SCIENCE

Published by the **American Association for the Advancement of Science (AAAS),** *Science* serves its readers as a forum for the presentation and discussion of important issues related to the advancement of science, including the presentation of minority or conflicting points of view, rather than by publishing only material on which a consensus has been reached. Accordingly, all articles published in *Science*—including editorials, news and comment, and book reviews—are signed and reflect the individual views of the authors and not official points of view adopted by the AAAS or the institutions with which the authors are affiliated.

The American Association for the Advancement of Science was founded in 1848 and incorporated in 1874. Its objectives are to further the work of scientists, to facilitate cooperation among them, to foster scientific freedom and responsibility, to improve the effectiveness of science in the promotion of human welfare, to advance education in science, and to increase public understanding and appreciation of the importance and promise of the methods of science in human progress.

Membership/Circulation
Director: Michael Spinella
Deputy Director: Marlene Zendell
Member Services: Rebecca Dickerson, *Manager*; Mary Curry, *Supervisor*; Pat Butler, Helen Williams, Laurie Baker, *Representatives*
Marketing: Dee Valencia, *Manager*; Jane Pennington, *Europe Manager*; Hilary Baar, *Associate*; Angela Mumeka, *Coordinator*
Research: Renuka Chander, *Manager*
Business and Finance: Jacquelyn Roberts, *Manager*
Administrative Assistant: Nina Araujo de Kobes
***Science* Member Services**
Marion, Ohio: 800-347-6969;
Washington, DC: 202-326-6417
Other AAAS Programs: 202-326-6400

Advertising and Finance
Associate Publisher: Beth Rosner
Advertising Sales Manager: Susan A. Meredith
Recruitment Advertising Manager: Janis Crowley
Business Manager: Deborah Rivera-Wienhold
Finance: Randy Yi, *Senior Analyst;* Shawn Williams, *Analyst*
Marketing: John Meyers, *Manager;* Allison Pritchard, *Associate*
Traffic Manager: Carol Maddox
Recruitment: Terri Seiter, *Assistant Manager*; Pamela Sams, *Production Associate*; Debbie Cummings, Celeste Miller, Rachael Wilson, *Sales*; Charlotte Longhurst, *European Sales*
Reprints Manager: Corrine Harris
Permissions Manager: Lincoln Richman
Exhibits Coordinator: Arlene Ennis

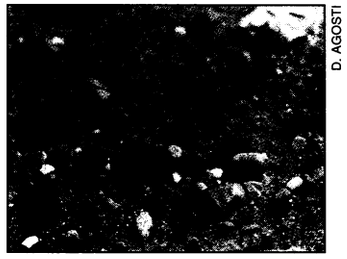
PRODUCT ADVERTISING SALES: East Coast/E. Canada: Richard Teeling, 201-904-9774, FAX 201-904-9701 • **Midwest/Southeast:** Elizabeth Mosko, 312-665-1150, FAX 312-665-2129 • **West Coast/W. Canada:** Neil Boylan, 415-673-9265, FAX 415-673-9267 • **UK, Scandinavia, France, Italy, Belgium, Netherlands:** Andrew Davies, (44) 1-457-838-519, FAX (44) 1-457-838-898 • **Germany/Switzerland/Austria:** Tracey Peers, (44) 1-270-760-108, FAX (44) 1-270-759-597 • **Japan:** Mashy Yoshikawa, (3) 3235-5961, FAX (3) 3235-5852
RECRUITMENT ADVERTISING SALES: US: 202-326-6555, FAX 202-682-0816 • **Europe:** Gordon Clark, (44) 1-81539-5211, FAX (44) 1223-302068 • **Australia/New Zealand:** Keith Sandell, (61) 02-922-2977, FAX (61) 02-922-1100
Send materials to *Science* Advertising, 1333 H Street, NW, Washington, DC 20005.

