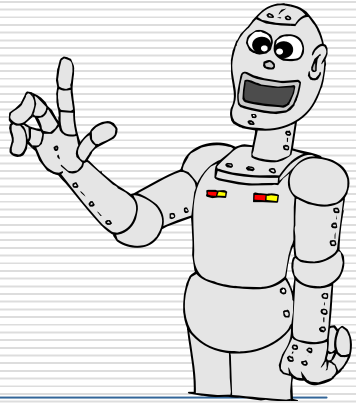
12 Appendix

12.1 Dimensions of Representative Robots

Manufacturer	Robot name	Width (cm)	Height (cm)	Length (cm)
Sony	AIBO	18	28	32
ActivMedia	PowerBot	63	47	84
	PeopleBot	38	112	47
	Pioneer3-DX	40	24	45
	Amigo-Bot	28	15	33
K-Team	Khepera	7	11	7
	Hemisson	12	variable	12
	Koala	32	20	32
Andros	F6A	72	127	127
iRobot	PackBot	41	8	69

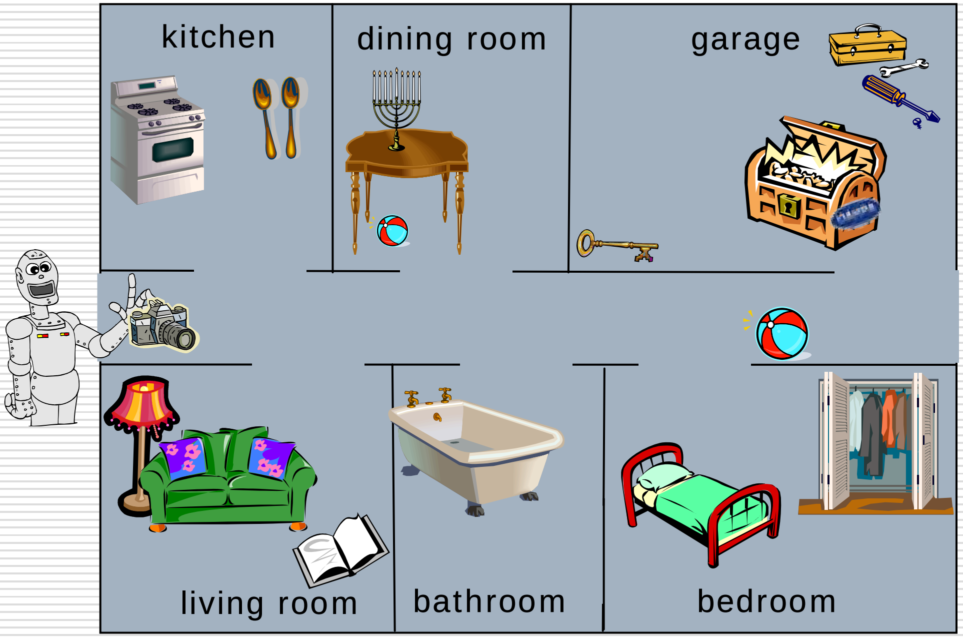
Table 5 Dimensions of Representative Robots





The DARPA
Grand Treasure Hunt Game

Mobile robots use navigation and visual recognition to uncover treasure in real world treasure hunt game.



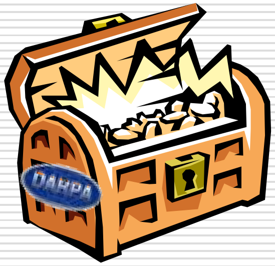
Find the $1 million treasure in less than 20 minutes.

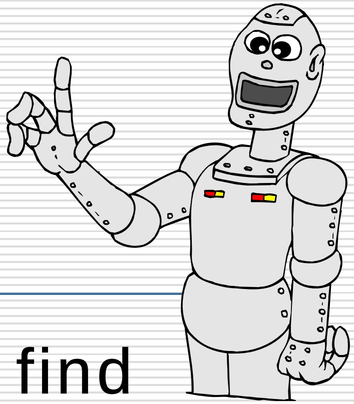
But… in order to uncover the identity and location of the treasure, the participants must ask for hints to find 10 other required objects along the way, some of which may reveal additional clues…

Of course, you could always ask the judges for more hints, but that takes more time…

**To “find” an object, robot must submit a photograph to panel of judges. Judges will impose any preconditions and assess if any violations have occurred.*

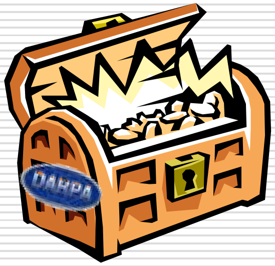
The Game

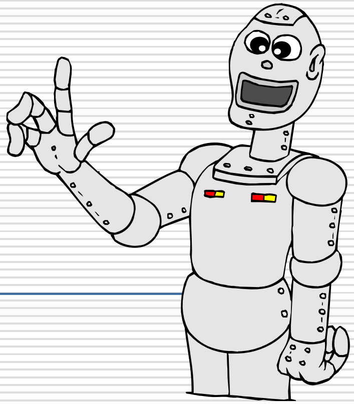




- Robots have 20 minutes to identify and find the location of a treasure somewhere within a single-story house
- The treasure can be any object, e.g., a rare vase, a box of old coins, a valuable painting
- Judges give the robots hints to find objects along the way that may reveal additional clues to the treasure
- Finding an object involves submitting a photograph of the object to a panel of judges by wireless communication
- Robots must operate autonomously
- Participants are encouraged to be innovative

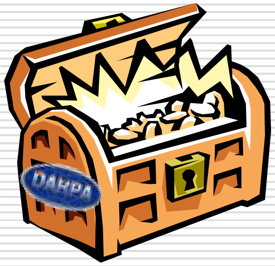
The Rules

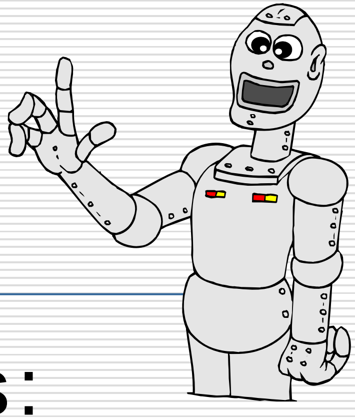




- No map of the environment will be provided
- General specs will be available (e.g., number & types of rooms, width of hallways)
- A list of 50 possible objects in the house will be provided in advance (e.g., cup, ball, chair, lamp, hammer)
- All objects will be placed on the floor and be no more than 3 feet in height
- Objects will not be placed in inappropriate rooms (e.g., a stove will not be placed in bedroom)
- Participants may:
  - Be a single entity or a team/swarm
  - Use GPS and other public signals
  - Access the static web
- Participants must:
  - Operate autonomously
  - Leave no trace of their presence
- The first robot to find the treasure within the allotted time wins!

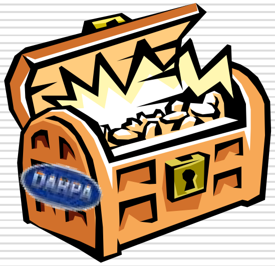
The Hints

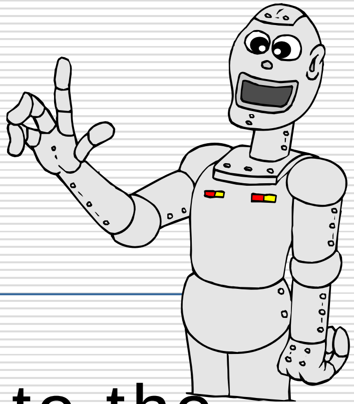




- Hints can be of multiple formats: text, audio, photograph, replica
- Hints can indicate specific objects or generic instances
- Robots may ask for additional hints, each hint being of a different format
- Hint formats will be randomly ordered
- Examples: *ball, ball under table, red ball in kitchen*

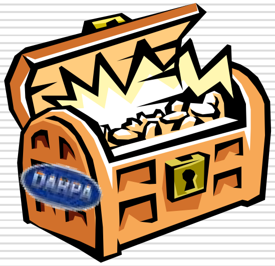
The Clues

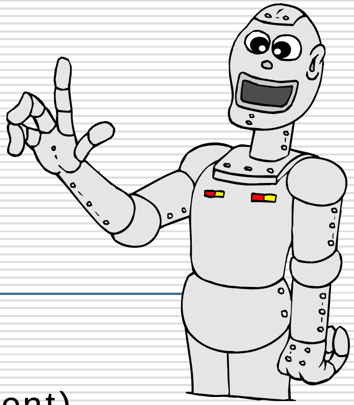




- When a robot finds an object, a clue to the treasure may be revealed
- Each clue provides a dimension of the treasure: object, color, shape, size, location
- Clues are implicit in that they refer to dimensions of other objects not yet found
- Examples: *the same color as object #4*

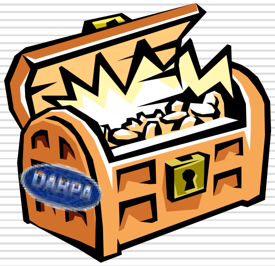
The Cognition

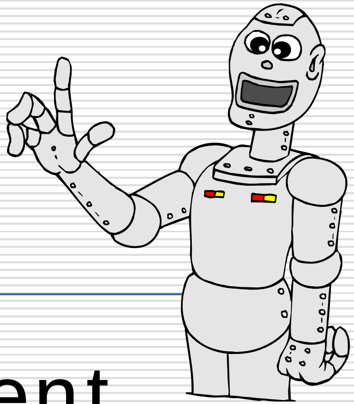




- Perception
  - Generic object recognition (e.g., identify a real ball in real environment)
  - Image understanding (e.g., identify a ball in a photo)
- Communication
  - Human-computer interaction (e.g., dialogue with judges & human team mates)
- Reasoning/Inference
  - Associating real objects to generic names (e.g., a physical instance of a ball is what is meant by the term “ball”)
  - Understanding context (knowing where certain objects are likely to be found, e.g., a stove would be found in a kitchen)
- Learning
  - Creating internal map of physical environment (e.g., house floor plan)
  - Associating objects with their locations (e.g., “remembering” that a tool box was in the garage)
  - Learning from experience (e.g., learning not to repeat a mistake twice, or remembering a shortcut into the bathroom from the bedroom)
- Knowledge Representation
  - Associating objects to their attributes and functionality
- Innovation
  - Discovering multiple ways to solve problem (example left to future robots)

The Credits



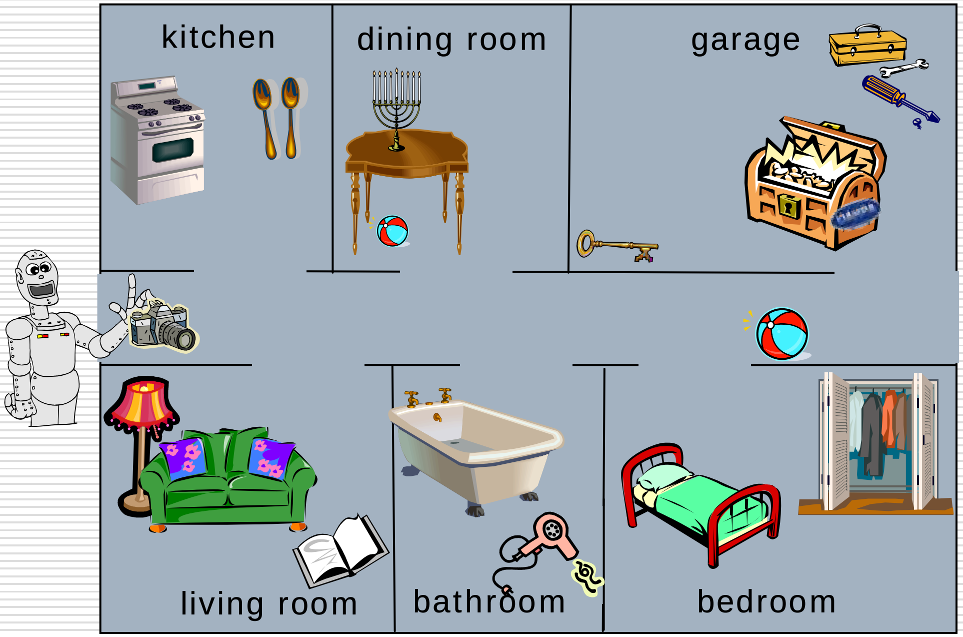


- Brainstorming & Initial Refinement Team
  - Michael Littman, Rutgers (Technical Lead)
  - Rodney Brooks, MIT CSAIL
  - Manuela Veloso, CMU
  - Laurie Damianos, MITRE (Support)
  - Randy Fish, MITRE (Support)
  - Laurel Riek, MITRE (Support)

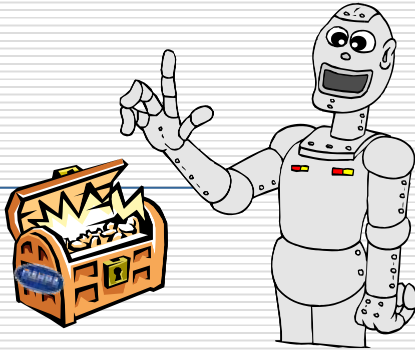
Grand Treasure Hunt Challenge

Proof-of-concept Treasure Hunt

- Assess challenge feasibility
- Establish baseline capability
- Scope interest



FY'06
Proposal



- Pre-Trial
  - Construct sample house
  - Implement communication protocol
- Pilot
  - Collaborate with local robotics teams (UMass Lowell, MITRE)
  - Test & refine
    - Robot/judge interaction
    - Clue format & transmission
    - Clue/object sequences
    - Evaluation: scoring algorithm
    - Room navigability
  - Modify & iterate
- Post-trial
  - Conduct survey of likely participants to assess where current research fits into challenge
  - Insert appropriate technological challenges into Treasure Hunt
- MITRE LOE: 6 SM

Test Taker
Taking the SAT

An autonomous system will take the SAT® and score in the 50th percentile of high school students taking the examination.

For this Grand Challenge, the winning system must take the SAT and score in the 50th percentile of high school students taking the examination the same year. Our investigations revealed that the Math, Reading, and Writing sections of the test require many of the capabilities IPTO seeks in a cognitive system.

