Technical forum

Machine intelligence and the Turing Test

Alan Mathison Turing, the British mathematician, philosopher, and logician, proposed in 1950 that if a computer could successfully mimic a human during an informal exchange of text messages, then, for most practical purposes, the computer might be considered intelligent.[1] This soon became known as the Turing Test (TT), most typically conducted as anonymous exchanges of English-language text between computers. A panel of judges poses questions to the contestants in order to determine which are human and which are programs. There have been many such experiments since Turing proposed the challenge, but there is considerable disagreement as to what passing the test means, and whether passing it tells us much at all.[2]

We use the TT here as a means of identifying artificial intelligence (AI) technologies that will have a pivotal role in creating more intuitive machine-human interactions. We have chosen six technologies supporting certain computer behaviors that could significantly increase the practical value of computers. In what follows we summarize and editorialize on where each of these technologies stands today, relying heavily on the findings of the conference and workshop "Machine Intelligence and the Turing Test"[3] held last year at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York.

The first technology we address is *natural language understanding* (NLU). We confess to a particular passion for this part of AI, so fundamental for communication and yet still full of nuances, poorly understood, and hard to symbolize.[4] Even if computers could understand plain English, this would just be the beginning.

Our second technology is *machine reasoning* (MR). The TT judges ask the contestants questions intended to flush out the mere mechanical responses of a computer. To fool the human judges, a computer will need to provide reasonable answers, answers that are relevant within the context set by earlier exchanges.

0018-8670/02/$5.00 © 2002 IBM

Since TT questions can be about any subject, contestants need a very significant knowledge base covering a wide range of human activities such as sports, politics, health, and food. In addition to the rational, objective knowledge about ourselves and our society, the computer's knowledge base needs to also include "rules of thumb," myths, "old wives tales" and urban legends, as well as the complex relationships between facts, theories, conjectures, and judgments. We need *knowledge representation* (KR) technology to represent this information in all its complexity.

Creating the on-line body of knowledge would be itself a daunting task, and *knowledge acquisition* (KA) is our fourth AI technology with direct application for any would-be TT winner. The manual effort required to capture this knowledge would be enormous, so computers need to be programmed to listen and learn automatically.

Two less prominent technologies complete our set. There is growing interest in the dynamics of dialog and the role of identity in computer-human exchanges. This has led to theories and experiments in *dialog management*,[5] and to experiments in making computers react to human emotions.[6]

Natural language understanding

Natural language understanding (NLU) is the technology enabling computers to extract meaning from text—easy and natural for humans, but notoriously difficult as computation. NLU is a key component of software that can accept commands and queries from humans in their own language and produce answers whose meanings are automatically extracted from electronic texts. Even limited comprehension has significant business benefit. For example, today's imperfect approximations to NLU are widely used in: abstracting of trends and important events from news sources; summarizing vast repositories of text; and supporting natural language queries for accessing online help.

Expectations for NLU are well beyond the limited capabilities of the first-generation technologies that are built into current search engines, classification engines, and information extraction systems.[7] Current techniques do not handle semantics very well because of the many-to-many mappings between syntactic forms and semantic meanings. One form can have many meanings (e.g., "time flies like an arrow"); whereas the same meaning can be paraphrased in different forms ("announce/unveil/come out with a new product"). This is usually no problem for humans because we interpret meaning relative to context—prior knowledge shared by author and reader (or speaker and hearer).

The grand challenge of NLU is to simulate the human capability to:

- Create a store of prior knowledge (see the section on knowledge representation, later)
- Create a representation of the meaning of the current text
- Integrate this representation into the knowledge store (see the section on knowledge acquisition, later)

A representation of the meaning of text is created "bottom up" (small semantic units are fitted together into larger representations) and "top down," by homing in on the specific topic the document discusses. Significant progress has been made with the semantics of the smallest units, such as people, places, organizations, and temporal expressions (successfully embedded in IBM text analysis products[8]) and, increasingly, relations in the sentence that hold among these are also recognized (e.g., announce [IBM, hard disk drive]). Both symbolic and statistically based machine learning techniques have been successful and continue to develop. Advances continue on disambiguation of common words[9] and in translation between pairs of languages.

Progress has been slower beyond the word and sentence level. One well-known requirement, for example, is pronoun resolution—determining what a word such as "it" or "they" refers to. Algorithms were developed over ten years ago and for a while they improved, but they seem to have reached a plateau. The next major advance in NLU will probably come from the use of inferences based on knowledge of the world. For example, in the text "IBM has unveiled the Ultrastar* 36Z15 . . . It is the fastest in the industry," we need to know that IBM is a computer manufacturer, that computer manufacturers produce computers and parts, and that machines have properties such as speed to determine that "it" refers to the disk drive rather than to IBM.

