featured research: bridging the worldwide healthcare gap

ICSI’s Srini Narayanan, head of the AI Group, is collaborating with UC Berkeley and the Hesperian Foundation, based in Berkeley, to bring accessible health care information to rural areas in developing countries. He and Matt Gedigian, a master’s student at the iSchool at UC Berkeley, are working on a Wikipedia-like Web site that would combine semantic search, digitized voice recordings, and multi-lingual translations of health care manuals to bring relevant information to users in poor communities around the world.

A Need for Digital Materials

Narayanan became interested in the work of the Hesperian Foundation in 2007 after talking with Madelaine Plauché, a postdoc in the Speech Group at the time. Plauché worked with TIER, a group of the UC Berkeley Computer Science Department that researches new technologies for developing regions. Through TIER, she had become familiar with Hesperian, a publisher of health care manuals designed for poor and semi-literate users. Hesperian’s primary publication, Where There Is No Doctor, has been in use since 1977, when it was translated from the Spanish text Donde No Hay Doctor. The manual lays out, in simple terms and with frequent illustrations, how to prevent, diagnose, and treat diseases. It has been translated into 80 languages and is used in 108 countries. The foundation also offers books on dental care, community and environmental health, women’s health, and a wide array of other issues. In 2008, the foundation shipped out 20,000 copies of its materials and users downloaded 300,000 free copies from its Web site.

After Narayanan read the text, the question for him became, he says, “Can we help this book be used by more?” Sending a copy of the book costs $30. While some copies are offered for free through the foundation’s Gratis Book Fund and all materials are offered in PDF form on its Web site, Narayanan saw the need for an easy-to-use version of the online books, a cross-referenced and searchable database that would deliver the right information to the right user at the right time.

Plauché was involved in the original plan to create a Hesperian Digital Commons, a Web site similar to Wikipedia that would tailor Hesperian materials to an individual user’s needs. In the last year, Narayanan and Gedigian have begun the work. The Rockefeller Foundation and the Bill & Melinda Gates Foundation, through the latter’s grant to Hesperian, funds the work.

Making It Relevant

At the root of the research is the question of how to “customize large amounts of information to one user’s needs,” Narayanan says. He and Gedigian are working toward searchable versions of Hesperian’s materials. So far, the researchers have uploaded and annotated ten Hesperian books in English of the 20 or so available. They are working to upload all Hesperian books in English to understand how difficult it will be to do the same for Hesperian books in different languages. They have also uploaded and begun tagging versions of Where There Is No Doctor in Spanish and Tamil, a language spoken in India, Sri Lanka, and Singapore.

To this end, Narayanan and Gedigian have created an ontology, a structured set of concepts and their relationships. For the researchers, this means thinking about the relationships between a disease, what causes it, and what cures it. The work is similar to that done in the FrameNet group: the researchers must think in terms of “frames,” structures of related concepts. The researchers must know the relationship between...
How far that little candle throws his beams! So shines a good deed in a naughty world.

- William Shakespeare (via Portia, in the Merchant of Venice)

In recent weeks, we have been saddened to read (and see) the devastation in Haiti following their earthquake. Natural disasters occur frequently, given the wide range of dangers in our world; but they are sometimes exacerbated by human action or inaction. This brings to mind the devastation in our own country that followed Hurricane Katrina. Many have cited the insufficient designs and their implementations for flood control as being the real cause of the major loss of life and property in that disaster. And in Haiti’s tragedy as in many other disasters, human choices (domestic or international) could be seen as the primary cause for much of the suffering.

But many other disasters are slower in occurring. Insufficient public health resources are the cause for far more deaths worldwide than terrorism or war. Even in our own highly developed and wealthy country, it has been estimated that 45,000 people die annually because of the lack of health insurance. Even more subtle is the effect of poor early childhood education, including of preschool age children; this leads to, among other things, poor language skills, which in turn leads to a larger poverty-stricken population; and consequently, many reductions in the overall quality of life in our country.

I hope that the reader excuses these departures into topics that I certainly have no credible background in. There is, however, a connection to science and technology, and in particular to what is now happening at ICSI. We certainly are a technical establishment, with goals that are primarily focused on progress in computer science (and on international collaboration – see the statement on our home page). As I’ve noted before in one of these columns (March, 2005), while technological and scientific progress is certainly not inherently evil (as would be suggested by the "Frankenstein complex"), it also is not destined to only promote good things – fire is used for both cooking and arson. Those of us who work in technical areas have to accept that the basic science and technology we develop will be used in ways that we will not approve of. Nonetheless, we continue, generally holding the belief that the problems of the future (and the present!) cannot be handled without continuing scientific progress.

