Feature

Tough Mining

The challenges of searching the scientific literature

Steven Dickman

The standard "front end" for biomedical literature search is MEDLINE and its Entrez query system. Huge, well-managed, and nearly exhaustive, MEDLINE and its 11 million references provide incredible ease and facility for anyone who can type a Boolean query. Though not quite a parallel for Google—which runs a kind of popularity contest for Web links in real time—the Entrez search has opened up the literature to anyone with a Web browser. To those who grew up chasing citations and papers through the aisles of a scientific library, Entrez is a dream come true.

And yet. Suspend disbelief and imagine for a moment a kind of literature search dream-tool. "Find me all references citing my gene of interest," you could ask. But why stop there? "Find me all references citing some or all of my four genes of interest with expression or in vitro data." And then, "Bring up the text of the paragraph in which these citations occurred so I can view them in context. And do it in real time."

Tools that can perform such searches would go beyond Google because they avoid the repetitiveness involved in multiple searches. And they would go beyond Entrez because they would search the entire medical literature in full-text format and not, as MEDLINE does, just the abstracts. Furthermore, they would go beyond both types of searches in that they would be at least somewhat intelligent.

Such text-mining efforts are the next frontier for both academic and commercial groups that have sprung up from Pasadena to Boston to Tel Aviv.

Steven Dickman is a freelance writer and president of CBT Advisors in Cambridge, Massachusetts, United States of America. E-mail: sdickman@cbtadvisors.com

DOI: 10.1371/journal.pbio.0000048

DOI: 10.1371/journal.pbio.0000048.g001

Figure 1. Barely Getting below the Surface
The four levels of information retrieval: Google and MEDLINE both use keywords to direct a searcher to documents. But the next level has been tough to crack. Improved software would allow biologists to jump from the Web or MEDLINE to specifics with a single query. (Adapted with permission from the MITRE Corporation.)

But how realistic is this venture? Text-mining and its more universal relative "information retrieval" are still in their infancy. The first paper on text-mining for biology was published only in 1997. Furthermore, because biological text-mining comes so close to the challenge of comprehending human language—arguably the most complex invention in the history of the planet—it is what computer scientists call a "hard problem." So even here, at the embryonic and fun stage in this technology's history, the outcome and especially the timing of improvement are impossible to predict.

Build or Buy?

Language-processing software tools have been successfully applied in text-mining of nonscientific sources, especially to newswire content. Computer programs can already perform all three levels of text-mining (Figure 1) effectively: *retrieving* documents relevant to a given subject; *extracting* lists of entities or relationships among entities; and *answering* questions about the material, delivering specific facts in response to natural-language queries.

Information retrieval and extraction can be performed on news data at success rates of 90%–95%, says Lynette Hirschman, a structural linguist. Question-answering has been reported in the literature at 85% accuracy, she notes, which is "amazingly good." The question is, how soon can these levels be achieved for biology?

Good thing for biologists that Hirschman has turned her energies in their direction. Hirschman works in Massachusetts at MITRE Corporation, a government-funded institution that pursues projects in the national interest, be they in defense and intelligence or, as in the case of text-mining, "anywhere we can move an entire field forward," says Hirschman.

The good news from news-mining is that improvement seems to arrive in direct proportion to the time and energy expended by the research community. Similar improvement has occurred in speech recognition by computers, she adds (Figure 2). When people took successively harder problems and worked on them for four or five years, she explains, it caused error rates to drop, as a rule, by a factor of two every two years.

One might think tackling the biomedical literature would be relatively easy, remarks Hirschman: biology jargon has a lot of prefixes and suffixes, which can be parsed more easily than verbs and adverbs; it is highly regular, with Greek-letter add-ons to gene or protein names signifying relatives or subtypes of the original proteins; and there are many resources available, such as databases and ontologies linking different biological terms.

Information retrieval and extraction can be performed on news data at success rates of 90%–95%. The question is, how soon can these levels be achieved for biology?