Information for Contributors appears on pages 112–114 of the 6 January 1995 issue. Editorial correspondence, including requests for permission to reprint and reprint orders, should be sent to 1333 H Street, NW, Washington, DC 20005.
Internet addresses: science_editors@aaas.org (for general editorial queries); science_letters@aaas.org (for letters to the editor); science_reviews@aaas.org (for returning manuscript reviews); membership@aaas.org (for member services); science_classifieds@aaas.org (for submitting classified advertisements)

LETTERS

Desert Ants

The Random Sample item "Hot ants" (14 Apr., p. 207) featured a recent article by W. J. Gehring and R. Wehner describing a remarkable thermal adaptation of desert ants (*1*). It is suggested that these animals may synthesize heat shock proteins (Hsp's) for protection even before leaving their nests to forage in the desert heat, that is, that they prepare "a pre-emptive strike." It is important, however, to distinguish experimentally between the normally abundant and essential heat shock cognate (Hsc) proteins (Hsc70) and the heat inducible Hsp70 family members. Hsc70 is more highly constrained evolutionarily than Hsp70, suggesting that they are functionally distinct (*2*). An alternative interpretation of the data of


D. AGOSTI

Hotfooting it. How does this ant, *Cataglyphis bombycina*, survive the desert heat?

Gehring and Wehner is that the ants have unusually high levels of the normal housekeeping protein Hsc70, which appears to be only modestly augmented at high temperatures by inducible Hsp70 in these animals. There must be more to the desert adaptation, as the characteristic in question is shared by another ant species from temperate woodlands. Perhaps desert ants are smart enough to return to the nests before their proteins melt down, and perhaps Hsc70 makes a contribution to their high critical thermal maximums and thermal resistance of protein synthesis reported. Such a response would be less amazing than the mental image of ants turning on their heat shock genes in anticipation of "hotfooting it" across the desert sands, but interesting nonetheless.

Lawrence E. Hightower
Department of Molecular and Cell Biology, University of Connecticut, Storrs, CT 06269–3044, USA

References

1. W. J. Gehring and R. Wehner, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 2994 (1995).
2. C. N. White, L. E. Hightower, R. J. Schultz, *Mol. Biol. Evol.* **11**, 106 (1994).

■

Oak Ridge's Strengths

Andrew Lawler's article about the Galvin report, "Report to stress research over close ties to industry" (News & Comment, 27 Jan., p. 446), includes a table describing DOE's multipurpose laboratories. The table indicates that the research focus of Oak Ridge National Laboratory (ORNL) is "nuclear physics, ion-beam, neutron scattering." This description does not reflect the scientific and technical competencies of ORNL as presented to, and used by, the Galvin task force. While ORNL has long-standing and pioneering research activities in the areas listed, it also addresses a wide range of scientific and technical challenges as directed by DOE and other federal agencies. The five broad programmatic themes of ORNL's research activities are energy production and end-use technologies; biological and environmental sciences and technology; advanced materials synthesis, characterization, and processing; neutron-based science and technology; and computational science and advanced computing.

Johnnie B. Cannon
Director,
Office of Planning and Management,
Oak Ridge National Laboratory,
Post Office Box 2008,
Oak Ridge, TN 37831–6251, USA
E-mail: cannonjb@ornl.gov

■

Performance of Text Retrieval Systems

In his article "Gauging similarity with *n*-grams: Language-independent categorization of text" (10 Feb., p. 843), Marc Damashek states that his *n*-gram information retrieval system, Acquaintance, performed "on a par with some of the best existing retrieval systems," on the basis of results of the third Text Retrieval Conference (TREC-3) (*1*).

The TREC Program Committee objects to this conclusion. On the basis of average precision (the generally accepted measure

of performance), Acquaintance was ranked 22nd out of 23 systems on the TREC-3 ad hoc task, and 19th out of 21 systems on the TREC-3 routing task. (Some groups presented multiple runs with their systems, and not all groups did both tasks).

Whereas it is true that Acquaintance did well on a few of the queries, this can be said for most of the participating systems. The nature of information retrieval dictates that any reasonable system will do well on some queries; the best systems do well on most of the queries. The conservative tests of statistical significance presented at the conference clearly show that Acquaintance belongs in the bottom set out of five sets of systems.

While the *n*-gram approach of Acquaintance may be valuable in some environments, it has not yet been shown that it is even an average system for general information retrieval.