For follow-on work, the next step is to meet with College Board and ETS to stimulate their support of this challenge, and then obtain sample tests and build an API and computer-readable test format.

Attached documentation:

Challenge Description
Taking the SAT Grand Challenge
Detailed description of the challenge, rules, and specifications

Supplemental Report
Report to DARPA IPTO on the SAT as a Grand Challenge
Results of MITRE's investigations into the merits of the SAT challenge

Supplemental Talking Points
Why Should College Board & ETS Support a Grand Cognitive Challenge Aimed at Building an AI Able to Ace the SAT?
Talking points for gaining the support of the test-makers

Briefing (modified from version used in AAAI Presidential Address)
Single-slide overview of challenge with supporting slides in more details

FY'06 Proposal
Plans for follow-on work [3 SM]

Taking the SAT®
Grand Challenge

Table of Contents

1 Overview

An autonomous system will take the SAT® and score in the 50th percentile of high school students taking the examination. The system must take the same test that is administered to high school students and observe the same rules as humans regarding help, access to external resources (no internet connection), and time constraints. The only significant difference between the human and computer versions of the test is that the computer version will be administered in a computer readable format, including diagrams and formulas.

The goal of this challenge is to foster the development of computer systems that are able to reason and communicate effectively in an open domain. The SAT is designed to test such abilities, and so makes an ideal target for advanced computer systems to aspire to.

1.1 Targeted Technology

The Taking the SAT Grand Challenge demands complex, integrated abilities in a cognitive system, focusing on two key areas: [1]

- Reasoning and Inference
  - Reason on the basis of text and diagrams (math sections)
  - Understand and take into account human perspectives and emotions (passage-based reading sections)
  - Draw conclusions based solely on the evidence presented (math and reading sections)
- Symbolic Communication
  - Understand and follow directions (math, reading, and writing sections)
  - Detect subtle textual clues (critical reading sections)
  - Develop and express ideas effectively (writing sections)

We anticipate that some of the test is within the capabilities of state-of-the-art technology.

2 Test Contents

Systems will be required to take the same SAT exam administered to human students. Participants are encouraged to learn more about the SAT directly from The College Board, but we present here a brief overview of the test contents.[2] This overview is presented merely as an introduction to the Grand Challenge, and it is in no way intended to limit what may be presented in the actual examination.

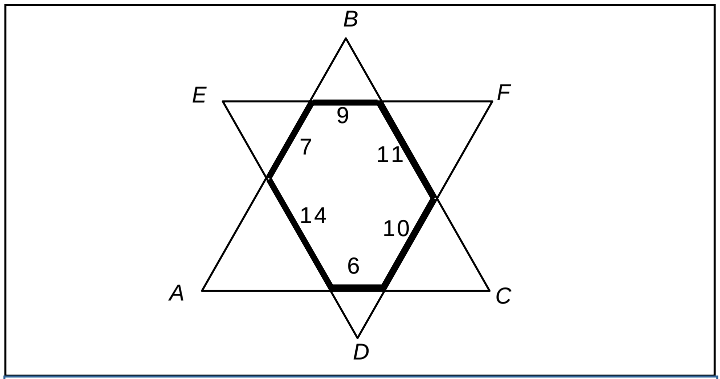
The new SAT® consists of three sections: Math, Critical Reading, and Writing.

[1] This assumes the system would be taking the new SAT Reasoning Test™. The older SAT would target a slightly different set of technologies.

[2] The following is from the *Official SAT Study Guide for the New SAT™*, College Board, 2004.

Math. The Math section contains mostly multiple choice questions, but there is also a small set requiring the test-takers to write in their answers. Formulas that are required to answer the questions are provided within the test materials themselves (e.g., $C = 2\pi r$, $V=\pi r^2 h$ etc.). The areas of mathematics covered by the exam are number operations, algebra, functions, geometry, measurement, data analysis, statistics, and probability. Questions are posed in a variety of ways, some involving little language, and some heavily language dependent. In addition to the question and the list of answer choices, many items also reference one or more mathematical statements, diagrams, tables, and charts. Diagrams and other non-textual data can appear in the question or the answer choices.

Here is a sample test item; in this particular case, the test-taker must write in the response:



In the figure above, equilateral triangles *ABC* and *DEF* intersect so that side *AB* is parallel to side *DF*. The numbers indicate the lengths of the sides of the polygon outlined in bold. How much greater is the perimeter of triangle *ABC* than the perimeter of *DEF*? (PSAT 2004)

Critical Reading. This section contains two questions types:

- *Sentence Completion* tests vocabulary and understanding the logic of complex sentences. Sample test item:

> In public, Henry was somewhat ------- toward his opponents; behind their backs, he was even more -------.
> (A) sympathetic..furious (B) amicable..disparaging (C) caustic..vitriolic (D) bitter..patronizing (E) imperious..unctuous (PSAT 2004)

- *Passage-based Reading* tests reasoning and inference, comprehension, and vocabulary in context. Passages vary in length and some items require reading and answering questions about pairs of passages. Sample test item:

Passage 1: "…Obviously, having a coelacanth in a tank would make someone a lot of money…."

Passage 2: " 'We need a live coelacanth in captivity,' said Mike Bruton from his base at the Two Oceans Aquarium…"

The comment in Passage 1 about "someone" implies which of the following about the Two Oceans Aquarium mentioned in Passage 2?

(A) It has plans to support coelacanth conservation programs.
(B) It could benefit financially from displaying a live coelacanth.
(C) It has great expertise in simulating the coelacanth's habitat.
(D) It might provide scientists invaluable access to live coelacanths.
(E) It would be the first institution to breed coelacanths in captivity.
(PSAT 2004)

Writing. This section contains four question types:

- *Identifying Sentence Errors* tests the ability to find mistakes in grammar, usage, and word choice.
- *Improving Sentences* and *Improving Paragraphs* test the ability to recognize and produce clear and effective writing.
- *The Essay* tests the ability to develop and support a viewpoint. Sample essay question: [3]

Think carefully about the issue presented in the following excerpt and the assignment below.

> Some people believe that there is only one foolproof plan, perfect solution, or correct interpretation. But nothing is ever that simple. For better or for worse, for every so-called final answer there is another way of seeing things. There is always a "however."

Assignment: Is there always another explanation or another point of view? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

3 Detailed Grand Challenge Rules

3.1 Test Format

The test will be presented as an XML form. Diagrams will be in GIF format and formulas will use MathML (http://www.w3.org/Math/).

[3] This sample essay question is from the *Official SAT Study Guide for the New SAT™*, College Board, 2004.

3.2 System Constraints

Each system must be containable on a single laptop computer, using any operating system. Each system can store as much data as will fit on the laptop's internal storage devices, and the data can be consulted during the test itself. Calculators are permitted by human students and equivalent facilities are permitted by the systems. The system is not permitted to store in memory any data from the test, including computations and processes performed during the test.[4] The system is not permitted to access the Internet nor is it permitted to communicate with anyone other than the Test Server (defined below). The hardware cannot be equipped with wireless connectivity devices.

3.3 Test Administration Day

Participants will arrive at a central site on the designated test day, with their systems on laptops. The test will proceed as follows.

Establishing Contact with the Test Server. The test administrators will establish a server, which we will call the Test Server to parallel the Test Supervisor who administers the SAT to humans. Participants will connect their systems to a local network and access the Test Server via an API that permits each system to communicate with the server (i.e., receive instructions and send notices, view the test items, and return answers). The network specifications and API will be made available to participants 90 days prior to test day. Participants will be given two hours to establish and test the server connection, which will include a brief practice SAT to verify system-server interoperability.

Timing. Five minutes before the test begins, the human participants will be required to vacate the test room containing the system laptops. Systems will have 3 hours and 45 minutes to complete the test. There are ten separately timed sections[5]:

One 25-minute essay
Six other 25-minute sections
Two 20-minute sections
One 10 minute section

The order of the sections varies and will not be announced in advance. One of the 25 minute segments of the test does not count toward the final score, but the test-taker does not know which one it is. This is the standard practice used by the test makers for calibrating against other editions of the SAT.

Communication and Recovery Efforts. At the start of the test, the Test Server will send a message to each participant "You may now begin the test. You have *n* minutes to complete this section of the test." (where *n* is determined based on the section being presented). When each system has completed the current test section and wishes to do no

[4] This is to protect test security, and mirrors the constraint on human test-takers, who are not allowed to use calculators with memory and are allowed to write only in their test booklets, which they are not allowed to take with them. If the PSAT is used for the Grand Challenge rather than the SAT, this restriction might be relaxed. Human test-takers can keep their PSAT booklets.

[5] This schedule applies to the new SAT®, and is the same as the one used for humans.

more work, it will send a signal to the server: “Bits Down” (the computer’s version of “Pencils Down”). If all systems signal their readiness to continue prior to the end of the scheduled test period, the Test Server will initiate the next section of the test. If one or more systems use the entire test period, the session will run to completion. At the end of the scheduled duration, the Test Server will cease accepting inputs and send a message, “Bits Down.” The Test Server will initiate the next section by announcing: “You have *n* minutes to complete the next section of the test.” Systems will be required to move on to the next portion of the test with no human intervention. After the final “Bits Down,” the Test Server will issue a “Test Completed” message. All of the Test Server’s messages will be visible on a monitor inside the test room and on monitors in the human’s waiting room.

There will be a short break at the end of each hour of testing time, which is the procedure followed with human test-takers. During these rest breaks, human participants will be permitted back into the test room only for the purpose of verifying that their systems are still running, and restarting crashed systems if necessary. The system must be back up running and reconnected to the server in the time allotted, or it will be disqualified from taking the next portion of the text. Human participants may try restarting their system during the next scheduled rest period.

3.4 Scoring

Answers will be scored by the College Board following standard procedures. For multiple choice questions, one point is awarded for each correct answer. For each attempted but incorrect answer ¼ of a point is subtracted from the total number of correct answers. No points are added or deducted for unanswered questions. For the student-produced answers in the Math section, no points are deducted for wrong answers. The essay is evaluated by two independent human judges, who can each assign a score anywhere from 0-6, for a maximum of 12 points combined. The raw scores are then equated to a scale ranging from 200 to 800 in order to adjust for minor differences between test forms.

4 Prizes

One million U.S. dollars will be awarded to the system that scores in the 50th percentile on the entire SAT, as compared to human students taking the same exam that year. For systems that do not score in the 50th percentile overall but do achieve the 50th percentile on the Critical Reading Section, the Writing Section, or the Math Section, $250,000 U.S. dollars will be awarded for each section in which the system scores in the 50th percentile.[6]

5 Practice Tests

One or more sample tests, with answer keys, will be provided to participants. As with the human practice tests, the format of the questions and the directions for each will remain the same across the practice and official tests.

[6] These $1M and $250K prize amounts are only suggestions. Further study and discussion with DARPA is needed.

6 Acknowledgements

This challenge was created in support of DARPA IPTO's effort to produce Grand Challenges in Cognitive Science.

Contributor	Affiliation	Email
Ronald Brachman		rbrachman@darpa.mil
David Gunning	DARPA IPTO	dgunning@darpa.mil
Selmer Bringsjord	Rensselaer Polytechnic Institute	selmer@rpi.edu
Michael Genesereth	Stanford University	genesereth@standford.edu
Lynette Hirschman	The MITRE Corporation	lynette@mitre.org
Lisa Ferro	The MITRE Corporation	lferro@mitre.org

Report to DARPA IPTO on the SAT$^{®}$ as a Grand Challenge

The MITRE Corporation

September 2005

Table of Contents

Overview

On behalf of DARPA IPTO, the MITRE Corporation investigated the potential of the SAT® as a Grand Challenge for cognitive systems. Our goal was to determine the feasibility of the task and whether it would meet DARPA's need to foster research and development in this area. The following report presents details of our findings on the merits of the test. A draft specification for running the actual challenge is contained in a separate document, "Taking the SAT® Grand Challenge."

Summary of Findings

Our research revealed that the SAT is a strong candidate. It supports many of the IPTO Grand Challenge criteria, and requires many, though not all, of the capabilities desired in a cognitive system. The SAT is strongest in addressing the areas of knowledge, reasoning, and symbolic communication; however, it is problematic in the area of language generation.[1] The SAT is geared for the college-bound high school student and assumes that rudimentary language generation skills are already present; the multiple-choice items focus only on fine-tuning advanced writing. The one language-generation task, the Essay, is probably beyond the capabilities of systems within the next 20 years. Nevertheless, the SAT scores favorably on five of the six IPTO Grand Challenge criteria, so we believe it does merit further investigation.

Action Items for Moving Forward

To make the SAT Challenge a reality, the sponsor of the challenge must acquire training and test materials. Our recommendation is that the sponsor work with College Board and ETS to gain their interest and support, and to request that they administer the test and score the results. Although the ultimate challenge should be focused on the SAT, an intermediate option is to administer the PSAT rather than the SAT. Students who take the PSAT are allowed to keep their test booklets, so there would no issues with test security. Whichever test is used, it will need to be converted into computer-readable format, something that College Board does not currently possess. However, ETS, who creates the SAT for College Board and both creates and administers the GRE, does offer the GRE online, indicating that there is an existing on-line format and a method for converting these tests.

For practice tests, there are several options for acquiring materials. There are independent publishers of practice tests such as Kaplan and Barron's who might be willing to sell their materials. College Board and/or ETS might provide practice tests as well. Furthermore, even if the Grand Challenge itself uses the new SAT, the old-style SAT could be used for training material, obtainable perhaps from either independent publishers or College Board. Finally, DARPA already has experience in hiring ETS staff to write equivalent test items for use in specific applications.

[1] Although the SAT fails to address two cognitive functions desired by IPTO, namely learning and self-awareness, these could be addressed by slightly modifying the design of the challenge so that the system must demonstrate self-driven improvement in test performance over a series of tests.

Further study is also needed to determine the exact metrics for awarding the grand prize. For example, in our specifications for the challenge, we suggest an overall score in the 50th percentile, but it may be necessary to require a minimum score in each of the three test sections, to discourage idiot savant solutions.

Why Give the SAT® to Computers?