But whereas extraction of person and place names from news text routinely reaches 93%, results in biology remain mired in the 75%–80% range. "It's a little depressing," warns Hirschman. "Even something as simple as a slash may imply two different entities or a single compound."

A chorus of assent greets her observation. Programmers eager to codify the rules of biology have been stymied by what one bioinformaticist calls "a sea of exceptions."

Moreover, there is a chronic lack of data that have been "marked up" by software or humans to indicate the roles played by some of the key words. This marking-up process, however it is done, is crucial for machine-learning tasks. Getting these data is both hard and expensive, says Hirschman. To move biology text-mining forward, she believes, requires organizing different academic and commercial groups so that they are at least working on the same problem. Only then can standards emerge that will allow progress in the field even to be measured.

This type of shared problem—known as a "challenge evaluation"—has become something of a "religion" in the speech and language community since the 1980s, says Hirschman. By putting out a set of data to train on and then issuing a "challenge" for each group to extract the same information or answer the same questions, "you compare apples to apples. In the process you build a research community."

Last year, Hirschman and others ran the very first challenge evaluation in biology, the KDD Cup (officially called the Knowledge Discovery and Data-Mining Challenge Cup). Six weeks in advance, the organizers gave participants a training set of 862 journal articles already included in the model organism database FlyBase, along with associated lists of genes and gene products, as well as relevant data fields from FlyBase. After building their software tools, the entrants were then asked to take a test set of 213 articles and pretend they were curators: the tools were supposed to determine whether the articles were appropriate for curation, based on whether they contained experimental evidence for gene expression products, including both RNA transcripts and proteins.

Eighteen participants took a shot at the KDD Cup and their results speak of the infant state of the field. On average, they could assign only 58% of the papers correctly and could determine whether relevant gene products were present only 35% of the time. The winning entrant, a joint group from the Israeli company Clearforest ("see the forest *and* the trees") and Maryland-based Celera Genomics, did better.

Their entry made the right decision to curate 78% of the time and the right call on the presence of gene products 67% of the time.

The winning group did so well by using a clever "trick," says Hirschman admiringly. Their program searched for figure captions and then applied multiple techniques to find those gene products they were looking for.

Getting Out of the "Bag of Words"

The techniques applied by Clearforest and others fall into two broad categories, statistical and heuristic. Statistical techniques are the next step up from keyword searches. They count words such as genes or gene products appearing close to one another, but apply no linguistic insights, such as whether an adjective modifies a noun. By contrast, heuristic approaches use hand-crafted rules designed for specific datasets: e.g., January, February, March, etc., are months; the word following "Mr." is a name; and so forth. This approach is labor-intensive but especially useful when there is only a limited amount of data—as is the case with single scientific papers or small groups of papers.

Some statistical approaches have been labeled with the nickname "bag of words" because they fail to account for grammatical relationships; e.g., "man bites dog" and "dog bites man" would drop the same three words in the bag. A key observation at the KDD Cup was that the most basic statistical approach, which counts word occurrences at the document level, is not sufficient unless it takes into account at least some higher-level context, such as the part of the paper from which the search terms are extracted.

Furthermore, the more hand-crafted rules there were, the better. Many of the top teams included biologists who applied their expertise to help create empirical rules that became part of the program instructions. This points to a general theme in machine learning: the greater the degree of human intervention, the better. The best programs are covered with fingerprints.

Access: A Wrench in the Works?

Although the march toward better text-mining systems is building momentum, there are two issues that could stop it in its tracks. The first is access.

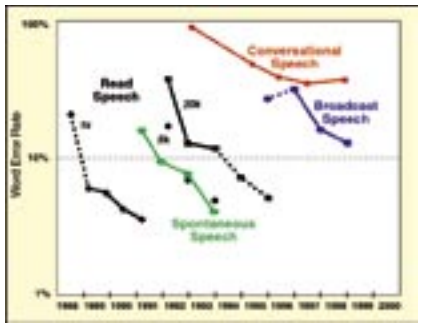
Experts in text-searching uniformly cite access as a key obstacle for developing better search tools. "Access is a bigger problem than algorithms" is how one machine-learning expert puts it, and a half-dozen others agreed.