And yet, for many of us, this is not enough. We want to help with “good deeds.” We give to charitable organizations, volunteer time and money for causes we believe in, and sometimes even lead in public action. These are personal choices ICSI staff have made outside of the Institute per se; but within ICSI’s official program, we also work toward what we see as the public good in specific, targeted ways. In our Algorithms Group, for instance, we have worked for years on finding genomic associations with disease. In our relatively new Vision Group, facial identification algorithms are being developed in order to reunite lost children with their families (a task that sometimes took as long as six months after Katrina). The Networking Group has long had a focus on developing the underlying science necessary to defeat attacks on the cyber structure of our country, which increasingly is at the core of our overall infrastructure, for instance the energy supply system. The group is also concerned with attacks by cybercriminals that can be particularly devastating for the financial security of vulnerable parts of society. Speech researchers have been looking at the nature of calls to emergency services, in hopes of increasing their efficiency, and have also begun studying the problems of improving the language skills of low socioeconomic status preschool children with the help of some current and upcoming technologies. And our AI Group researchers have been working on improving the access to public health information in poor populations; this is the topic of the lead article in this Gazette and is an increasing focus at the Institute.

Finally, for many years ICSI has supported the Berkeley Foundation for Opportunities in Information Technology (BFOIT). BFOIT is the primary Berkeley effort to support the advancement of underrepresented minorities and women into computer science and related fields. This is a very local and very direct way for Berkeley scientists and engineers to help in a wonderful cause. I’d like to give credit here to ICSI’s Founding Director, Jerome Feldman, for involving ICSI in this worthy enterprise. Please see www.bfoit.org to learn more about BFOIT.

I’d like to close with one last pointer: each year, our annual activities report for the previous year is published; it can be found in pdf form through the pull-down menu on our Publications page. If you are interested in looking at the range of ICSI research for a year, please take a look. It usually comes out within a month after the publication of the March Gazette.
The first annual call for Brazilian applications to ICSI's Visitor Program has been issued, and the research visitors are on track to begin late spring 2010.

ICSI held its annual BEARS Open House February 11. Professor Kistie Asanovic, head of the Architecture Group, presented his group’s work on photonic-based memory, Collin Baker led a demonstration of the latest improvements to the FrameNet Project, and Luke Gottlieb demonstrated Joke-O-Mat, a project by members of the Speech Group that won the Multimedia Grand Challenge at the ACM Multimedia conference in October 2009. Scientists from all groups presented posters summarizing recent results of various research projects. The open house is held annually in conjunction with UC Berkeley EECS Annual Research Symposium (BEARS).

Eran Halperin of the Algorithms Group and Tel Aviv University received one of six 2010 Krill Prizes for Excellence in Scientific Research. The prizes are given annually to faculty members at universities in Israel, and come with a $10,000 award.

Liz Shriberg of the Speech Group was selected as one of six Fellows of the International Speech Communication Association (ISCA) in 2009 for her significant contributions to the field of speech communication science and technology.

Board of Trustees member Wolfgang Wahlster will be the first chairman of the TZS Scientific Advisory Board. TZS is a new German company using data analysis to address urban management challenges. Wahlster was chosen because he is “one of the world’s preeminent artificial intelligence experts,” according to Peter Liebhart, TZS’s CEO.

Jerome Feldman of the AI Group received a 2009 Berkeley Citation. He was cited for work in computational cognition and neural networks, leadership in interdisciplinary research, and support of the Berkeley Foundation for Opportunities in Information Technology (BFOIT), an ICSI program that supports underrepresented minorities and women in science and engineering.

ICSI Speech Group alum Matthew Aylett has developed text-to-speech software that allows film critic Roger Ebert, who lost his voice box four years, to speak in an approximation of his own voice. “It still needs improvement,” said Ebert in a report aired on CBS, “but at least it sounds like me.” Aylett’s Scotland-based company, CereProc, broke down voice recordings from Ebert’s film commentary into individual sounds, which are reassembled when Ebert types a sentence into the software. Ebert can also adjust the emphasis and intonation of the synthesized sentences. CereProc offers a number of other voices produced from text, including voices with Irish, Scottish, Southern English, and American accents. Ebert’s new voice debuted on the Oprah Show in March.