As a Grand Challenge, the SAT Reasoning Test™ (aka "the SAT") aligns favorably, though not ideally, with the IPTO Grand Challenge criteria:[2]

- *Clear and compelling demonstration of cognition?*
  Yes, it is typically a non-gameable proxy for a range of problems requiring knowledge, reasoning, and communication skills, but…
  No, it does not require self-awareness, and
  No, it's not clear that the challenge requires an autonomous learning capability on the part of the system.
- *Clear and simple measurement?*
  Yes, the metrics are already established, and the tests are already being created. Furthermore, there are human benchmarks against which to measure system performance.
- *Decomposable and diagnostic?*
  Yes, there are discrete components of the test designed to target unique abilities. Scores are not Pass/Fail, and performance on each item should point the way to needed improvement. Furthermore, ETS is interested in making the tests more diagnostic.
- *Ambitious and visionary, but not unrealistic?*
  Yes, scoring in the 50th percentile should be possible in 10-20 years if enough resources are committed to the task.
- *Compelling to the general public?*
  Yes, the public is familiar with standardized tests and would be interested in learning the results of the challenge: would a computer score as well as my child?
- *Motivating for the research community?*
  Yes, the problem has a high "cool factor," the work can begin right away, and continuous testing could be conducted over the web.

The weakest alignment is in the areas of learning and self-awareness. There are some Passage-Based Reading questions that test for knowledge acquired from the passage, but since all the system does with the knowledge is answer a question, this is probably too narrow a definition of "learning" for IPTO's needs. In general, the nature of the test is that there is no active learning that occurs during the test itself. As all high school students are painfully aware, what you know by the morning of the test is all you have to rely on. The only way to demonstrate learning would be to structure the challenge in

[2] See the report from the *DARPA IPTO Grand Challenge Brainstorming Workshop #1*, January 12-13, 2005, prepared by the MITRE Corporation.

such a way that the system would take one test, get feedback on its performance on each item, determine what it needs to improve, obtain what it needs from online sources or by asking its human creators, and then take another test. This departs from the way in which humans take the test, and thus undermines much of the appeal of the challenge, although humans taking SAT-preparation courses undergo much the same learning process. Finally, one might argue that although the test itself does not demonstrate learning, it is by its very nature supposed to demonstrate the *capacity* for learning, since it is used as one measure of a person's aptitude for higher education. Unfortunately, the SAT's predictive ability is a controversial claim in human assessment, and becomes more so in system evaluation, because we have no data to base conclusions on.

Another significant disadvantage of this challenge is that it does not compare favorably with the immediate military relevance of something like the DARPA autonomous ground vehicle challenge. That is, while the DoD does need autonomous vehicles, it does not need computers that take tests, and multiple-choice tests at that. Real-life problems seldom come with a selection of possible solutions, one of which is guaranteed to be correct. By choosing the SAT as a grand challenge, DARPA would be gambling that the technology required to achieve success on the SAT would be equally useful for practical applications. Thus, because the predictive ability of the SAT cannot be guaranteed when applied to computers, we attempt to do the next best thing by describing below how the individual components of the SAT are likely to align with specific cognitive abilities sought by IPTO. These capabilities are typically knowledge, reasoning, and symbolic communication.

How the SAT® Challenges Computer Technology

The new SAT Reasoning Test™ consists of three sections: Math, Writing, and Critical Reading. We base the following analysis on the new test as opposed to the former SAT that was used until 2004. The old SAT had only two sections, Mathematics and Verbal. Verbal is now called "Critical Reading" and it now contains paragraph and paired-paragraph (see page 11) reading items in addition to the longer single passages of the old SAT. Analogies have been eliminated. There was no Writing section in the old SAT. The following synopsis of the new SAT is based on information drawn from the *Official SAT Study Guide for the New SAT™*, College Board, 2004. The examples are drawn from the study guide as well as a 2004 PSAT and the sample SAT available for download by registered users at www.collegeboard.com.

Math Section

We anticipate that the purely mathematical component of each math item will not pose difficulty for today's technology, but significant development effort will be required to build systems with the ability to recognize the type of mathematical problem being presented (e.g., set theory vs. algebra), to translate the problem into a solvable mathematical representation, and to work out the solution. In terms of cognitive abilities, all the math items require knowledge of mathematics, and a component that does well on them should have utility in any application requiring mathematical problem solving based on human-readable data. There are four other features of the math items which increase the complexity for computer systems. Each of the features conspires to make it increasingly more challenging for a computer to understand the problem to be solved. The four features are as follows:

A. Language. All of the items have some natural language. It's never the case that the item contains only numerals and other mathematical symbols plus a list of answer choices. The simplest use of language is found in conditionals statements of the form "If <equation> which of the following is the value of <variable>?" and "If <equation> what is the value of <variable>?" While it will be possible for system developers to anticipate many of these common question templates, there appear to be many novel structures as well, so that some degree of language understanding will be required for these items. Items that have this feature but not B, C, or D below will be the easiest for computers and will demonstrate only limited communication abilities.

B. Supporting text/math. Some items contain an extra statement, typically mathematical, that fulfills the same supporting role as a diagram would. It will require the computer systems to resolve references to both the supporting material and the answer choices, but still falls toward the easiest end of the spectrum. For example:

> $2x - 5y = 8$
> $4x - ky = 17$
>
> For which of the following values of k will the system of equations above have <u>no</u> solution?
>
> (A) -10 (B) -5 (C) 0 (D) 5 (E) 10

C. Supporting diagram. Some items have one or more diagrams or tables. Usually the diagram/table is part of the question, though it can appear in answer choices as well. While there has been some work in diagram understanding, the enormous variety of graphical representations encountered will pose significant challenges for systems. They will have to recognize what is being represented – which can be anything from a floor plan to a line graph – and accurately relate the visual information to the linguistic and mathematical information. The cognitive ability demonstrated here is the ability to acquire information from broad-domain cross-modal sources. Items with supporting diagrams will fall toward the more difficult end of the scale.

D. Real world concept. Some items refer not just to mathematical concepts, but also to non-mathematical objects or concepts such as rugs, wire, dough recipes, etc. This feature is always found in "Story Problems," but also in shorter items that do not involve "stories" per se, but use real-world objects to set up problems about geometry, quantities etc. For example:

> If as many 7-inch pieces of wire as possible are cut from a wire that is 3 feet long, what is the total length of the wire that is left over? (12 inches = 1 foot)

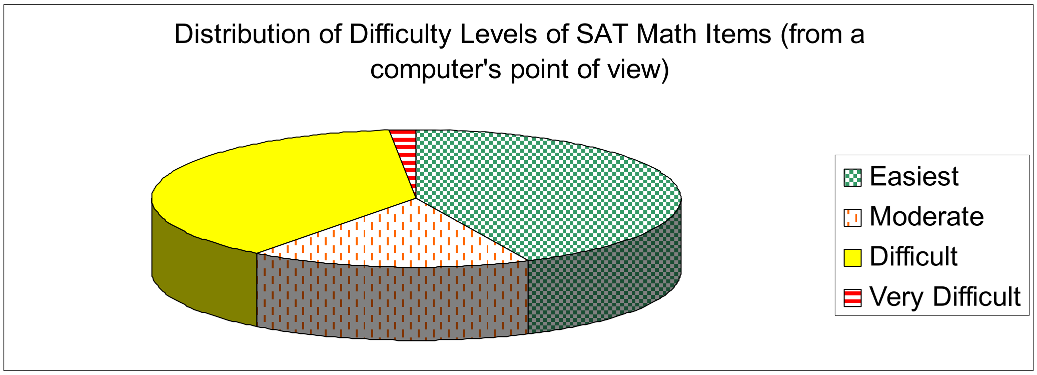
Here the system must be able to generalize on the basis of a specific example, and to identify which features of an object are relevant to the task at hand, e.g., that the important feature of “wire” in the above example is that it has length, not that wire is usually metal. Thus, knowledge and reasoning are required to understand the question, even before the process of solving the problem can begin. However, given the preponderance of a small set of mathematical terms that will also be present, this is not likely to be as difficult as an open-domain language understanding task. Items with real world concepts but none of the other complicating features thus fall into the mid-range of difficulty.

As said above, feature A appears in all question types, and the others can be combined in various ways, so that the more of these features an item has, the more difficult it will be for today’s technology. We analyzed the free sample SAT test available from the College Board website and found that the above features combine to create multiple levels of difficulty from the point of view of a computer (note that these features may or may not have anything to do with how humans would perform). The following table shows how these features combine, and the number of items at each difficulty level.

Table 1. Difficulty Levels of SAT Math Items (from a computer’s point of view)

Text	Supporting math/text example	Supporting Diagram	Real World Concepts	**Difficulty Level**	**Total Items**
√				Easiest	20
√	√			Easiest	5
√			√	Moderate	10
√		√		Difficult	19
√		√	√	Difficult	3
√	√	√	√	Very Difficult	1
					58

From a computer’s point of view, a good number of the items are on the easier end of the scale, promising that some traction will be obtained on the SAT Math section in the early years of the challenge. Equally encouraging is that most items will continue to push the technology for some time to come.


Distribution of Difficulty Levels of SAT Math Items (from a computer's point of view)

Writing Section

The Writing section of the SAT contains three types of multiple-choice questions and an essay question. The three multiple-choice question types are Identifying Sentence Errors, Improving Sentences, and Improving Paragraphs. Each multiple-choice item type poses increasing levels of difficulty for computers, though once again, this measure of difficulty may have no connection to how humans perform. The cognitive capability targeted by these multiple-choice items is communication, and a component that did well on this task would have utility as a means of polishing computer-generated text. However, with their emphasis on correcting and improving complex, sophisticated text, these items ignore the more immediate need, which is to have computers generate rudimentary text.

In **Identifying Sentence Errors**, the test item presents a sentence with four portions underlined. The goal is to identify which portion, if any, contains an error. E.g.:

The other delegates and him immediately accepted the resolution drafted by the neutral states.
A B C D
No error
E

Some of these errors are within the capabilities of existing grammar-checkers. We input 23 of these items into Microsoft Word and ran its Grammar Checker. Four of the items had no error, which Word's checker concurred with. Nineteen had errors, and the checker found four of them (including the one in the above example). Thus, its accuracy was 34%. Clearly, even this simple task is not completely solved by today's systems, but it does fall toward the easiest end of the spectrum.

The item known as **Improving Sentences** is designed to test one's ability to recognize and write clear sentences. The test presents a sentence and four alternative ways of wording a portion of the sentence, and the test-taker must choose one, or opt to leave it as is (always the first choice). For example:

Laura Ingalls Wilder published her first book and she was sixty-five years old then.

(A) and she was sixty-five years old then
(B) when she was sixty-five
(C) at age sixty-five years old
(D) upon the reaching of sixty-five years
(E) at the time when she was sixty five

These types of "errors" appear to be outside the capability of today's grammar-checkers. We supplied 15 items to Microsoft Word's Grammar Checker, and it had no suggestions that corresponded to the target problems and their solutions. Two of the items had no errors, so at most, we could say the software achieved 13% accuracy. Given sufficiently large training corpora of well-written texts, it might be possible to build systems that perform well on this task. We do not consider this a "cheap trick" because humans also probably perform the task based on what "sounds right." Given that this technology is not as far along as Grammar Checkers, this items falls in the mid-range of difficulty, but its utility might be limited to editing existing text.

The third type of multiple-choice Writing item is **Improving Paragraphs**. The test-taker is presented with a draft passage (actually several paragraphs long) and is asked questions about various ways to improve it. Some items involve improving sentence structure and word choice and so bear some resemblance to the “Improving Sentences” item. However, in this item such questions always require processing the context of the altered sentence, for example, replacing an ambiguous pronoun with a full noun phrase:

…**(6)** Consumers have the right to buy whatever they want. **(7)** They should consider the effects of their choices. **(8)** In the last several years, hundreds of thousands of workers in United States industries have lost their jobs. **(9)** They represent billions of dollars of lost wages and taxes.

Which of the following best replaces “They” in sentence (9)?

(A) The consumers
(B) These lost jobs
(C) The industries
(D) Those arguments
(E) The United States

In other cases, the changes can involve adding sentences, such as transitions or conclusions, or connecting two sentences without changing the meaning, and without creating any awkwardness, redundancy, or grammatical errors. Some items suggest adding various discourse connectives such as “however” or “also,” the correct choice of which depends on the viewpoint of the author being expressed in the rest of the passage. For example:

In context, which is the best version of the underlined portions of sentences 6 and 7 (reproduced below)?

Consumers have the right to buy whatever they want. They should consider the effects of their choices.

(A) (As it is now)
(B) Consumers certainly have the right to buy whatever they want, but they should consider
(C) Consumers certainly have the right to buy whatever they want, regardless of
(D) Although consumers have the right to buy whatever they want, they also consider
(E) Apparently, consumers have the right to buy whatever they want. If only they would consider

Another item type asks the test-taker to choose topics for additional paragraphs that would strengthen the writer’s argument.

The Improving Paragraphs item is a very challenging task, as it requires genuine understanding of the passage on many levels – the grammar, the meaning, and the writer’s overall viewpoint and purpose. This task falls into the more difficult end of the scale, and as with the other multiple-choice writing items, it only challenges the ability to improve sophisticated text, at a time when computers are not yet generating such text.

The **Essay** is the only section requiring actual language generation. This alone is extremely challenging, but has the advantage from a Grand Challenge point of view that it is testing a desirable cognitive capability – the ability to develop and express ideas effectively. The SAT essay differs from summarization and report-writing tasks because the goal is not to collect and present facts. Rather, the test-taker must state a viewpoint and support it with examples from his or her own internal knowledge. Here is an example Essay item:

Think carefully about the issue presented in the following excerpt and the assignment below.

> Some people believe that there is only one foolproof plan, perfect solution, or correct interpretation. But nothing is ever that simple. For better or for worse, for every so-called final answer there is another way of seeing things. There is always a "however."