The present "balkanized" situation for text-processing is filled with "dead ends" and "short circuits" in information flow among biologists, says David Lipman, head of the United States' National Center for Biotechnology Information, which runs PubMed, the MEDLINE database, as well as the National Library of Medicine and other critical resources in biology and bioinformatics. It is as if readers are marine biologists on a coastline whose beaches are 98% private. At best, asking permission to view every article slows down the work. At worst, there are some important tools one can never build owing to the missing context. MEDLINE itself would be much more powerful if it were based on full text, experts say.

Owing to lack of access, says Hirschman, "we miss a great deal by not having large corpora of full-text articles" included in the design of both the KDD Cup and the next challenge evaluation, called BioCreative, being held later this year. Many of the relevant biological data are found outside abstracts, but getting access to full text is complicated at best. For manual searching, researchers traditionally fall back on portal-hopping: jumping from one full-text subscription (to *Nature*, *Science*, or *Cell*, for example) to another, or from one portal (HighWire, Web of Science) to another. That way, many scientists routinely obtain access to as many as 80% of the journals they need. The rest they can usually request via interlibrary loan or order as photocopies online. However, this approach fails for most automated search programs. Just sorting out the permissions and keeping up with changes in the portals dramatically increase the headaches for anyone trying to build a search tool.

Laying Heisenberg to Rest

The second threat to text-searching programs ever becoming widely useful has more of the ring of linguistics jargon. The so-called " ontology problem" threatens successful searching based on the very specific



DOI: 10.1371/journal.pbio.0000048.g002

Figure 2. The Impact of Challenge Evaluations (and Investment Dollars)
Driven by investment and competition as well as the pressure of regular challenge evaluations, error rates in speech recognition have dropped steadily, to the point where the technology has become standard from directory assistance to travel to financial information. Error rates drop by a factor of two every two years as challenge evaluations attract wide participation. (Graph adapted with permission from the MITRE Corporation.) Source: Pallett D, Garofolo J, Fiscus J (2000) Measurements in support of research accomplishments. Communications of the ACM: Special section on broadcast news understanding.

nature of biological terminology.

The issue here is not only that scientists are truly terrible about sticking to established terminologies. "Scientists would rather share each other's underwear than use each other's nomenclature," as biochemist Keith Yamamoto is fond of saying. Consequently, the scientific literature is a hodgepodge of identical or overlapping terms. A naïve text-parsing program does not know whether "cat" refers to the catalase gene, the chloramphenicol transferase gene, or a household animal.

The challenge is to build an ontology describing all the important relationships so your computer program can navigate among them without asking you what to do. Consequently, an ontology would prescribe rules for understanding the interactions among genes based on the appearance of certain verbs ("inhibit," "express"), nouns ("agonist"), or phrases. Although within each narrow scientific subdiscipline it may be possible to build exquisitely useful text-mining tools, as soon as programmers broach the borders of the narrowest subfields, they will run into a kind of Heisenberg uncertainty principle of linguistics and science. Every toolmaker

is faced with the ontology problem in one respect or another, especially when the tool is meant to be a general one.

David Gilmour, chief executive officer of Tacit Inc., a knowledge management company in Palo Alto, California, is an industry veteran of exactly this war "and I have scars all over my body to prove it," he says. The issue in a nutshell, he explains, is that "ontologies scale poorly, and by the time they are useful," that is, large enough to capture most of the possible relationships among words, "they are unmaintainable."

Hirschman acknowledges that keeping up with the literature and new terminologies is challenging. Adapting tools to new domains has traditionally been one of the "critical stumbling blocks" for text-processing technology, she says. The dynamic growth of biological terminology does not help. There are 50–100 alterations *every week* to the nomenclature section of mouse genome database Web page.