Leah Hitchcock-Ybarra of the Admin Group gave birth to a baby girl, Isabella Marie Ybarra, on Friday, February 26. Isabella weighed 6 lbs, 10 oz at birth. Congratulations to Leah and her family!

Speech Group scientist Dilek Hakkani-Tür was quoted in The LA Times explaining some of the technical difficulties of transcribing spoken language to text (Speech and Handwriting: Where’s the App” Tracy Wood, Feb. 7, 2010). Scientists in the Speech Group are actively researching ways to make machine transcription of human speech more accurate, as well as creating algorithms to assist computers in making sense of the resulting transcriptions.

Charles Fillmore of the FrameNet Group was the invited speaker at the FrameNet Masterclass and Workshop. The workshop, held in Milan, Italy, in December 2009, was held in conjunction with the 8th International Workshop on Treebanks and Linguistic Theories.

Rainer Böhme, a DAAD-sponsored postdoctoral Fellow in the Networking Group, was selected as a co-winner, with Tyler Moore of Harvard University, of the first Gordon Prize in Managing Cybersecurity Resources by the University of Maryland’s Robert H. Smith School of Business. The prize is named after Lawrence Gordon, an Ernst and Young alumni professor and a pioneer in cybersecurity economics research.

Dan Klein, an affiliate of the Speech Group and a professor at UC Berkeley, has won an Ōkawa Foundation research grant award for 2009. This award is given annually to select scientists in the U.S. and Asia and includes a $10,000 unrestricted grant.
aspirin and fever, and the relationships of those terms to a disease like the flu. They then tag these terms so that a search engine will return relevant results.

Much of this work can be done automatically. When text is scanned into the Wiki, it can be checked against lists of diseases and drugs. “Malaria,” for example, will be automatically tagged as a disease and “penicillin” as a drug. Some symptoms can also be automatically tagged. More complicated are the causes and consequences of diseases: many of these must be annotated by hand.

Terms are tagged with a property so that they will show up in property searches, which look not just for a word but for the relationship the word has to its page. For example, an article on the flu contains the phrase, “Aspirin or acetaminophen helps lower fever and relieve body aches and headaches.” This page will show up under a property search with “treated with” as the property and “aspirin” as the value. Such searches are helpful when, for example, users need to know what they might have if they are running a fever, or what viruses they might catch through contact with blood.

After tagging the books in English, the researchers want to do the same in five other languages. In addition to Spanish and Tamil, there is preliminary work being done in Pular, a West African language, and Narayanan hopes to work with French and Wolof, another West African language, in the future. French and Spanish, says Narayanan, would be the first or second language of most of the people he wants to reach.

The researchers chose Tamil as one of the first languages to upload to the digital commons because the script and grammar are dramatically different from those used in English and Spanish. Using Tamil, Narayanan hopes to work out low-level technical difficulties with Unicode, the standard that seeks to represent every character in every human language in a standardized way. While they work, the researchers are developing technical tools and extensions to open source software to make their work easier. Many of these tools make uploading PDFs onto their Web site easier, for example by allowing them to move captions and figures around. It is possible that in future, they will develop tools to allow them to use optical character recognition for languages not currently recognized.

**Looking Forward**

According to a study from New York University, over 90 percent of those who bought Where There Is No Doctor did so to teach or train others, and almost 95 percent bought the book to promote health within a community. While the initial set of online materials is designed to deliver highly personalized information to individual users, Narayanan and his colleagues are also looking forward to community-wide uses. They hope one day to be able to automatically produce flyers, pamphlets, and radio broadcasts with information gleaned from multiple books to respond to specific health threats and every day health problems. Further down the road, the researchers also hope to look into whether video, audio, or text information is most persuasive, particularly for users who are semi-literate.

Narayanan’s work is part of a larger effort by Hesperian, funded by the Bill & Melinda Gates Foundation, to update their materials for the 21st century. Hesperian has asked its partners, who translate its materials into local languages, to comment on how efficient the materials are. There will be larger field tests in 2010.