Assignment: Is there always another explanation or another point of view? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

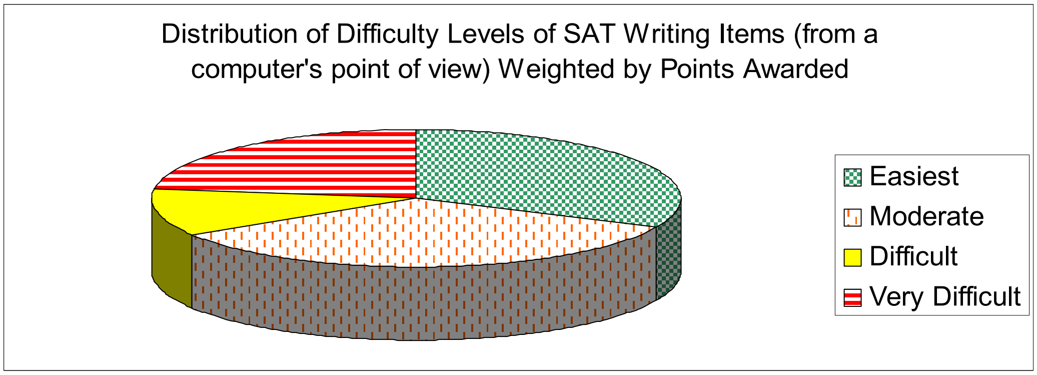
This type of behavior is far beyond today's systems. For the foreseeable future, computers are not going to have opinions on such matters as the above topic. At most, they could fake a view point, and gather evidence from their knowledge base to support it. The Essay is worth six points, and it is judged by two independent judges, so the most points one can achieve is 12. The only way to score 0 points is to write nothing at all, or write something that is not on topic. Today's system can be guaranteed at least 1 point from each judge by simply converting the essay question into a thesis statement, e.g.: "There is always another explanation or another point of view." But earning any points beyond that will be very difficult for many years to come. A large knowledge base will be required, combined with sophisticated reasoning and language generation abilities – something approaching science fiction proportions. It is possible, however, that a system could win the Grand Challenge by performing well on the Math, Reading, and multiple-choice writing items, and still fail to demonstrate the ability to generate text of any utility.

The free sample SAT test available from the College Board website has the following distribution of Writing items. We've indicated the relative difficulty level of each, based on the preceding analysis of what would be most challenging for computer systems. Note that while there is only one Essay item, it is worth 12 points, whereas the other items are worth one point each.

Table 2. Difficulty Levels of SAT Writing Items (from a computer's point of view)

	Difficulty Level	**Total Items**
Identifying Sentence Errors	Easiest	17
Improving Sentences	Moderate	18
Improving Paragraphs	Difficult	6
Essay	Very Difficult	1
		42

The following pie chart takes into account the 12 points awarded to the essay in showing the percentage of easiest versus most difficult writing tasks.



Critical Reading Section

The Critical Reading section contains two basic item types: Sentence Completion and Passage-Based Reading. **Sentence Completion** items contain a sentence with one or two blanks to be filled in. Of these, the College Board's prep material identifies two varieties: "Vocabulary-in-Context" and "Logic-Based." "**Vocabulary-in-Context**" questions require the test-taker to know the definition of the words, and some even read much like dictionary entries:

> "A judgment made before all the facts are known must be called _____"
> (A) harsh (B) deliberate (C) sensible (D) premature (E) fair

These and other simple fill-in-the-blank items will be fairly easy for statistical systems trained on large corpora and systems with dictionary definitions in memory. This is probably the only item type in the reading section that some reasonable performance will be possible with little or no genuine language understanding.

"**Logic-Based**" Sentence Completion items, on the other hand, require the test-taker to know the definition of the words, plus understand the overall logic of a complex sentence. Introductory and transitional phrases are key for this type ("but," "although," etc.) as well as negatives. See, for example, the importance of "despite…" in the following:

> Despite their _____ proportions, the murals of Diego Rivera give his Mexican patriots the sense that their history is _____ and human in scale, not remote and larger than life.
>
> (A) monumental..accessible (B) focused..prolonged (C) vast..ancient (D) realistic..extraneous (E) narrow..overwhelming

Such items also require the reader to be aware of how the presence of a given word (concept) in one part of a sentence implies what is most sensible in another part. For example, consider the impact of "reversal" in the following:

The Supreme Court's reversal of its previous ruling on the issue of state's rights _____ its reputation for _____.

(A) sustained..infallibility (B) compromised..consistency (C) bolstered..doggedness (D) aggravated..inflexibility (E) dispelled..vacillation

Thus, even in this most simple of tasks (from a computer's point of view), the SAT will require systems to do more than "cheap tricks"; these items will likely require some logical representation of the sentence a well as vocabulary knowledge. As a crude test of the usefulness of simple string matching for this task, we choose three Logic-Based items, entered substrings containing the answer choices into Google, and counted the number of hits. We chose a substring that would best take advantage of nearby meaningful words. For example for the penultimate example above, we tried "monumental proportions," "focused proportions," "vast proportions" etc. The correct answer received the highest number of hits in only one of the three items. It must also be remembered that the systems will not have access to such large corpora during the actual test itself.

Passage-based Reading involves reading one or two passages and answering questions about the material. The passages can be short or long. The paired passages often ask questions requiring the reader to compare or contrast the two. The College Board's test prep material identifies three question types: "Literal Comprehension," "Vocabulary-in-Context," and "Extended Reasoning."

Literal Comprehension questions are intended to test for information acquired through reading. These do not appear to be simple "factoid" (i.e., "who-what-when-where" questions). The Literal Comprehension questions do not occur often, and when they do occur, they typically require that one has read and understood multiple non-contiguous sentences in the passage and how they relate to one another, and also require recognizing some sophisticated paraphrasing in terms of vocabulary and structural variety. Other examples require understanding a single very long complex sentence. For example:

…School might frankly *be* the place where one reads the books that are a little off-putting, that have gone a little cold, that you might overlook because they do not address, in reader-friendly contemporary fashion, the issues most immediately at stake in modern life but that, with a little study, turn out to have a great deal to say…

In [the lines above], the author describes a world in which schools teach books that are

(A) interesting
(B) celebrated
(C) uncontroversial
(D) not obviously relevant
(E) not likely to inspire

These Literal Comprehension questions come closest to the types of questions that are addressed by current question answering systems, so they are the easiest of the Passage-Based Reading items. But they are also much more difficult in the level of reasoning and language understanding they require.

Vocabulary-in-Context asks about the meaning of a word as it is used in the passage. Unlike the Sentence Completion vocabulary questions discussed above, these focus more on words that have several meanings, and in this case the passage "won't necessarily use the most common meaning," according to the test makers. For example:

> … That is the state of reading, and books, and literature in our country, at this time.
>
> In [the line above], "state" most clearly means
>
> (A) government
> (B) territory
> (C) condition
> (D) scale
> (E) mood

This item is closely related to word-sense disambiguation, which has been a topic of research in natural language processing for many years but continues to be a difficult task for systems. It is a fundamental skill in language understanding and often functions as a roadblock to making progress in higher-level language processing tasks.

Extended Reasoning requires the test taker to draw conclusions from the information in the passage, or to evaluate the information. It includes:

- Identifying the function (rather than meaning) of a word, stylistic device, example, punctuation, etc.
- Identifying the overall theme.
- Identifying the author's tone, attitude, viewpoint, purpose.

For example:

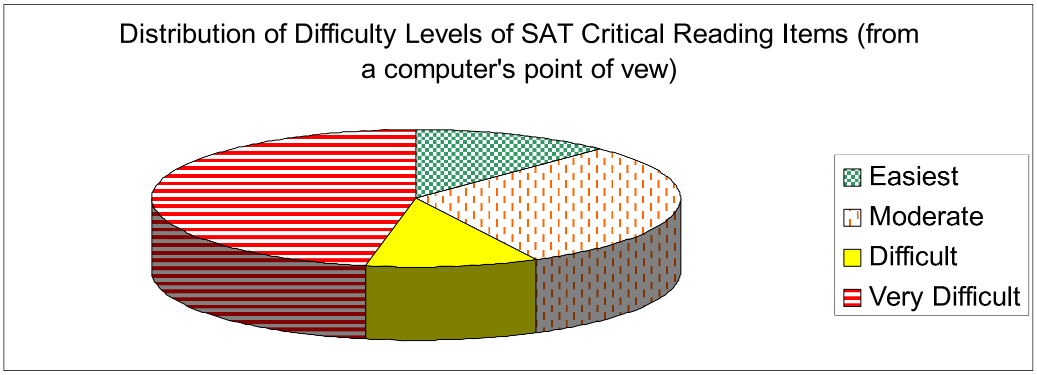
> …Years later, it may even be hard for them to remember if they read *Jane Eyre* at home and Judy Blume[1] in the class or the other way around.
>
> [1] *Jane Eyre*, by Charlotte Brontë, is a nineteenth-century novel. Judy Blume writes contemporary young adult novels.
>
> In [the lines above], the author cites *Jane Eyre* and Judy Blume primarily in order to
>
> (A) propose that a love of reading might blur a commonly perceived distinction
> (B) show that younger readers cannot distinguish between literature of different eras
> (C) argue that most modern novels have no lasting impact on readers
> (D) observe that classic literature has great appeal for even reluctant readers
> (E) indicate that certain works are interchangeable

From a computer's point of view, the Extended Reasoning items are a demanding test of higher-level language understanding and the ability to detect messages, information, and perspectives that are not overtly stated in text. A system that could perform well on this item should have many components applicable to intelligence analysis.

The free sample SAT test available from the College Board website has the following distribution of Critical Reading items. We've indicated in Table 3 the relative difficulty level of each item type, based on the preceding analysis of what would be most challenging for computer systems.

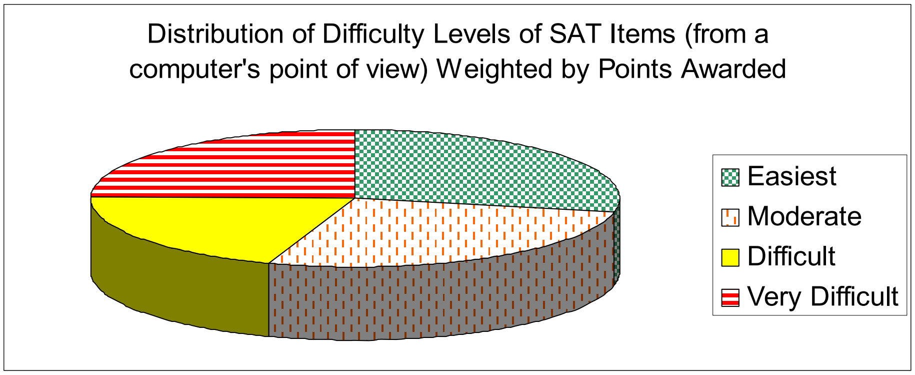
Table 3. Difficulty of Levels of SAT Critical Reading Items (from a computer's point of view)

	Difficulty Level	Total Items
Sentence Completion		
Vocabulary	Easiest	8
Logic-Based	Moderate	11
Passage-Based Reading		
Literal Comp	Moderate	9
Vocabulary	Difficult	7
Extended Reasoning	Very Difficult	31
		66



Summary

The following pie chart is an aggregate of the preceding charts analyzing the level of difficulty present in each section of the SAT. It appears that there's roughly equal distribution of the four difficulty levels (weighted by points awarded). There are enough items on the easiest end of the scale to make some progress in early years and draw participants into the test, and given the compartmentalized nature of the test, good performance on just a few items types could translate directly into useful applications. In addition, there are many difficult items to keep the challenge open for years to come. Further study is needed to determine the score required for the grand prize to be awarded. For example, DARPA's long-term needs might not be served by a system that is able to win the prize by performing at close to 100% proficiency on the easiest and moderate items.



Acknowledgments

Selmer Bringsjord of Rensselaer Polytechnic Institute and Michael Genesereth of Stanford University were instrumental in the discussions leading to this report.

Why Should College Board & ETS Support a Grand Cognitive Challenge Aimed at Building an AI Able to Ace the SAT?

Selmer Bringsjord selmer@rpi.edu
Lynette Hirschman lynette@mitre.org
Lisa Ferro lferro@mitre.org

draft of 6.15.05 6:18 pm

Our request to the College Board/ETS:

To become a partner with DARPA in running an annual "Grand Challenge Cognitive Competition" by providing on-line access to the current year's SAT for a set of "artificial students" and to grade their responses based on the norms developed for the grading of that year's SAT.

We propose that some number of competing artificial students (intelligent systems) would "take" the SAT: that is, access a server via an API that would permit each system to view the test items and return answers, within the time constraints designated for each section.

DARPA's goals in this undertaking are:

- To define an enticing yet ambitious Grand Challenge for cognitive systems that will foster ground-breaking research over the next ten to twenty years.
- To compare the state of the art in intelligent systems against human performance, using the SAT as a well-understood measure of aptitude.

We believe that this activity will not only encourage the development of intelligent systems, but that it will also provide significant benefits to the College Board and ETS. We believe that development of such systems would allow the College Board to:

1. Improve SAT Assembly
    - Better understand what lies behind item difficulty and pre-screen items via an automated system
    - Debug test items

2. Generate SAT Items
    - That are diagnostic
    - That will be able to support intelligent tutoring applications

3. Demonstrate Educational Soundness of SAT
    - Educate the public about what SAT tests are and why a machine can't do it right now
    - Demonstrate the primacy of reasoning over raw knowledge in the SAT

1. Improve SAT Assembly

(a) **Understand and Predict Item Difficulty**. The difficulty of items on the SAT (and, for that matter, on the GRE, LSAT, etc.) is determined statistically, rather than cognitively. An item whose difficulty is unknown is given to the test-taking population in the experimental mode, and, in general, how this population performs on the item determines its difficulty. Note that 'performs' here pertains only to answers that are right or wrong: there is little information about the thinking behind the selection of an option. This approach doesn't disclose what *cognition* behind the scenes enables a test-taker to discover the key, or fall for a distractor, in the context of a given test item. Research aimed at meeting the SAT grand challenge could lead to the use of computational systems that decompose item answering into discrete processes, thus providing insight into characteristics that can make an item easy or difficult. In addition, such systems can be used to pre-screen test items for difficulty.

(b) **Perform Quality Control Checks**. Work devoted to meeting this challenge will support technology that can "debug" the SAT to an unprecedented high level of accuracy. You will be able to submit an item to high performing systems to find out if the systems are consistent in their responses, if there really is only one clear key, and so on. So the DARPA-sponsored research could provide ETS with technology that is the ultimate insurance policy against an embarrassing item turning up in the *New York Times*. (This is a parallel to what Noah Friedland discovered in Halo I, using a portion of the chemistry GRE: that the publishers benefit from having their texts scrutinized by attempts to build machines able to assimilate them, because this produces a high-end editing/cleansing not obtainable by normal, human-only methods.)