The complete semantic task includes understanding the relations among entities discussed in the text, the actions and events they are engaged in, and temporal and causal sequences. There has been some success in understanding the main events (*who* did *what* to *whom*) in specific and narrow domains,[10] but expanding to larger domains remains a challenge.

Statistical "top down" techniques for topic identification have been developed mainly within the information retrieval community, with notable successes by IBM.[11] Current classifiers group large collections of documents into more specific topics, automatically producing a taxonomy like that of the Yahoo!** service.[12] These techniques continue to be crucial for identifying the "aboutness" of a document.

The progress of NLU to date has been encouraging in the areas of syntactic parsing, language-pair translation, semantic analysis in narrow domains, and statistically based information retrieval. Now is the time to concentrate on a deeper semantic understanding of text in larger domains. The domain-independent and complete NLU required for TT-like tasks will remain elusive for many years, but incremental progress can be made, and measured, within broadly defined domains and with respect to specific tasks.

Machine reasoning

The human capability for reasoning is another important aspect of our intelligence that machines have not fully captured. What we call "reasoning ability" is in fact a bundle of different abilities, such as:

- Simple inference within the KR system. For example, if we are told that Clyde is an elephant, we can deduce that he probably has four legs and that he needs food, water, and oxygen to survive.
- Search within the KR system. For example, find an African animal that is large, gray, and four-legged, with big-floppy ears and a long prehensile nose.
- More difficult inference. Some kinds of inference, such as theorem-proving, are much more difficult for humans than the kind of KR-based inference described above. There seems to be a qualitative difference in the effort required, and it may be that different mechanisms come into play.
- Planning and problem solving. In domains such as chess and factory scheduling, computers already exhibit problem-solving performance that is better than human performance. However, humans still excel at tasks requiring broad, diverse knowledge, flexibility, and the ability to learn, generalize, and transfer skills from one domain to another.
- Plan recognition and the ability to reason explicitly about plans. Humans can create plans; they can also recognize and explain what another person is trying to do. They can examine their own plans and explain costs, risks, and alternatives to clients or coworkers.
- Creativity. Some say that creativity is just competent problem solving that happens to lead to a surprising result, and in that sense, machines can sometimes be creative, but we believe fundamental elements of creative problem solving are still missing.
- Applying recipes. It appears that we humans store our knowledge of procedures in the form of recipes or scenarios, rather than as rigid programs. This same knowledge can be used to produce new plans and to recognize the structure and components of other plans that we encounter. This is another kind of knowledge that we must represent and store effectively.

The expert-system tools of the 1980s and early 1990s developed much of the basic machinery for machine reasoning. The Soar[13] system, developed at Carnegie Mellon University by the late Allen Newell and his students, explored ways of combining rule-based problem solving with powerful learning and chunking mechanisms, so that the system's performance would improve over time. However, much still remains to be done to make this problem-solving more flexible and use knowledge of all kinds to guide the problem-solving process.

Knowledge representation

For problems that require breadth of understanding—what we sometimes call "common sense"—current computing systems fall far short of human ability. The most critical missing piece is the ability to deal with large amounts of knowledge of many kinds, and to make that knowledge effective in perception and problem solving.[14]

Many kinds of knowledge are required for human-like capability. Predicate calculus can in principle be used to represent all the types listed below, but efficiency concerns push the system to more specialized representations for some of these types:

- Declarative statements
- Linguistic knowledge
- Procedural knowledge
- Naive physics
- Recognition knowledge
- Tactile/kinesthetic knowledge
- Visual knowledge
- Social knowledge

It seems unlikely that any single approach to knowledge representation (KR) will adequately cover all of these areas. Merely *representing* and *storing* each kind of knowledge is not sufficient; we must also make the knowledge *effective*. Each kind of knowledge requires appropriate representation, machinery for efficient search and inference, and some way to acquire and digest knowledge. Then all the knowledge types must fit into an architecture that allows them to work together effectively.

Declarative knowledge figures most prominently in the TT, which is focused on natural language input and output. There are several approaches both inside and outside IBM (most notably the Cyc** knowledge base[15]) for building a declarative KR system with powerful search and inference capabilities. These systems include a large base layer of knowledge that spans most domains—physical objects, materials, people, organizations, common actions and behaviors, and so on. They are then enriched with more specialized knowledge for each domain of interest. Although specialists may be required to build the most fundamental "roots of the universe" knowledge, it needs to be easy for nonspecialists to extend the knowledge base.