But in the meantime, Narayanan says, “If we can help more people reach [these materials], I construe that as being a great thing. It makes sense to do, and we believe we have something to offer.”
Vern Paxson has been leading one wing of ICSI’s Networking Group for almost a decade, pushing back against spam, credit card theft, malware, and network attacks. One central tenet of the group’s work is empiricism; they believe that in order to truly understand a problem, you must first measure it. Two such recent projects of the group have been receiving well-deserved attention lately, and we’re happy to describe them both here. Judo is a joint project between ICSI and UC San Diego that has come up with a novel way to fight spam, and Netalyzr is an easy-to-use and very robust network health analysis tool.

Judo

Most spam on the Internet comes from botnets, which are networks of subtly compromised computers that can be hijacked remotely for malicious purposes. The Judo team – Christian Kreibich, along with Paxson and Nicholas Weaver – has studied the pattern of some of these botnets by deliberately infecting computers with bots, and running them in a controlled environment (made up of both real and virtual machines) to learn their behavior.

The researchers have discovered that these botnets have a very specific, limited range of spam messages they can produce, because they use templates to generate their e-mail messages. The researchers have been able to infer the contents of these templates by examining the output of each botnet; moreover, they can use this template against its own botnet and actually filter out all spam it generates with virtually no false positives.

This approach has significant advantages over existing spam filtering techniques. Current implementations are not necessarily able to catch all spam from any given botnet, and can also accidentally mark messages as spam because of their content, even if they are legitimate e-mails – including the researchers’ own e-mails discussing their spam filtering techniques.

Some recent press coverage has painted the results in an almost holy light. Paxson characterizes some of this exaggeration as a reflection of Judo’s emotional resonance; it uses spam’s own technology against itself. Kreibich points out that this is still a reactive measure, and therefore can never be perfect. Judo is only effective once a botnet has been captured and analyzed, and only effective against mail generated by that specific template. This method will enable us to catch up quickly to spam in the wild, but will not be able to prevent it.

Netalyzr

Net neutrality has been a hot topic lately, touching on subjects such as censorship, peer-to-peer networking, Internet accessibility, and more. Up until now, no one has had any quantitative sense of just what sort of neutrality – or lack thereof – the Internet currently offers. Weaver and Kreibich have developed a tool that can detect the health and openness of a network connection. In the planning stage, the team decided to make a single comprehensive diagnostic tool that was both easy to use and robust enough to detect a very wide range of network disruptions. They use a Java applet to check the consistency of IP addresses, correct DNS resolution and hidden proxies, port filtering, IPv6 support, and much more.

Netalyzr has already tested over 100,000 Internet sessions, and the data it provides have often proven surprising. It has found network configurations at several sites where devices in the network have imposed traffic controls that operators themselves didn’t know about; it has even identified this kind of connectivity restriction at networking conferences. The fact that even networking conferences don’t always have totally open connections shows how much net access is being limited, because of deliberate censorship or simple hardware configuration. Netalyzr has also revealed problems with the handling of fragmentation, poor DNS performance, and deliberate manipulation of DNS results.

While the debate about the future of the Internet continues, these concrete measurements of network neutrality, security, and performance provide an invaluable resource for determining what the Internet is actually like today.

ICSI acknowledges generous NSF support for this work; grant CNS-0722035 on the work of Netalyzr, and grants NSF-0433702 and CNS-0905631 on the work of Judo.
Using data collected in 110 languages over 30 years, ICSI’s Paul Kay and Richard Cook, along with other researchers, have supported and expanded Kay and Brent Berlin and Paul Kay’s 1969 hypothesis that the world’s languages conform to a universal pattern as they develop names for colors. In The World Color Survey, released in February by Stanford’s CSLI Publications, the researchers argue on the basis of the new data that there are indeed universal constraints on the ways in which languages name colors and how color naming systems evolve.

When Berlin and Kay’s Basic Color Terms was published in 1969, the prevailing doctrine held that local cultural factors determine how colors are named in different languages, specifically that “each culture has taken the spectral continuum and has divided it upon a basis which is quite arbitrary.” Berlin and Kay found that color terms in different languages tend to be based on a limited number of colors, and that languages acquire new names for colors in a partially fixed order.

In 1976, Kay, Berlin, and the late William Merrifield established the World Color Survey (WCS) project to investigate these hypotheses more fully and to respond to (legitimate) criticisms. Critics argued that Basic Color Terms had been based on a small number of speakers, mostly of written languages spoken in developed societies, who were also speakers of European languages and who were currently living in the U.S. The WCS project gathered data on 110 unwritten languages in their own localities to test Berlin and Kay’s hypotheses. While finding general confirmation of the hypotheses of universality and evolution of color naming systems, the project has also made several significant revisions.