2. Generate SAT Items

(a) **Semi-automatically Generate Test Items**. Cognitive systems able to solve SAT problems would pave the way toward systems able to generate SAT problems. It is currently a large expense to have humans write the items on these tests. ETS has sponsored some work in automatic generation, but there would be a *major* advance if the R&D in question takes place.

(b) **Create Diagnostic Tests and Intelligent Tutoring Systems**: Imagine a system able to generate test items targeted at specific student weaknesses. This is what points 1a and 2a would lead to. Such a system could be used to generate diagnostic test items for standardized tests (2a above), or operate in a tutoring mode with a student. Since the system would "know" both the question and the answer, it could provide diagnostics to guide the student to the correct answer, or help them to understand why their choice was incorrect. ETS, at least in the past, has considered spinning off companies that produce educational products (e.g., courseware), so if this business model is still in play, what DARPA proposes to do might well produce knowledge and software that could be sold by these companies.

3. **Demonstrate Educational Soundness of the SAT**. The grand challenge research effort would lend support to the validity of the SAT. It will be many years before a system can perform credibly on the SAT. This will give the public a chance to understand what the SAT is testing – reasoning rather than raw knowledge, and why computers are so limited at the former. This should give greater appreciation to the fact that the SAT is testing reasoning skills.

Take the SAT®

An autonomous system will take the SAT® and score in the 50th percentile of high school students taking the examination.

- ***The system must***
  - ***take the same test that is administered to high school students***
  - ***observe the same rules as humans regarding help, access to external resources (no internet connection), and time constraints****

*The exception is that the test must be administered in a computer readable format, including diagrams and formulas.

Cognitive Systems and SAT®

The new "SAT® Reasoning" test demands complex, integrated abilities in a cognitive system:

- Reasoning and Inference
  - Reason on the basis of text and diagrams *(math sections)*
  - Understand and take into account human perspectives and emotions *(passage-based reading sections)*
  - Draw conclusions based solely evidence presented *(math and reading sections)*
- Communication
  - Understand and follow directions *(math, reading, and writing sections)*
  - Detect subtle textual clues *(critical reading sections)*
  - Develop and express ideas effectively *(writing sections)*

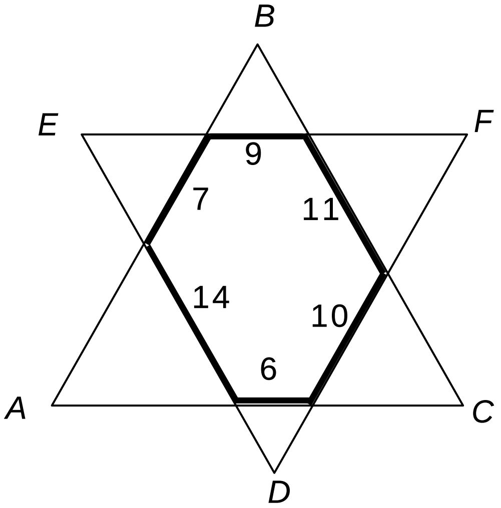
Overview of the New SAT®

- Math
  - Multiple Choice answers
  - Student-produced answers
- Critical Reading
  - Sentence Completion
  - Passage-based Reading
- Writing
  - Identifying Sentence Errors
  - Improving Sentences and Paragraphs
  - Essay Writing

Source: *The Official SAT Study Guide for the new SAT™*, College Board, 2004.

Math

- Reference formulas are provided by test materials ($C = 2\pi r$, $V=\pi r^2 h$ etc.)
- Number operations, algebra, functions, geometry, measurement, data analysis, statistics, probability



In the figure above, equilateral triangles *ABC* and *DEF* intersect so that side *AB* is parallel to side *DF*. The numbers indicate the lengths of the sides of the polygon outlined in bold. How much greater is the perimeter of triangle *ABC* than the perimeter of *DEF*? (PSAT 2004)

Critical Reading

- Sentence Completion tests vocabulary and understanding the logic of complex sentences.

In public, Henry was somewhat ------- toward his opponents; behind their backs, he was even more -------.

(A) sympathetic..furious (B) amicable..disparaging (C) caustic..vitriolic (D) bitter..patronizing (E) imperious..unctuous (PSAT 2004)

- Passage-based Reading tests reasoning and inference, comprehension, and vocabulary in context.

Sample Passage-Based Reading

Passage 1: "…Obviously, having a coelacanth in a tank would make **someone** a lot of money…."

Passage 2: " 'We need a live coelacanth in captivity,' said Mike Bruton from his base at the Two Oceans Aquarium…"

The comment in Passage 1 about "someone" implies which of the following about the Two Oceans Aquarium mentioned in Passage 2?

(A) It has plans to support coelacanth conservation programs.

(B) It could benefit financially from displaying a live coelacanth.

(C) It has great expertise in simulating the coelacanth's habitat.

(D) It might provide scientists invaluable access to live coelacanths.

(E) It would be the first institution to breed coelacanths in captivity.

(PSAT 2004)

Writing

- Identifying Sentence Errors tests ability to find mistakes in grammar, usage, and word choice
- Improving Sentences and Paragraphs tests ability to recognize and produce clear and effective writing
- The Essay tests ability to develop and support a viewpoint

Sample Essay Question*

Think carefully about the issue presented in the following excerpt and the assignment below.

> Some people believe that there is only one foolproof plan, perfect solution, or correct interpretation. But nothing is ever that simple. For better or for worse, for every so-called final answer there is another way of seeing things. There is always a "however."

Assignment: Is there always another explanation or another point of view? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.

* Source: *The Official SAT Study Guide for the New SAT™*, College Board, 2004.

Administering the SAT® to Computers

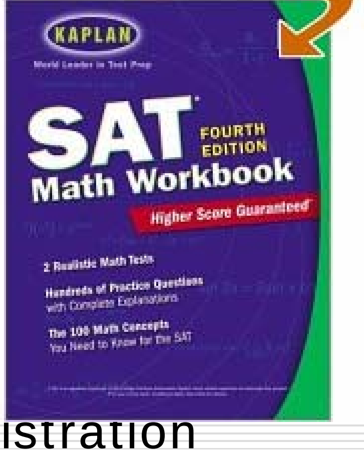
- Participants arrive at central site on test day, with their systems on laptops
- Systems will access a server via an API that permits each system to view the test items and return answers
- Answers will be graded by College Board following standard procedures

SAT® Challenge: FAQ

- Will the computer be allowed to use a calculator? *Yes, and so are human students.*
- Will the computer be allowed to store as much information as it wants? *Yes, that's what it is good at.*
- Will the computer have to write an essay? *Yes, though we expect initial results to be poor*
- How will you prevent cheating? *An oversight committee will evaluate each entry for compliance to the spirit of the challenge.*

Taking the SAT®

FY'06 Proposal


KAPLAN
World Leader in Test Prep
SAT
FOURTH EDITION
Math Workbook
Higher Score Guaranteed
2 Realistic Math Tests
Hundreds of Practice Questions with Complete Explanations
The 100 Math Concepts You Need to Know for the SAT

- Work with College Board and ETS
  - Stimulate their interest and support
  - Facilitate agreements for materials and test administration provisions
- Obtain sample tests
  - College Board & ETS
  - Kaplan & Barron's practice materials
- Build and test an API and computer-readable test format
  - Graphics
  - Formulas
  - Questions and answer choices
  - Submitting the selected answer
- MITRE LOE: 3 SM

Report Generator
Handy Andy, the DARPA Essayist

Automated AI Systems Compete Against Invited Human Contestants

The Handy Andy Challenge is to produce a multi-page report on any topic in response to a user request. It involves at least three subtasks: understand the request, find appropriate content, and produce an informative and well-organized write-up. The assessment will be based on both human and automated measures, maintaining two essential criteria: ranking of reports produced by such metrics will need to remain stable across different sets of judges, and reports that are in fact similar should get similar ranks.

For follow-on work, we propose to design a feasibility pilot applied to After Action Reports.

Attached documentation:

Challenge Description
Handy Andy: A Cognitive Grand Challenge
Detailed description of the challenge, rules, and specifications

Supplemental Report
MITRE supplement which focuses on possible applications and evaluation metrics

Briefing (used in AAAI Presidential Address)
Single-slide overview of challenge with supporting slides in more details

FY'06 Proposal
Plans for follow-on work [3 SM]

Handy Andy: A Cognitive Grand Challenge

Paul Cohen, Beatrice Oshika, Kathy McKeown, David Waltz[1]

June 23, 2005

The Handy Andy Challenge is to produce a multi-page report on any subject. This document describes how the challenge will be administered, how contestants will be scored, how the challenge will be made more difficult each year, and what makes it an effective challenge and a worthy successor to Turing's test.

Imagine cognitive systems that can satisfy requests like these:

> As a representative of a NGO, I need a "cultural briefing" on the people of Qatar.
>
> How do airplanes fly?
>
> I know about the Turing Test, and I've heard that other great scientists have issued challenges in the past (e.g., Hilbert). Write me a catalog/history of these challenges.
>
> I invited eight people to dinner, two of them vegetarian, and I'd like to cook some sort of ethnic food, can you suggest a menu?
>
> Produce a manual to help me understand the process of buying a house in California.
>
> What is retinitis pigmentosa and what can I do about it?
>
> I am nine years old. What should I eat if I hate cheese?

These requests involve quite different technologies, some of which do not yet exist. If the Handy Andy Challenge is to write an essay that satisfies all aspects of any such request, then it will suffer the same fate as Turing's test, which is so absurdly out of reach that few researchers even think about it. If instead Handy Andy is administered as a graduated series of challenges, then we might use these to encourage new technologies for "killer applications" of the World Wide Web – applications that depend on *understanding* documents on the Web.

The Challenge clearly has at least three parts: comprehending the request, finding appropriate content for the essay, and producing a well-organized, informative essay. Note that the Challenge is to *produce* an essay, not write one de novo. In the early years of the Challenge, contestants will be encouraged to compile essays from relevant material on the Web. However, the Challenge is easily extended to include other tasks such as original writing, forming and expressing an opinion, comparing several positions, learning from previous requests and constructing a report largely (or entirely) from previously-learned knowledge.

[1] This proposal has been shaped over the last few months by Ed Feigenbaum, Ed Hovy, Kevin Knight, Craig Knoblock, Kristina Lerman, Daniel Marcu, Tom Mitchell, Tim Oates, Roger Schank, and Wei-Min Shen. Some liked it, others liked it less, but all contributed ideas and suggestions.

The Challenge Protocol

The Handy Andy Challenge is an open competition for artificial systems and an invitational competition for humans. Human and artificial systems compete in each of several leagues or tracks. Some leagues will be appropriate for children. All the contestants are required to produce three essays in the course of three hours. All are provided access to the Web. The essays are scored by a panel of expert judges according to criteria discussed later in this document.

Cash prizes are awarded for the best essays in each league and a grand prize is awarded to the artificial system that performs best in all the leagues. More challenging leagues get bigger prizes. In the event that an artificial system beats the best human in its league it will receive a bonus prize – and be required to compete in a "higher" league next time!

Perhaps the most compelling way to assess the quality of essays from artificial systems is to compare them with essays authored by humans. These might be found on the web or in books, or commissioned from college students who need summer jobs, but we think it will be more exciting and engaging for the public, and methodologically more sound for these essays to be produced by human contestants in the Handy Andy Challenge.

Human contestants must be invited. In the first years of the Challenge, they might be nominated by teachers or principals of schools local to Washington DC.

Leagues

All contestants must produce essays, but different leagues emphasize different aspects of the challenge and different levels of competence. We anticipate having four General Leagues in the first year of the Challenge: Elementary, Middle, Secondary, and College.

General League. The task for the General League is to produce "fact-based" essays. Fact-based essays are made of expository sentences and paragraphs about things that are true; for example, an essay on the life of Picasso should mention that he was a prolific painter, an inventor of many styles, influenced by African art and by Cezanne, married several times, and so on. A fact-based essay is not expected to offer an original opinion, create an elegant phrase, construct a history, compare and contrast, or argue a case. These aspects of essay production (and others) are stressed in Specialist Leagues, discussed shortly. Nor must the essay be original; in fact, General League contestants are encouraged to assemble essays from material available on the Web.

Special Leagues. The task for special leagues is to produce essays, just as it is for the General League. The Special Leagues stress different aspects of producing essays, though none relaxes the requirement that contestants produce essays.[2] Here are some suggestions for Special Leagues:

[2] AI has a disconcerting habit of aiming low and the goal of a Grand Challenge is to make us aim high, even if we don't succeed at first. If contestants are allowed to enter leagues that solve parts of the Handy Andy problem, then they will, and we may never see systems that solve the whole problem. So, we require all systems to solve the whole problem, if poorly.

Historical/Narrative League. Here the challenge is to get the historical facts in the right order and draw the causal connections appropriately. For example, an essay on Darwin and Natural Selection should make clear that the theory was developed over several years and was triggered by observations Darwin made as a naturalist aboard HMS Beagle.

Comparative Essay League. Comparative essays are organized around comparisons and are judged according to how well they identify and present the elements of comparisons. For example, one might be asked to compare Texas barbecue with Kansas City barbecue, or ocean liners with airplanes, or the political systems of the US and Canada.

Original Writing League. The challenge is to write (as opposed to compile) an essay. For example, contestants might be asked to form an opinion, or thesis, on the role of alcohol in the book *The Sun Also Rises*. Contestants are scored on the fraction of their essays that are original, and on the quality of the language. We expect humans to do much better than machines in this league, but we hope that some machines might perform as well as elementary-school children.

Human-Machine Collaboration League. In this league, essays are produced by humans and machines working in concert. Scoring this league is slightly more complex than scoring the others: First, a winner is selected based on the criteria discussed below, but the winner gets a prize only if it is better than, say, the 75th quantile of human contestants in the appropriate General League. In this way we hope to control for the possibility that the real work was done by the human member or the human-machine collaboration. An alternative (though weak) control is provided by allowing much less time to human-machine teams – say, ten minutes instead of an hour.