Knowledge acquisition

A system needs hundred of thousands to several million knowledge elements to approximate the knowledge of a human being. The challenge is to automate the KR process using a variety of techniques, including "learning" (as in statistical modeling and machine learning), in both supervised form (where the answer is provided to the learning algorithm) and unsupervised form (where the system observes the data without knowing the answer and has to infer it). To illustrate, a team at the University of Pennsylvania manually parsed a million words and provided parse trees for about 40000 sentences. Their *Treebank*[16] has been used by many researchers worldwide to create and improve parsers for broad domain English for various applications. Similarly, the creation of annotated databanks for other purposes will be a key ingredient in improving the state-of-the-art of the component technologies. At the other extreme is the knowledge in the Cyc knowledge base—a million facts manually entered over 15 years with 450 person-years of effort. One Cycorp researcher, when pressed, estimated that Cyc contains perhaps 2 percent of the required knowledge. But Cyc is at an inflection point and can start exploring methods to automate the KA and extract knowledge from the tens of terabytes of on-line text available on intranets and the Internet.

Another crucial aspect of KA is that new knowledge is needed on a daily basis, so the process of acquiring it has to be intrinsic to any system to keep up with the demands of deployed applications. The efficient creation of a whole cycle of knowledge update, from statistical learning to manual acquisition, is key to managing the total cost of operating these "intelligent" systems and a fundamental activity in creating the technologies.

Dialog management

A dialog is a sequence of interactions between participants with a shared context and a shared set of goals. Dialog management refers to the analysis of user utterances in the context of the current discourse, figuring out an appropriate response, and conveying it to the user.

Existing dialog systems differ in the degree to which users can take the initiative and steer the conversation. Directed dialog systems force users to constrain their input and stay on predefined dialog paths. Mixed initiative systems constrain the user input only when it is imperative to have a clear understanding of user intentions (e.g., to get confirmation before executing a stock purchase). Over the last 20 years much progress has been made in replacing rigid hierarchical directed dialog systems with mixed initiative systems offering a more open mode of conversation.[17] Of course, we are still a long way away from completely open user initiative dialogs as characterized by the TT.

Narrow domain dialog systems are finding widespread use inside and outside IBM for a whole gamut of applications ranging from buying stocks to finding information to directory assistance. The big challenge confronting dialog researchers is to build systems that can converse with humans about topics not limited by a few predefined forms or templates. Another issue is one of meta-knowledge: Does the system know how much it knows or does not know? Can its behavior degrade gracefully when it encounters the limits of its knowledge? Promising new approaches combining statistical information retrieval, information extraction, and dialog systems may help answer the above questions.

Emotion

An intelligent system may benefit from having access to information about the intentions or the emotional states of humans. Systems can already begin to recognize and use affective information in a variety of forms. Ultimately, systems might do this by observing facial expressions and body language, recognizing patterns in physiological measures, analyzing the affective content of text and speech, and inferring emotional states from interactive behavior.

Understanding what people are feeling can guide the computer interaction with its users, from help messages to the use of true-to-life computer-generated speech. Human conversational partners who do not give affective cues—tone of voice, choice of words, gestures—are perceived as flat and unresponsive partners. Without affective cues, misunderstandings abound. However, today it is difficult to find reliable indicators of emotion that are not obscured by individual differences in the way humans experience and express emotion.

From the user's perspective, giving our systems the ability to understand and appropriately respond to affective content may raise computers from their current socially inept role to a role more consistent with human conversational expectations.

Epilogue

Most of the AI technology used in products today is based on linguistic models of knowledge and linguistic processing techniques. An important next step will be to combine linguistics with a large database of assertions representing commonsense facts about the world, in the hopes of producing much more human-like reading and conversational systems. Systems like Cyc and NETL[18] aim to give machines common sense by amassing a large collection of commonsense assertions, then reasoning about them with the help of a logic engine.

Although there is no doubt as to the business value of this direction, it is interesting to note that memorized linguistic assertions play only a small part in intelligent human behavior, and formal first-order logic has no significant role in human cognitive processing. This opens the way for approaches that seek to incorporate aspects of humans-as-systems beyond the purely linguistic, both to enhance the linguistic skills of our systems and to develop skills that may not be accessible with purely linguistic approaches. One such approach creates learning machines that discover the facts by themselves much as humans do. Common sense is acquired by the machine sensing its environment directly and learning from that experience. In order to learn about catching a baseball, the machine might interact with the physical world through sensors and effectors designed for vision and motion. Common sense involved in a particular linguistic domain might be acquired by reading texts and conversing with humans.