The TREC Program Committee: Donna Harman, *Chair, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA;* ***Chris Buckley,*** *18 Dickinson Court, Plainsboro, NJ 08536, USA;* ***Jamie Callan,*** *Department of Computer Science, University of Massachusetts, Amherst, MA 01003–4610, USA;* ***Susan Dumais,*** *Bellcore Laboratories, Morristown, NJ 07974–0636, USA;* ***David Lewis,*** *AT&T Bell Laboratories, Murray Hill, NJ 07974–0636, USA;* ***Steve Robertson,*** *Department of Information Science, City University, London ECLV OHB, United Kingdom;* ***Alan Smeaton,*** *Dublin City University, Glasnevin, Dublin 9, Ireland;* ***Karen Sparck Jones,*** *New Museum Site, Computer Laboratory, Cambridge University, United Kingdom;* ***Richard Tong,*** *Verity, Inc., 1500 Plymouth Street, Mountain View, CA 94043, USA.*

References

1. D. K. Harman, Ed., *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST SP 500-225, National Institute of Standards and Technology, Gaithersburg, MD, 1995).

The content analysis of stored texts and documents represents a crucial component in automated document retrieval and question-answering systems. In principle, such a content analysis ought to be based on sophisticated syntactic and semantic methodologies leading to a complete understanding of the respective text content. Unfortunately, complete linguistic text understanding systems are too brittle to be usable in open-ended text environments with unrestricted subject coverage. For this reason, word- or phrase-based analysis methods are used in operational retrieval systems where each text item and each search request is represented by sets of possibly weighted terms (words and word phrases) that collectively characterize text content (*1*).

Damashek proposes to simplify the text analysis step still further by giving up word and phrase recognition and using instead sets of overlapping strings of *n* characters (*n*-grams) derived from the existing document texts for the representation of text content. Although a decomposition of running texts into overlapping *n*-grams is in principle trivial, such a representation is too rough and too ambiguous to be usable for most purposes. Necessarily, a decomposition of text words such as HOWL into HOW and OWL raises the ambiguity of the text representation and lowers retrieval effectiveness. Damashek shows that the *n*-gram characterization is good enough for certain rough classification tasks. But for more demanding tasks, such as information retrieval, the *n*-gram analysis can lead to disaster.

The results obtained with Damashek's Acquaintance system in the TREC studies (where many different retrieval approaches compete against each other in an objective retrieval evaluation environment) are sufficient proof of the inadequacy of the *n*-gram text representation method (*2*). For ad

hoc retrieval, where individual user queries are processed against existing text collections, Damashek states that "Out of 50 queries considered, the measured precision [of the *n*-gram Acquaintance system] exceeded the median (across 34 participating systems) only twice." He goes on to say that his system does, however, perform "on a par with some of the best existing retrieval systems" in the routing task, where certain relevant items are identified a priori, and the task consists in retrieving new items similar to the previously known relevant ones.

Actually, Damashek's system scored below the median for 22 queries out of 50, and for an additional 18 queries out of 50, the system scored only as well as half the other participating systems. A reasonable performance was obtained by Damashek for only 10 queries out of 50. Overall, Acquaintance was ranked near the bottom (28th out of 34 systems participating in the TREC studies) with an average precision performance of 0.2641, compared with a precision performance of 0.4068 for the best performing system.

Gerard Salton
Computer Science Department,
Cornell University,
Ithaca, NY 14853, USA

References

1. G. Salton, J. Allan, C. Buckley, A. Singhal, *Science* **264**, 1421 (1994).
2. D. K. Harman, Ed., *Overview of the Third Text REtrieval Conference (TREC-3)* (NIST SP 500-225, National Institute of Standards and Technology, Gaithersburg, MD, 1995).