The Polymath League. Contestants in this league eschew web sources and write their reports based largely on information they already know. To score essays in this league we need to distinguish material in memory from material taken from the web. One way to do this is to require the contestants to submit a draft essay before they are given access to the web, and another essay thereafter.

Expert Leagues. While the Handy Andy Challenge is to produce an essay *on any topic*, we recognize that some areas of human endeavor are enormous and it will do little harm to the universality criterion (see Section) to require essays *on any topic within a huge area*. We have in mind areas of professional expertise such as medicine, astronomy, art history, and such.[3] We would prefer not to introduce expert leagues in the first years of the Challenge for fear that their focus might become narrowed and the universality criterion lost.

[3] We are discussing a Handy Andy system for the National Library of Medicine with its director, Dr. Donald Lindberg.

Essay Topics

The topics of the essays are not known to the contestants before the competition. Each topic is described in English and in a formal language. (In later years the formal representations might be phased out, but in the first few years of the Challenge it will enable contestants to participate without having to understand idiomatic natural language.)

Because most topics can be approached in many ways, a good essay assignment will identify both a topic and how it should be approached. It's one thing to ask for an essay about the automobile, another to specify the effects of automobiles on the environment.

Each league will be given appropriate topics; for instance, the Elementary league might be asked to write about nutritious foods, or the right way to behave on a play date; while the College league must assess the legacy of Dr. Martin Luther King, or provide a short history of Islamic astronomy.

Topics for the Specialist Leagues will be selected to stress the functionality that defines each league. For instance, topics for the historical/narrative league should have a clear historical story to be told; the history of steam power would be good, drought-resistant plants for small gardens probably wouldn't be.

Scoring

Each essay will be scored by human judges. The judges will be recruited from the institutions that supply the human contestants; i.e., they will be teachers and professors. Some of them will help us develop scoring criteria. To the extent possible, Handy Andy systems should be evaluated according to the criteria that teachers apply to their students.

Judges will be blinded and, in particular, they will not know whether the essays were authored by humans or artificial systems.

Two principles will guide the development of scoring methods: First, each essay will receive a score that is a simple, easily understood weighted sum of several criteria. Second, some weights will be different in different leagues. For example, for the General League, we envision awarding up to 100 points for an essay, divided like this:

15 points: The content of the essay should be relevant to the assigned topic

15 points: Certain facts or events must be reported in the essay

15 points: The essay should incorporate multiple sources. It shouldn't recapitulate or copy a single source on the web.

20 points: The essay should not be redundant. Copying several articles that say roughly the same thing is bad. Alternatively, the essay should be concise.

20 points: The essay should be well organized. It should not simply append material found on the web.

5 points: Copying material verbatim from the web is acceptable (at least in the first years of the challenge), but the essay should cite its sources.

5 points: The essay should contain some language generated by the contestant, not copied from the web.

5 points: The essay should make an original assertion, something not explicitly stated in any of the sources (the assertion needn't be in English in the first year of the Challenge, it could be in some formal language).

Special Leagues have the same criteria but may add others and adjust the number of points assigned to each. Some of the Special Leagues require an essay to do something – make comparisons, construct histories, write original text – and in these cases a large number of points will be associated with these tasks. If the instructions are to compare two things or concepts, then the essay should include comparative statements. If the instructions are to provide a manual or a procedure, then it ought to be possible to execute the procedure effectively on the basis of the essay.

Comparisons and Evaluations

One would like to assess how well the artificial systems perform relative to "less cognitive" technology, and to human contestants and each other. Google is an excellent example of less cognitive technology, the more so because it is really very good. We will run a simple Google-based essayist as a comparison, as follows: Google will be run on the content words in the essay assignment and the text fields of the top three hits will be appended to produce an essay, which will be scored by the judges just as all the contestants' essays are.

It is very desirable to compare artificial systems with humans at different grade levels, to say, for example, that the best artificial system is as good as the average fifth grader overall, or on particular dimensions. The comparison is difficult because human contestants are graded relative to their age level (otherwise the college kids would always beat the fifth graders) so there isn't a single scale on which all contestants – human and machine – can be compared. The problem is that judges will be inclined to say, "this is very good organization [or whatever] *for an elementary school kid*," whereas a college student would be marked down for the same performance. If the elementary school child and the college student need very different absolute levels of performance to get the same score, how can we judge the "grade level" of an artificial system? A related problem is that the assignments given to younger contestants will be less challenging. Children read and write more slowly than college students, and they have less world knowledge, so the elementary General league will set essays on kid-friendly themes and accept shorter and less thorough essays. What does it mean to say that an artificial system is as good as elementary school students? Does it mean the system can only respond to kid-friendly themes and can only produce short essays? Clearly not.

Finally, artificial systems are very likely to exhibit what Piaget called decalage: The lagging of particular skills relative to age-group norms. An artificial system might outperform a college student in its ability to find relevant material, but fall below a fifth grader in its ability to organize the material.

These factors preclude straightforward assessments of machines' grade levels. We could always ask teachers to assess the grade levels of machines on various aspects of the General and Special League tasks, but this is indirect – it doesn't involve a direct comparison between human and machine contestants. We will continue to work on this issue.

Ongoing Administration of the Handy Andy Challenge

A problem that can be solved in a year probably isn't a grand challenge, so we must consider how Handy Andy will be administered over several years. Robocup provides an excellent model. The robotic soccer community has a 50-year goal, to beat the reigning human world champion soccer team. Each year, the community elects a steering committee to moderate debate on how to modify the rules and tasks and league structure for the coming year's competition. It is the responsibility of this committee to to steer the community toward its ultimate goal in manageable steps. The bar is raised each year, but never too high; for instance, this year there will be no special lighting over the soccer pitches. It isn't all fun and games: To play in the annual Open competition, one also must present a paper at the colocated Symposium. Prizes are awarded to research just as they are to on-field competition.

Although the robotic soccer community faces tougher problems each year, the fundamental task never changes: Play soccer. This combination of an easy-to-state task and a graduated series of technical challenges, set by experts in the field, has produced remarkable progress.

We recommend essentially the same model for the Handy Andy challenge. The fundamental task is to produce an essay. New technical challenges can be introduced every year. One way this will happen is to move Special League challenges into the General League. Here are some examples:

- Require a significant fraction of the essay to be written rather than copied from web sources.
- Require that multiple sources be integrated in an essay.
- Require the contestant to make and report inferences about aspects of the assigned essay subject.
- Require the contestant to compare objects, events, assertions, etc.

Another approach is to do away with devices that were intended to make the Handy Andy problem relatively easy and accessible; for example,

- Do away with the formal specification of the essay topic, provide the topic only in English.
- Limit the access of contestants to the web, make them rely more on stored knowledge.

New Special Leagues will function as incubators for even more challenging problems. Examples include:

- Grade an essay according to the criteria used by human judges, achieve high concordance with human judges' scores.

- Given an essay that argues for a position (e.g., global warming is the result of human activity) produce an essay that argues for the opposite position and specifically challenges particular assertions in the original essay.

Is Handy Andy a Cognitive Grand Challenge?

A grand *cognitive* challenge wouldn't deserve the name if success didn't depend on comprehension, understanding, semantics, content. You can't measure comprehension the way you can measure power consumption and processor speed. You have to measure something that *depends* on comprehension, instead. This is how it works in elementary school: After a child reads a paragraph she answers questions about it, and if her answers are correct, we say she understands. You can make the test less challenging by providing multiple-choice questions, more challenging by requiring the child to write a summary from scratch. Our test – produce a five-page report on any subject – is as challenging as any elementary school comprehension test, indeed, it is a common challenge for college students. In fact, our test involves two comprehension tasks: First, understand the user's query; second, understand material on the Web well enough to use it to respond to the query.

Of course, one can imagine not-very-cognitive solutions to the Handy Andy Challenge, but if the scoring methods are constructed properly, these should not receive high scores. For example, we expect the Google-based essayist (described above) to score poorly on criteria such as conciseness, good organization, making novel inferences, writing original sentences; and to score not consistently well on recall and precision in fact-based essays.

Moreover, it is the responsibility of the Handy Andy steering committee to make the Challenge more cognitive each year. This means emphasizing stored knowledge, common sense inference, semantic analysis, and learning in successive versions of the Challenge. Although we can't solve the problem today, we may hope that by a graduated series of challenges we might steer Handy Andy technology to the point that it can conduct the kind of semantic analysis and comparison implicit in one of Turing's questions, "In the first line of your sonnet which reads 'Shall I compare thee to a summer's day', would not 'a spring day' do as well or better?"

Learning was not criterial for Turing, but it is for us. We want our system to learn to produce better reports. So much knowledge is required to do this well, so much can be learned. For instance, good reports have good structure, they aren't redundant, they use rhetorical devices, they introduce new vocabulary; all these writing techniques might be learned from examples of good reports or by advice-taking. However, all depend on a more fundamental skill, the ability to read and understand text. Most of the world's knowledge is encoded in text, and learning to read makes increasing amounts of this knowledge available to children. We have to be a little careful here, as children can "read" text they do not understand. They can parse the sentences and even get the intonation of dialog right, but they don't know all of what's being said. Understanding is not a binary predicate, one understands more or less of a text, depending on how much one knows about the subject. By judicious choice of essay subjects and scoring criteria, the Handy Andy steering committee can actually steer contestants toward learning to read and understand text, and thus to have access to the world's knowledge.

Is Handy Andy an Effective Challenge?[4]

An effective challenge makes people pay attention and change what they are doing. For many reasons the Turing test is not an effective challenge to artificial intelligence. We believe the Handy Andy Challenge accomplishes the goals of Turing's test more effectively than the test itself.

The Universality Criterion

A defining feature of our cognitive grand challenge, one it shares with Turing's test, is its *universal scope*. You can ask about the poetry of Jane Austen, how to buy penny stocks, why the druids wore woad, or ideas for keeping kids busy on long car trips. Whatever you ask, you get five pages back.

The universality criterion entails something about evaluation: We would rather have a system produce crummy reports on any subject than excellent reports on a carefully selected, narrow range of subjects. Said differently, the challenge is first and foremost to handle any subject, and only secondarily to produce excellent reports. If we can handle any subject, then we can imagine how a system might improve the quality of its reports; on the other hand, fifty years of AI engineering history leaves us skeptical that we will achieve the universality criterion if we start by trying to produce excellent reports about a tiny selection of subjects. It's time to grasp the nettle and go for *all* subjects, even if we do it poorly.

The Web already exists, already has near universal coverage, so we can achieve the universality criterion by making good use of the knowledge the Web contains. Our challenge is not to build a universal commonsense knowledge base, but to make better use of the one that already exists. We accept that machines cannot understand Web pages, today, and that our system will produce crummy reports at first; yet the answer is not to give up on universality, but rather to work on better comprehension and producing better reports.

The Come-as-you-are Criterion

Turing's test requires simultaneous achievement of many cognitive functions and doesn't offer partial credit to subsets of these functions. As we have said repeatedly in this proposal, we favor a graduated series of challenges, each just *slightly* out of reach. We begin with challenges to today's technology: The first challenge *intentionally* is within striking distance of current information retrieval and text summarization methods. Unlike Turing's test, an all-or-nothing challenge of heroic proportions, completely out of reach today, we begin with technology that is available today and proceed step-by-step toward the ultimate challenge. Consequently, we do not require a preparatory period to build commonsense knowledge bases, ontologies, inference engines, and the like. We

[4] This section borrows from Cohen's unpublished essay "If not Turing's test, then what?" which is available from the author.

aren’t in the position of waiting for some prerequisite (e.g., a “critical” amount of knowledge in Cyc which will enable Cyc to read). This is a strong methodological point because those who wait for prerequisites usually cannot predict when they will materialize. Our approach is to “come as you are” and proceed through a series of increasingly-stringent challenges.

The Ample Rope Criterion

We have been asked, “Why five pages? Some queries can be answered with a single word, others require many pages. Five seems arbitrary.” In response we say the challenge is to write a report, not provide a one-word answer to a question, and the required length should be sufficient for the system to make significant mistakes. This criterion might be satisfied by requiring three pages or seventeen, the number won’t matter as long as it gives the system enough rope to hang itself.

Other Criteria

The Handy Andy Challenge satisfies many other methodological requirements, such as transparency, diagnosticity, continuous testing, and the like. In addition, it could have very practical consequences. Much of the world’s knowledge is in textual form, and increasing amounts of it are on the Web, yet machines understand essentially none of it. This problem doesn’t have a quick and easy solution. However, the Handy Andy Challenge can steer researchers to develop technologies to provide increasingly sophisticated kinds of understanding of documents on the Web. Perhaps the world doesn’t need an artificial essayist, but it could certainly use technologies that understand written material well enough to write an essay.

Supplement to Handy Andy:
A Cognitive Grand Challenge

1 Introduction

This is a supplement to the Handy Andy Challenge (Cohen et al. 2005) (HAC) that describes a possible DARPA Grand Challenge problem of getting a computer system to produce a multi-page report on any subject in response to a user query. The HAC task is aimed at dramatically accelerating the pace of research in cognitive artificial intelligence, and is suggested as a successor to Turing's test for ascribing intelligence to machines. The idea is that a system that has "understood" a subject should be able to produce a report on it of at least the same quality as humans can. This supplement suggests a possible application domain and relevant metrics.

The HAC is conducted as a competition where humans and machines compete in various tracks, called leagues. Contestants are required to produce three reports in the course of three hours on a topic that they have not been told about beforehand. All are provided access to the Web. The resulting reports are scored by a panel of expert judges according to a variety of criteria. Systems will be compared against each other as well as against humans. Topics will described in English as well as (at least initially) in a formal language. A variety of different leagues are suggested. The General League requires "fact-based" reports, made up of expository text about things that are true. The Historical/Narrative League requires getting the historical facts in the right order and drawing appropriate causal connections. The Comparative League requires identifying and presenting the elements of comparisons. The Original Writing League requires writing (rather than compiling) a report. The Human-Machine Collaboration League involves reports produced by humans and machines working in concert. In the Polymath League, contestants are required to use information that they already know, rather than exploiting information from web sources.

Presumably, a human or artificial system that produces a high-scoring, novel report on a subject will exhibit some level of intelligent understanding of the subject matter. Such a system would have to understand what the question means, then find relevant information from knowledge sources relevant to the topic, then assimilate and organize the information for inclusion in the report, and then produce the report in natural language. The level of linguistic sophistication needed for a system to understand the topic/question and generate the report text may be constrained in initial versions of the task.