Textpresso, Anyone?

Staying within one's narrow domain, then, could be a recipe for success, as long as the vocabulary and user questions remain tightly constrained, especially if there is a way to tiptoe around the access problem. That is apparently the case at Wormbase, though the newly available tool there, called Textpresso, is still being built. The motivation for Textpresso was simple, says Hans-Michael Mueller, a postdoctoral fellow in the lab of Paul Sternberg at Caltech in Pasadena, California, where Wormbase—the genetic database for the nematode worm *Caenorhabditis elegans*—is curated. "We want the user to be able to avoid going to the library to read all those papers [on genes and proteins] that your favorite gene interacts with. That is very tedious." The other goal is equally recognizable in the biology community: no mere mortal can hope to keep up with the burgeoning literature, even in the relatively narrow field of worm biology.

Mueller, a nuclear physicist by background, called Textpresso "a search engine for full-text searches of abstracts and articles" that can help find answers to more challenging queries than simple keyword searches.

Mueller and his team use human "taggers" to mark up the corpus of text to indicate categories like "biological processes" ("late larval activation"), "genes" (let-7), and "molecular functions." Then, like the Clearforest-Celera program, Textpresso searches for combinations of categories in the same or neighboring sentences. The ontology relating the expressions and categories to one another is based both on scientific and common sense as well as linguistic components. In less than two years of work, Mueller and his team have already marked up 3.9 million terms in 16,000 abstracts and 2,000 full-text papers. A typical search asks a question such as "What can be found out about the negative regulatory aspects of a genetic network in the pharynx?" Answers emerge in the form of citations, abstracts, and, if available, a paragraph or so from the text of the relevant paper. Textpresso went up—unpublicized—on the Web in February this year and already receives a couple of hundred hits a day, a big number in a field of about 2,000 researchers. Mueller estimates that Textpresso is 95% accurate and that about 35% of the relevant papers have been included.

Textpresso needs full-text access to be as good as it is, says Mueller. "We noticed" that drawing on full text "greatly increased the chances of a true hit," not a false positive. He managed to avoid the access issue by claiming a kind of "curator's privilege." Only the curators see the full text. Once the data are on the Web, users can only get at most a paragraph, which falls within fair use, said Mueller. If a user happens to subscribe to the journal in question, it is possible for him to click through, the publisher's portal and see the paper.

Whereas Textpresso works exclusively on worm genetic data and commercial players like Clearforest are just beginning to hunt for biological applications, a handful of companies have begun to market text-searching products to academic biomedical scientists. One such product is called QUOSA, for query, organize, share, and analyze. The software had its commercial launch in late 2002. Put simply, the program—available on an institution-wide basis and already installed for hundreds of researchers at Massachusetts General Hospital and the Dana-Farber Cancer Institute in Boston—allows a search across one's own documents. A front end for the literature that cooperates with MEDLINE, QUOSA pulls in and prioritizes full-text papers. The program first allows the user to search for the relevant files and download them in full-text format to the extent permitted by her library's subscription agreements and licenses. Once it becomes second nature to users, they rave about it. Like the best of the first-generation software, QUOSA allows users to make connections they would not have otherwise made. Like so many other early software products, its long-term success will hinge on demand as well as improvements made in the upgrades.

Because of the ontology problem, improvements in searching in the next couple of years are likely to result from the application of ever-better techniques within existing domains. Collaborations among Wormbase, Flybase, and other model-organism database groups will help improve all their search tools. MEDLINE itself may benefit from more advanced search techniques, though these will be restricted to abstract searches.

The big unknown for predicting further development of text-search tools is the path publishers will take. If each publisher or portal such as Reed-Elsevier or HighWire were to license or develop its own tool for searching its own content, the result might be better than the status quo, but would still be unsatisfying. Running the same search three times on three different subsets of content might be better than running it 15 times—but wouldn't it be easier to run it just once? ■