Other emerging areas include the study of machines that are based on models of brain behavior, and although these approaches are relatively immature and more speculative in nature than the technologies based on linguistic models, they may allow us to make new, significant, and perhaps revolutionary progress in AI. On the other hand, although we doubt that AI technology based on purely linguistic models will ever be mistaken for a human, or ever pass an unconstrained TT, we believe that it will make the largest AI contribution to business and society for years to come.

*Trademark or registered trademark of International Business Machines Corporation.

**Trademark or registered trademark of Yahoo! Inc. or Cycorp, Inc.

Cited references and notes

1. A. M. Turing, "Computing Machinery and Intelligence," *Mind* **59**, 433–460 (1950).
2. A formal TT yearly contest, sponsored by Hugh Loebner and The Cambridge Center for Behavioral Studies, accords a $2000 prize and medal to the most human-like computer contestant. Among the most well-known critics of the contest is Marvin Minsky, a professor of computer science at the Massachusetts Institute of Technology (MIT) who is considered by many to be "the father of AI." Minsky has wittily sponsored a "Minsky Loebner Prize Revocation Prize."
3. In May 2001 the IBM Academy of Technology and IBM Research held a conference and workshop on "Machine Intelligence and the Turing Test." The conference speakers were: Jaime Carbonell (Carnegie Mellon University), Barbara Grosz (Harvard University), Jerry Hobbs (SR International), John Laird (University of Michigan), Doug Lenat (Cycorp, Inc.), Michael Mauldin (Virtual Personalities & Carnegie Mellon University) and Rosalind Picard (MIT). The IBM organizing team comprised: Joe Bigus, Ian Brackenbury (chair), Scott Fahlman, Joe Londa, Clifford Pickover, Yael Ravin, and Alan Webb. The authors gratefully acknowledge the contributions to the workshop report by: Nancy Alverado, Scott Fahlman, Charles Peck, and Steve R. White, fragments of which are given here in condensed form.
4. Y. Bar-Hillel, "Automatic Translation of Languages," *Advances in Computers*, D. Booth and R. E. Meagher, Editors, Academic Press, New York (1960). This classic article on the NLU challenge is still often cited today.

5. See, for example, the *Proceedings of the ANLP/NAACL 2000 Workshop on Conversational Systems*, Seattle, WA, May 2000.
6. R. W. Picard, "Toward Computers That Recognize and Respond to User Emotion," *IBM Systems Journal* **39**, Nos. 3&4, 705–719 (2000).
7. "Enterprise Portals: Web Interfaces for Employees, Partners, and Customer Communities," META Group (September 1999).
8. Intelligent Miner for Text, IBM Corporation, http://www.ibm.com/software/data/iminer/fortext.
9. See http://www.itri.brighton.ac.uk/events/senseval.
10. See http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html.
11. This technique has been successfully used by customers and partners to route customer e-mail to the right expert, based on the e-mail content. See http://domino.research.ibm.com/comm/wwwr_thinkresearch.nsf/pages/email198.html.
12. Yahoo! is a large Web site featuring vast, manually maintained, taxonomies covering all manner of general-interest topics such as DIY (Do-It-Yourself), medicine, cooking recipes, arts, and sciences. Find out more at http://www.yahoo.com.
13. J. E. Laird, A. Newell, and P. S. Rosenbloom, "Soar: An Architecture for General Intelligence," *Artificial Intelligence* **33**, 1–64 (1987).
14. We are indebted to S. E. Fahlman of IBM and Carnegie Mellon University (http://www-2.cs.cmu.edu/~sef/) for contributions to the sections on KR/KA and machine reasoning.
15. The Cyc knowledge base is an extensive knowledge base and inference engine system, with a core of over 1 000 000 hand-entered assertions (or "rules") designed to capture a large portion of what we consider knowledge about the world. The effort was pioneered by Doug Lenat in 1984. For more information, see the Cycorp, Inc. Web site at http://www.cyc.com/.
16. See the Linguistic Data Consortium page at the University of Pennsylvania, http://www.ldc.upenn.edu/.
17. See http://www.darpa.mil/ipto/research/com/index.html.
18. S. E. Fahlman, *NETL: A System for Representing Real World Knowledge*, MIT Press, Cambridge, MA (1979).

Accepted for publication May 16, 2002.

I. Brackenbury
IBM Software Group
Somers, New York

Y. Ravin
IBM Corporate Technology
Somers, New York