2 Possible Application Domain

It would be useful to emphasize the production of reports that are of interest to researchers and that also have relevance to DARPA. A possible common domain is education and training, covering reports on general topics as used in educational testing in schools, and also reports on topics that can help support military *training*.

Such training reports, which need to be in a form where they are clearly and quickly understood by different levels of users, are currently produced by organizations like the Army's Training and Doctrine Command (TRADOC). According to the General Accounting Office (GAO 2003), maintaining training and readiness of forces is a substantial challenge for TRADOC. Further, the laborious effort required for humans to produce such reports precludes the mass production of customized reports on different topics. Automating such reports can therefore have considerable practical utility

In addition, there are extensive existing training materials which could be made available to the various tracks for system training or as on-line resources for report writing. Finally, such an application lends itself to task-based evaluation, since there are trainers who can assess how well people are being trained by human-produced versus machine-produced materials.

Accordingly, examples in the HAC proposal (What is retinitis pigmentosa? Write a catalog history of the Turing Test) can be extended to include topics for which there exists a current need:

1. *What is the Geneva Convention on the Rules of War?* (ArmyStudyGuide-000144 2005)
2. *Produce a short manual explaining how to respond to a depleted uranium contamination.* (ArmyStudyGuide-0024 2005)
3. *What is the correct procedure for cave search using a Packbot?*
4. *Give me an overview of Washington, DC, focusing on evacuation routes.*

Example 1 involves assembly of facts that are well-known. Example 2 involves a particular training procedure that is currently in place. Example 3 involves a custom report that the training organization may not have prepared in advance. Example 4 focuses on a report of a particular type (a city tour), with a focus on a particular facet (transportation).

Although not explicitly addressed in the original proposal, it is crucial to allow for multimedia content (e.g. maps, photos) in the reports. Such content can be very effective in helping make information clearly and quickly understood to trainees. In addition, enhancing the task to allow for narrated briefings on a topic will add another dimension of difficulty to the challenge, although the first few systems to have such a capability have recently emerged (Andre et al. 2005).

The HAC proposal accommodates different aspects of reports and differing levels of system capabilities through the mechanism of Leagues (e.g., fact-based, historical narratives, comparative essays, etc.), with tasks of increasing difficulty over time. Consider a report on a city, e.g., Washington, DC. The report might include a history of the city, highlights of its museums and monuments, tours of particular neighborhoods, and information about entertainment, dining, and transportation. Such a report would include a variety of different genres (historical narrative, geographical information, listings, reviews, etc.). Basic reports of this kind could be easily created by assembling information from the Web, allowing for a suitable first year task. The task could be made substantially more difficult by requiring zooming in to particular topics in more detail, e.g., the Capitol, or the British siege of Washington, or by demanding a facet that requires reasoning and synthesis on the fly, e.g., transportation to and from health care facilities.

Such a city report can be tied to the application of *after-action analysis*. Consider a city report in the Historical Narrative League which zooms into a historical event, such as the 1814 siege of Washington, DC by the British. The events include several preventive measures taken: the Americans evacuated treasures from the White House, and also set fire to the Washington Navy Yard to prevent them from falling into British hands. The British, on arrival in the city, set the White House on fire, as well as the Capitol and other buildings. A challenging historical narrative could discuss whether these preventive measures were adequate, in other words, offer an after-action analysis of the government's response. Such *after-action reports* related to exercises or historical events are widely used throughout the government, and some of them are accessible on the Web. Reports which target more recent events are also of considerable interest, e.g., a report in the Comparative Essay league that compares the state and federal government's responses to hurricane Katrina.

After-action reports can be used as case-studies for use in training military and civilian personnel. For the topic of search and rescue missions in a part of a city affected by a crisis, the report might include information about availability of equipment such as robots that can help in this process. Equipment often is reused in such situations for new purposes. A technician might have received prior training on how to operate a particular robot, the Packbot, but the robot may not have been used outside in driveways and landscaped areas. Before sending the robot into the area outside a building, it would be helpful for the technician to have at least some understanding of the robot's operating characteristics and how the robot might function in the new area. Having a Handy Andy generated report on "Remote Controlling Robots on Uneven Surfaces", or "The Physical Characteristics and Limitations of the Packbot" could be quite useful. All these characteristics call for custom, on-demand, anytime, anywhere report generation.

3 Possible Metrics

The reports produced by artificial systems in the HAC will be compared against each other, as well as against reports produced by humans. The HAC also requires that today's systems be able to participate in the challenge, with increasingly stringent challenges in subsequent years. These characteristics of human comparison of system performance against human performance, in a progression of increasingly more difficult tasks, are shared by a number of evaluation activities that have been carried out in AI and language processing communities, e.g., question-answering evaluation in the Text Retrieval Conference (TREC 2004), and summarization evaluation in the Document Understanding Conference (DUC 2005). Based on the experience and lessons learned from such activities, we can consider a variety of methods of evaluation:

The HAC proposal suggests that subjective grading, as traditionally used in pedagogical situations, will be used to assign points to essays. Subjective grading of this kind is compelling and can easily be appreciated by the public. However, human graders can easily distinguish computer generated from human results, and tend to penalize the former. For example, in mixed sets where a subject sees both a human and a machine translation, human translations are rated higher in clarity and machine translations lower than in non-mixed sets (Falkedal 1991). The design of HAC could test whether mixed sets are a problem, and if they are, subjective grading could be used, if desired, without mixed-sets.

In addition to subjective grading, we would like to propose several other methods:

1. *Objective assessment* of human comprehension of reports. The hypothesis here is that a better-produced report will allow humans, on an average, to more accurately answer questions about the topic. This assessment doesn't directly test how well-written the report is; it's theoretically possible that a badly written report could result in adequate comprehension by a human. This assessment can be carried out by posing reading comprehension tests to humans who read the reports. The cost here is in the preparation of question and answer keys for each report topic. The metric used is percentage answers correct.

2. *Content-based comparison* of reports. System reports can be compared against human or other system reports using measures that assess similarity of content. These measures range from word- and phrase-level comparisons to comparisons at the level of concepts. Recent work in summarization evaluation (Nenkova and Passonneau 2004) has developed a weighting scheme for concepts based on the number of reference summaries which include them. The type of task will determine what sorts of comparison is appropriate; for example, if the report in the General League is to list facts,

the concepts being compared will be such facts. Also, depending on the task, the reports being compared are likely to diverge in their textual form; essays which reproduce similar web sources are likely to diverge less than those which inject their own content. Tools to help humans in any such comparison, such as the SEE summary element alignment editor (Lin 2001), as well as automatic comparison software such as ROUGE (Lin 2004) and BE (Hovy et al. 2005) have been explored extensively in the DUC conferences.

Automatic essay grading based on content analysis (Hearst 2000) (Burstein et al. 2003) (ETS 2005) extends the more generic content comparison methods with techniques based on linguistic features specifically tailored for essays and the topics and points contained in them.

The metrics used for content-comparison involve similarity scores of various kinds. Content-based evaluation can be more or less costly than objective assessment. The method depends on availability or construction of gold standards in the form of reference reports that can be compared against. One issue here is incorrectly penalizing a good report because it bears little resemblance to a reference report. The use of multiple reference reports, and an experiment correlating content-based scores with scores from subjective grading or objective assessment can help address this.

3. *Task-based evaluation* involves testing the utility of the report generator in some task. In the case of reports to support training, if the report involves generation of a short manual, one can assess how well trainees performed their training exercises using the computer-generated versus human-generated manuals. (Lennox et al. 2001) provide an example of such an evaluation in the case of generated brochures. Task-based evaluations can provide very useful feedback to funding agencies, but are less useful as diagnostics for developers. They are also relatively expensive to conduct because they require the involvement of human experts.

Metrics developed for HAC will need to have at least two desirable properties, following (Donaway et al. 2000): the ranking of reports produced by such metrics will need to remain stable across different sets of judges; and reports that are in fact similar should get similar ranks. Further, such metrics should address both 'informativeness' of material in the report as well as the overall coherence of the report in terms of organization, argumentation, fluency, etc.

4 Additional Issues

Progress evaluation: A general problem with competitions that offer tasks of increasing difficulty over time, is that progress is often difficult to measure. If the competition includes a core set of metrics that remain unchanged, in addition to others which may change, it is a reasonable challenge to require that performance on these metrics increase monotonically with task difficulty. However, particular tasks may need their own specialized metrics that provide insight into particular aspects of task performance, for example, metrics to assess degree of comparison.

Complexity of reports: Reports can vary considerably in terms of the degree of authorship provided by an agent. At one end of the spectrum, as expected in the Original Writing league, are reports which have a high degree of authorship; these are more expository and expressive of opinions, as essays and op-ed articles typically are. At the other end of the spectrum are extractive reports which recycle information which is already there on the Web. Between these two extremes, one can find reports that conform to some particular 'boilerplate', e.g., technical, military, or business documents where an overall structure and format is dictated by the business rules of the particular organization. These differences suggest an ordering of HAC reports in terms of complexity, with the less complex kinds being taken on in earlier years.

Complexity of topics: In addition to complexity associated with the type of report, the relation between the topic and the information available (in memory, or on the Web) is also important. The issue of what makes a particular topic 'hard' for a machine or a human in the HAC needs to be explored. In part, this will depend on the relationship between information in the topic and information used to construct the report. In information retrieval, one important predictor of topic difficulty for systems is the ambiguity of the topic with respect to the document collection being searched against. (Cronen-Townsend et al. 2002) have characterized the ambiguity of a topic with respect to a document collection based on a *clarity score* which measures the relative entropy between the language usage associated with the topic and the language of the collection. A less ambiguous query returns highly ranked retrieved documents that are on a single topic; those documents are characterized by unusually high probabilities for a small number of topical terms. While this notion may or may not carry over into HAC, performance on HAC may be used to explore and derive empirical measures for topic complexity/hardness.

Veracity of information: Another important issue is the veracity of the information in a report, especially when information from the Web is being used. Rather than restrict the competition to particular sources, veracity can be factored into the evaluation methods. When subjective grading is used to assess a report, points should be given for accuracy of information. Reading comprehension and content comparison suppose an answer key for the former and a reference report for the latter, both of which are required to be accurate. Likewise, task-based evaluation will require accuracy for correct execution of the task. Informativeness measures will have to factor in accuracy.

Plagiarism: The HAC competition will have to guard against plagiarism. While this is likely to be more of a problem in extractive, compiled reports, it could be a problem in any of the leagues, including the Original Writing and Polymath leagues. As (Clough 2003) illustrates, plagiarism can cover a wide range, from cases of lifting or paraphrasing language (words, phrases, passages, or entire texts) without acknowledgement, to more indirect cases, involving lifting of ideas or arguments without acknowledgement, or reliance on secondary sources while attributing only primary sources. Tools exist for detection of simple cases of plagiarism, and the HAC might consider using some of these. The HAC will benefit from contestants agreeing to an honor code and indicating that they have read and agreed with a statement warning against plagiarism. A concise statement of three simple rules for avoiding plagiarism, covering the use of language, ideas, and collaborative discussion or authorship in report-writing, can be found in (Cornell 2000).

Sharing: It is of course imperative that all participants in HAC use a common set of metrics so their systems can be compared; further, that these metrics be available, to the extent that they are automated, to them at any time. To compare different metrics and evaluation methods, it will also be useful to create a testbed where scoring methods and tools can be assembled and tested.

References

(Andre et al. 2005) André, E., Concepcion, K., Mani, I., and van Guilder, L, Autobriefer: A System for Authoring Narrated Briefings. O. Stock and M. Zancanaro, eds., *Multimodal Intelligent Information Presentation* , 143-158, Springer, 2005.

(ArmyStudyGuide-0024 2005) http://ppt.armystudyguide.com/index.html 0024.ppt

(ArmyStudyGuide-000144 2005) http://ppt.armystudyguide.com/index.html 000144.ppt

(Burstein et al. 2003) Burstein, J., Chodorow, M., and Leacock, C. Criterion: Online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence,*Acapulco, Mexico.

(Clough 2003) Clough, P. Old and New Challenges in Automatic Plagiarism Detection. http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf

(Cornell 2000) Office of the Dean of Faculty, Cornell University. *The Code Of*

Academic Integrity and Acknowledging The Work Of Others. http://web.cornell.edu/UniversityFaculty/docs/AI.Acknow.pdf

(Cronen-Townsend et al. 2002) Cronen-Townsend, S., Zhou, Y. and Croft, B. Predicting query performance. *Proceedings of the Twenty-Fifth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2002),* 299–306.

(Donaway et al. 2000) Donaway, R. L., Drummey, K. W., and Mather, L. A. A Comparison of Rankings Produced by Summarization Evaluation Measures. *Proceedings of the Workshop on Automatic Summarization*, 69-78.

(DUC 2005) Document Understanding Conference. http://www-nlpir.nist.gov/projects/duc/

(ETS 2005) e-rater. http://www.ets.org/research/erater.html

(Falkedal 1991) Falkedal, K. Evaluation Methods for Machine Translation Systems: An Historical Overview and Critical Account. Report 1991, ISSCO, Universite de Geneve.

(GAO 2003) Government Accountability Office. Defense Management: Army Needs to Address Resource and Mission Requirements Affecting Its Training and Doctrine Command *GAO-03-214 February 10, 2003.* http://www.gao.gov/docdblite/details.php?rptno=GAO-03-214

(Hovy et al. 2005) Hovy, E. and others. BE: Basic Element Analysis. http://www.isi.edu/~cyl/BE/

(Lennox et al. 2001) Lennox, A.S., Osman, L. M., Reiter, E., Robertson, R., Friend, J., McCann, I., Skatun, D., and Donnan, P. The Cost-Effectiveness of Computer-Tailored and Non-Tailored Smoking Cessation Letters in General Practice: A Randomised Controlled Trial. *British Medical Journal* 322:1396-1400.

(Lin 2001) Lin, C-Y. Summarization Evaluation Environment http://www.isi.edu/~cyl/SEE/

(Lin 2004) Lin, C-Y. ROUGE: a Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out* (WAS 2004), Barcelona, Spain, July 25 - 26, 2004.