Response: I sincerely regret any offense occasioned by my characterization of Acquaintance's performance on the TREC routing task in my article and offer my specific apologies to any participant so offended. TREC is a state-of-the-art showcase of retrieval methods. Using the performance data available at the time my article was written (the data presented in the article), I interpreted the performance cup as being "half full." I viewed a presentable showing as an indicator of above-average performance in the world at large. This ultimately led me to the unfortunate choice of words, "on a par with some of the best existing retrieval systems," which made no explicit mention of the world at large. There is no question that certain systems participating in TREC consistently exhibit superior performance and, at TREC-3, Acquaintance was not one of them. My statement could easily have been perceived as suggesting performance comparable to that of the best systems participating *in* TREC, a claim that would have been ludicrous in view of the evidence presented in the article itself.

Salton raises several points that demand close scrutiny. As it pertains to document content analysis, the contention that an exhaustive list of *n*-grams is "too ambiguous to be usable for most purposes" can be shown to be untenable. By construction, every *n*-gram in a document shares $n - 1$ characters with each of its immediate neighbors. For $n > 1$, this redundancy of the *n*-gram representation makes it possible to unambiguously reconstruct a significant fraction of the document's text, implying that much of the information conveyed by the document remains intact. Given no more than a list of the *n*-grams in the document, one can easily show, by automatically identifying and concatenating those *n*-grams that can be uniquely paired with a predecessor or successor, that on average approximately 60% of a typical narrative English document can be reconstructed for $n = 5$ (with low variance across documents). This figure rises to 70% for $n = 6$ and 80% for $n = 7$. The argument can be applied to the example cited by Salton, showing that the trigrams HOW and OWL uniquely specify the four-character

"document" HOWL. Historically, work in this field was limited almost exclusively to trigrams ($n = 3$), which typically support reconstruction of less than 10% of the text, so skepticism is understandable. A categorical dismissal based on "the inadequacy of the n-gram text representation method," however, is unwarranted. No such reconstruction is possible based solely on a word list.

The system described in my article is not speculative; Acquaintance has been successfully applied to a wide variety of difficult tasks since 1991. Its demonstrated competence, language independence, and garble resistance should answer any concerns about "disaster."

The approach taken by this system—involving an unfamiliar combination of n-gram processing and vector-space transformation—may well be counterintuitive. Also, performance tradeoffs differ somewhat from those normally encountered in this field. My hope is that serious discussion of these matters will lead to a fuller understanding of how such techniques can complement more familiar approaches.

Marc Damashek
Department of Defense, Fort George G. Meade, MD 20755, USA

Amateur Fossil Hunting

I read Carol Potera's article "Amateur fossil hunters dig up trouble in Montana" (News & Comment, 14 Apr., p. 198) with concern. While the specific incident described is troubling, tens of millions of dollars worth of one-of-a-kind fossils sit in museums throughout the world because of the generosity of many amateur fossil hunters. Untold other fossils sit rotting in the sun because of their location on government controlled land and the fact that collecting by amateurs is illegal. If it weren't for an amateur fossil hunter and her rock shop, Jack Horner might not have achieved such prominence as a paleontologist.

Don Miller
2615 Longwood Drive, Wilmington, DE 19810, USA

■

High-Frequency Outer Hair Cell Motility: Corrections and Addendum

In our report "High-frequency motility of outer hair cells and the cochlear amplifier" (31 Mar., p. 2006) (*1*), we proposed an answer to the question of what powers the cochlear amplifier (*2*) at high frequencies. In subsequently analyzing a generalized network representing n excited and m nonexcited outer hair cells (OHCs) in the electrical environment of the organ of Corti, we obtained some unexpected results that make it necessary to correct our report and to offer new information.

The voltage representing the intracellular potential of an excited OHC is inversely proportional to n and has a steep midfrequency drop off with a high-frequency terminal slope of unity (approximately 6 decibels per octave). Simple unity slope is obtained only if $n = 1$, that is, if there is only one active hair cell in the network. It has been argued (*3*) that this assumed unity-slope high-frequency roll off of the cell's receptor potential makes it difficult to see how the voltage-driven electromotility might be expressed at high frequencies. We now see that the midfrequency roll off is, in fact, faster, and the response is smaller than that expected for a single cell. As a consequence, it would appear even less likely than previously assumed that high-frequency electromotility driven by the cells' own receptor potential is the basis of the cochlear amplifier. In line with our previous argument, the voltage gradient between the extracellular fluid spaces of the scala media