The new book revises the model of how color-naming systems evolve. The current model admits the possibility of a language with only two color terms, as previously documented, but the survey did not document any such languages. In such a language, there is one term comprising white and all the “warm” colors and another comprising black and all the “cool” colors. A three-term language separates a “warm” term from the earlier warm-white term, producing a system of white, warm, and black-cool terms. Further stages of development successively divide these categories until a six-term language comprises separate terms for black, white, red, yellow, green and blue. Terms such as brown, purple, orange, pink, and gray tend to be added late in development, but not at precisely predictable points.

A separate chapter is devoted to the analysis of each of the 110 color terminologies. Each chapter contains several tabular and diagrammatic summaries of the data as well as an overall descriptive summary and evolutionary categorization of the color naming system. The 110 individual language chapters are preceded by chapters dealing with the history of the subject and the methods of the survey.

The book is authored by Paul Kay, Brent Berlin, Luisa Maffi, William Merrifield, and Richard Cook.

visiting scholars

Since its inception, ICSI has had a strong international program consisting primarily of ties with specific countries. Current formal agreements exist with Brazil, the European Union, Finland, Germany, Spain, and Switzerland. In addition, we often have visitors associated with specific research and projects.

AI
Terry Regier
Hiroaki Sato
Carlos Subirats (Spain)
Kyoko Ohara

ALGORITHMS
Jan Baumbach (Germany)
Oliver Kramer (Germany)
Jörg Lässig (Germany)
Shuai Cheng Li
Richard Röttger
Benjamin Satzger (Germany)
Dirk Sudholt (Germany)

ARCHITECTURE
Chris Batten
Sascha Hunold (Germany)

CAMPUS AFFILIATION / OTHER
Oliver Günther (Germany)
Anniina Hutunen (Finland)
Christoph Göbel (Germany)
Tommi Lampikoski (Finland)
Jouni Similä (Finland)
Marko Turpeinen (Finland)
Holger Ziekow (Germany)

NETWORKING
Maria Bermudez-Edo
Rainer Böhme (Germany)
Joos-Hendrik Bose (Germany)
Joakim Koskela (Finland)
Boris Nechaev (Finland)
Jarno Rajahalme (Finland)
Pasi Sarolahti (Finland)
Kyriakos Zarifis
Youquan Zheng

SPEECH
Carolin Mende (EU - Amida)
Carlos Vaquero

VISION
Nicolas Cebron (Germany)
Carl Henrik Ek (EU - Amida)
Sanja Fidler
Yang Hua
Ruihua Ma


THE INTERNATIONAL COMPUTER SCIENCE INSTITUTE (ICSI) is the only independent, non-profit US lab solely conducting open, non-proprietary, pre-competitive research in computer science. Affiliated with the University of California campus in Berkeley, ICSI’s mission is to further research in computer science through international collaboration, and further international collaboration through research in computer science. ICSI provides a haven for computer scientists to conduct long-term research without commercial limitations and with few faculty pressures. ICSI has significant efforts in these major research areas: Internet research, including Internet architecture, related theoretical questions, and network services and applications; theoretical computer science, including applications to bioinformatics; artificial intelligence, particularly for applications to natural language understanding; natural speech processing; computer architecture; and computer vision.

BOARD OF TRUSTEES
Javier Aracil
Hervé Bourlard
Deborah Estrin
Graham Fleming
Greg Heinzinger
Clifford Higgenson
Nelson Morgan
David Nagel
Peter Norvig
Prabhakar Raghavan
Stuart Russell
Scott Shenker
Eero Silvennoinen
David Tennenhouse
Wolfgang Wahlster

The ICSI Gazette is published twice a year by the International Computer Science Institute (ICSI). All material in The ICSI Gazette is copyrighted by ICSI and should not be used without explicit prior written permission.

EDITOR-IN-CHIEF
Maria Quintana
SENIOR EDITOR
Nelson Morgan
DESIGN
envision ID
COPYWRITER
Chris Switzer
LAYOUT & ADD’L COPYWRITING
Jacob Wolkenhauer
PHOTOGRAPHY
Jaci Considine