(Nenkova and Passonneau 2004) Nenkova, A. and Passonneau, R. Evaluating Content Selection in Summarization: the Pyramid Method. *Proceedings of the Human Language Technology conference / North American chapter of the Association for Computational Linguistics* (NAACL-HLT 2004).

(TREC 2004) Voorhees, E. M. Overview of the TREC 2004 Question Answering Track. http://trec.nist.gov/pubs/trec13/t13_proceedings.html

Acknowledgments

Contributor	Affiliation	Email
Christy Doran	The MITRE Corporation	cdoran@mitre.org
Inderjeet Mani	The MITRE Corporation	imani@mitre.org
Beatrice Oshika	The MITRE Corporation	bea@mitre.org
Laurel Riek	The MITRE Corporation	laurel@mitre.org

Handy Andy
The DARPA Essayist

- Produce a multi-page essay on any subject
- Compete against students at all levels of education
- All essays graded by teachers
- Cash prizes for all winners, bonuses for beating humans

Competition Overview

- Automated AI systems compete against invited human contestants (students from DC area schools, colleges)
- Multiple leagues stress particular technologies; e.g., fact-based, compare/contrast, construct history, etc.
- Explicit evaluation criteria; essays graded by teachers who are blinded to source of essays
- Human and AI winners get prizes, AIs that beat humans get *big* prizes

Evaluation Criteria – General

- All contestants may compile essays from online sources, scored on several, weighted criteria, for example:
  - Content should be relevant
  - Certain facts or events must be reported
  - The essay should incorporate multiple sources, not cribbed from one source.
  - The essay should not be redundant.
  - The essay should be well organized.
  - The essay should cite its sources.
  - The essay should contain some language generated by the contestant, not copied from the web.
  - The essay should make an original assertion, something not explicitly stated in any of the sources

Leagues

- Fact Based – fact-based essays on subjects selected by teachers
- Historical/Narrative – combine several sources but get the narrative structure right
- Comparative Essay – compare and contrast things, ideas, events
- Original Writing – write something original, don't just compile others' writing
- Polymath – write from knowledge in memory, limited access to sources on the web
- Human/Computer Team – human and machine produce better essays, faster, than humans alone
- Experts – specialize in very broad areas of human expertise (e.g., National Library of Medicine online sources)

Multi-Year Development Plan

- Each year's problem is harder, but only slightly out of reach
- Early years allow compiling essays from online sources
- Later years require essays to include original writing, increasingly based on stored knowledge, containing original inferences and opinions, etc.

Administration

- Steering committee sets rules, creates special leagues, each year to raise the bar
- Approach Turing test functionality via graded challenges, each just *slightly* out of reach
- Steering committee works with educators and other stakeholders to develop essay topics (hidden until competition) and grading criteria
- The task is always to produce an essay about any subject; never limiting the test to component technologies or specialized subjects

Why Handy Andy is a *Cognitive* Grand Challenge

- Success depends on *comprehending* the essay question and material on the Web – making the Web a resource for machines to read
- Steering committee makes the test more cognitive each year, emphasizing stored knowledge, inference, original writing, forming opinions, comparing and contrasting, etc.
- Systems must *learn* from material on the web, eventually they'll have to write essays based only on what they know

Credits

- Paul Cohen, ISI (Technical Lead)
- Dave Waltz, Columbia
- Kathy McKeown, Columbia
- Beatrice Oshika, MITRE (Support)
- Laurie Damianos, MITRE (Support)

Handy Andy Report Generator:
The DARPA Essayist

FY'06
Proposal

- Challenge: Generate report on any topic per user request, using on-line resources

- Feasibility pilot to focus on constrained application with
  - Clear report structure
  - Static corpora of resources
  - Reference set of good examples

- Possible application: After Action Reports
  - Synthesis from various sources
  - Narrative of events over time
  - Genuine generation, not cut-and-paste
  - Evaluation against reference examples

- MITRE LOE: 3 SM

colleagues, Donald Michie (who later founded the Department of
talking often about the possibility of computing machines (1) learn
possible solutions, guided by rule-of-thumb principles (Michie is
search', a heuristic being any rule-of-thumb principle that cuts dow
Park Turing illustrated his ideas on machine intelligence by referen
comment is chess programming (in particular minimax and best-fir

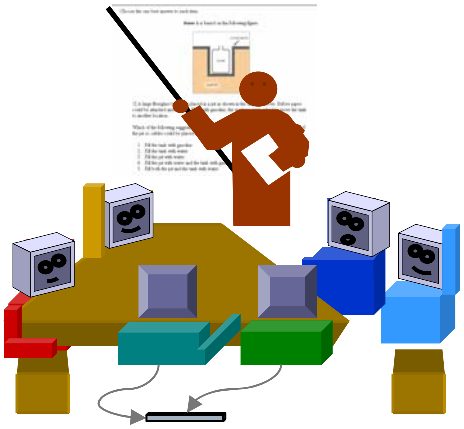
Electromechanical versus Electronic Co

With some exceptions - including Babbage's purely mechanical en
machines were electromechanical. That is to say, their basic compo
relatively slowly, whereas the basic components of an electronic co
operate extremely fast. Electromechanical digital computing machi
at Harvard University, George Stibitz at Bell Telephone Laborator
Zuse belongs the honour of having built the first working general
functioning in 1941. (A program-controlled computer, as opposed
plugs etc.)

Relays were too slow and unreliable a medium to make practicable
valiant efforts in this direction by Aiken). It was the development
possible.

Reading to Learn
The Scholastic Grand Challenge

An autonomous system will learn from a textbook and answer the questions in the book chapter-by-chapter



This Grand Challenge focuses on having systems “learn” by reading a textbook and passing incremental, chapter-by-chapter tests. This idea arose at the January DARPA Grand Challenge Workshop in discussions of “reading to learn”. Michael Witbrock and Lynette Hirschman put together a one-day follow-on workshop in Seattle, hosted by Bill Dolan at Microsoft.

The next step in developing the Scholastic Grand Challenge would be to write a prototype "Young Computer's First Reader." This would be a textbook written in simple English, focused on a constrained and structured subject matter (perhaps evolution or geology), consisting of at least four chapters, plus associated problems, along with an evaluation methodology. This could be used to attract participants to demonstrate the feasibility of the Scholastic Grand Challenge. Our plan is to work with Michael Witbrock and possibly other participants in the Scholastic Grand Challenge Working Group, in order to develop the reader and test performance of one or two current systems on such a sample reader.

Attached documentation:

Challenge Description
Scholastic Grand Challenge
Short, 2-page write up describing challenge as developed during the workshop

Briefing (used in AAAI Presidential Address)
Output of workshop brainstorming

FY’06 Proposal
Plans for follow-on work [6 SM – other participants funded separately]

Description of the proposed Scholastic Grand Challenge
August 25, 2005
Lynette Hirschman and Lisa Ferro

Background

This description is based on a workshop held in Redmond, WA on June 22, hosted by Bill Dolan at Microsoft, organized by Lynette Hirschman and Michael Witbrock, with the following participants: Judy Bundy, Murray Burke, Jaime Carbonell, Bill Dolan, Sanda Harabagiu, Chris Manning, Andrew McCallum, and Jean-Michel Pomerade.

The attached slides outline the proposed Scholastic Grand Challenge. The objective of this Grand Challenge is to create an autonomous system that can read and understand a textbook on any subject – the system capabilities are assumed to be generic, such that the system, like a person, could tackle any subject provided that it is pitched at an "appropriate level." For example, it might be necessary to start out with beginning science before tackling college chemistry. The system will be designed to learn from the textbook, and to demonstrate its understanding by answering the questions in each book chapter, as it progresses from chapter to chapter, just as would be required of a human student taking the course.

Overview

The task is framed so that the system's primary source of input will be the textbook; it may be necessarily in the early stages to write such a textbook ("A Young Computer's First Reader"). The system, like a human learner, will have access to external sources while it is trying to master the material. And like a human student, it will be allowed to have access to a (human) teacher for limited periods, to answer questions and to provide feedback on how to answer questions on homework or on a test. This feedback will be given in a virtual classroom setting – which means that all participating systems will "hear" both each other's questions and also the teacher's answers. The time allocated for such interaction will be limited, and rules will be created to ensure fairness, so that each system gets to ask an equivalent number of questions. These rules and time limits will also ensure that systems cannot elicit fragments of a knowledge based from the teacher. All system-teacher interaction will be conducted in natural language.

Rules

Course Procedure

Systems participating in the Scholastic Challenge will have no prior knowledge of the subject matter of the course. The "course" will be administered based on a textbook, one chapter at a time. There will be a specified amount of time to "study" each chapter. This will be followed by a virtual classroom session with a human teacher, who will answer questions raised by the student-systems. The systems will then be asked to work a set of problems (the "testing" phase). The teacher grades the systems on the problems and returns the test; systems scoring below a cut-off threshold on the test will not be permitted to advance to the next chapter. Following the testing, there will be another question-answer session with the teacher.

Pre-Course Mode:
Before beginning the course, the system will have access, as would a student, to whatever resources it can find, and it may download materials at liberty.

Course Mode
Once the course begins, the system will be sequestered, without network connectivity. The course will be designed to build test progressive competency, so that an instructional unit is based on the previous units, and the tests associated with a unit may include questions that test back to earlier chapters, or draw on material in earlier chapters.

It is assumed that this test will be administered "off-line." This is done for two reasons: to make sure that the system has internalized the materials – it cannot just try to do a rapid search of the Web to find matching question/answer pairs; and second, the system cannot access other humans to answer the questions for it (by, for example, entering a chat room and asking the question in the chat room).

Milestones

Year 1: Proof of Concept
The systems will read a book written specifically for this task: A Young Computer's First Reader"

Year 5: Pass a test for 10-year olds.
The winning system will score a passing grade (65% or better) in a course designed for ten year olds.

Year 10: Score 85% of human performance on standardized college level test.
The system will read a textbook for a course and score at least 85% of human performance on the test provided at the end of the course. The test will be a standardized test, to provide human performance benchmarks for comparison.

Defense Advanced Research Projects Agency DARPA

Scholastic Grand Challenge

Participants:

Judy Bundy
Murray Burke
Jaime Carbonell
Bill Dolan
Sanda Harabagiu
Chris Manning
Andrew McCallum
Jean-Michel Pomerade

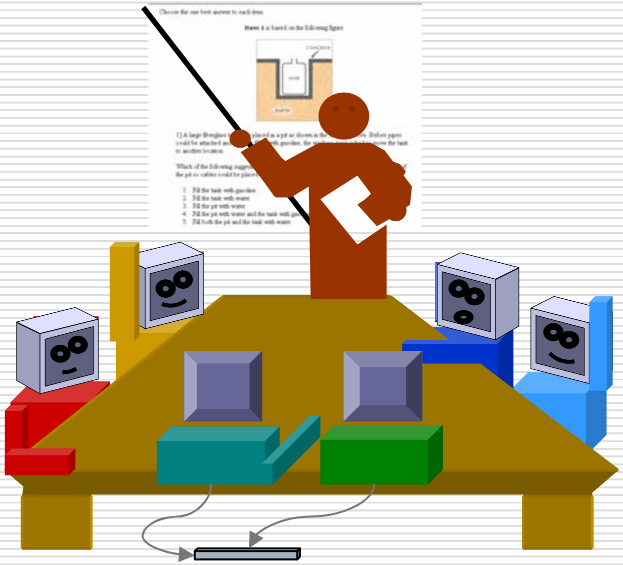
Organizers:

Lynette Hirschman
Michael Witbrock

June 22, 2005
Redmond, WA

INFORMATION PROCESSING TECHNOLOGY OFFICE

Read and answer questions from a textbook


Choose the one best answer to each item.
Item 1 is based on the following figure
CONCRETE
TANK
EARTH
1] A large fiberglass t... placed in a pit as shown in the ... Before pipes could be attached an... with gasoline, the ... to move the tank to another location.
Which of the following sugges... the pit so cables could be placed...
1. Fill the tank with gasoline
2. Fill the tank with water
3. Fill the pit with water
4. Fill the pit with water and the tank with ga...
5. Fill both the pit and the tank with water

An autonomous system will learn from a textbook and answer the questions in the book chapter by chapter

Grand Challenge Rules

- ***Entrants do not know subject area of test***
- ***Test sequence will be administered to all entrants one chapter at a time, after reading the chapter.***
- ***Time allowed for each chapter's reading and testing will be specified***
- ***At the end of each chapter's test, systems scoring below threshold are not allowed to continue***
- ***Interaction with a teacher occurs in an open "classroom" before and after each chapter's test****
- ***Additional sources not allowed during the test sequence except for natural language interaction with the teacher***

* Open classroom interaction means that all participating systems will have access to the questions asked by the other systems as well as the answers provided by the (human) teacher to those questions; all interaction is in natural language

Milestones

- Year 1: *Proof of concept: systems read a book written specifically for this task: "A Young Computer's Reader"*
- Year 5: *Winning system scores at 65% on a test at the fifth grade level**
- Year 10: *Read a textbook and score 85% of human performance on a standardized test on any college level topic*

* There are multiple dimensions of complexity for the middle year goals that still need to be worked out – see notes for further details

Test Plan

- Starting point: anything on the web ok to download
  - Before knowing specific subject matter
  - On the assumption that the test uses non-factoid questions
- Systems are sequestered (no network connectivity) as soon as learning-testing begins
- Test is designed to test progressive competency
  - Assume instructional units build on one another
  - Questions in later chapters test back to earlier chapters
- Testing cycle
  - Systems read a chapter of book
  - There is a period of (natural language) dialogue w teacher
    - All systems have access to that dialogue (the “classroom”)
  - Systems take test at end of chapter
  - Systems get their scores (and answers)
  - Low scoring systems may be eliminated
  - Systems are allowed a period of questions to understand test results (via the “classroom”)
  - Go on to next chapter

Scholastic Grand Challenge:
A Young Computer's First Reader

FY'06 Proposal

- Develop a primer to test feasibility of Scholastic Grand Challenge
  - 4+ chapters of material
  - Problems (and solutions)
  - Evaluation methodology
- Demonstrate performance of systems
- Team: Witbrock, 1-2 other members
- MITRE LOE: 6 SM