COMMENT

ILLUSTRATION BY JONATHAN BURTON

Search needs a shake-up

On the twentieth anniversary of the World Wide Web's public release, **Oren Etzioni** calls on researchers to think outside the keyword box and improve Internet trawling.

Two decades after Internet pioneer Tim Berners Lee introduced his World Wide Web project to the world using the alt.hypertext newsgroup, web search is on the cusp of a profound change — from simple document retrieval to question answering. Instead of poring over long lists of documents that contain requested keywords, users need direct answers to their questions. With sufficient scientific and financial investment, we could soon view today's keyword searching with the same nostalgia and amusement reserved for bygone technologies such as electric typewriters and vinyl records.

But this transformation could be unreasonably delayed. As a community, computer scientists have underinvested in tools that can synthesize sophisticated answers to questions, and have instead focused on incremental progress in lowest-common-denominator search. The classic keyword search box exerts a powerful gravitational pull. Academics and industry researchers need to achieve the intellectual 'escape velocity' necessary to revolutionize search. They must invest much more in bold strategies that can achieve natural-language searching and answering, rather than providing the electronic equivalent of the index at the back of a reference book.

Today, that 'book' is distributed over billions of web pages of uneven quality, and much effort has been directed at ranking the most useful results. Such engines readily index billions of documents, but overwhelm their users with millions of results in response to simple queries. This quandary only worsens as the number of web pages ▶

grows, and as web access shifts to mobile devices with tiny screens.

Moving up the information food chain requires a search engine that can interpret a user's question, extract facts from all the information on the web, and select an appropriate answer.

The big search engines have taken tiny steps in the right direction. Google directly displays film showing times and weather forecasts in response to particular queries, but this is a drop in the ocean of possible search queries. Bing markets itself as a 'decision engine' rather than a 'search engine', but with the exception of its airfare predictor, its differences from Google are limited. Wolfram Alpha's 'computational knowledge engine' provides remarkably sophisticated answers to certain questions. For example, when asked "am I too drunk to drive?", Wolfram Alpha offers to compute your blood alcohol level based on factors such as your weight and the number of drinks consumed. Unfortunately, this approach applies to only a very limited set of pre-specified questions (it fails to answer the similar question "have I had too much to drink?").

Microsoft acquired Powerset, a start-up company developing a natural-language search engine, for upwards of US$100 million in 2008. And in 2010, Apple bought Siri, a start-up with an iPhone app that answered natural-language questions about films, restaurants and other local services, for a price speculated to have been more than $200 million. But these efforts are dwarfed by companies' investments in keyword search. In short, search engines have yet to develop general-purpose question-answering capabilities.

INFORMATION EXTRACTION

Work on the challenge of automatically recovering factual information from text began in the 1970s. Early information-extraction systems were hand-crafted to particular genres and very narrow topics. For example, the JASPER system extracted financial information from Reuters newswire text in the 1980s. Broadening their scope was both labour-intensive and error prone. In the 1990s, a more automated class of information-extraction system emerged. Instead of relying on hand-crafted rules to glean facts from sentences, the systems generated the rules automatically based on a curated collection of example sentences. This automated approach is more streamlined and less error prone, but still requires careful manual effort to create an example collection for each topic of interest.

In 2007, my lab introduced open information extraction — methods that scale to any topic and to arbitrary English sentences. The basic idea is remarkably simple: most sentences contain highly reliable syntactic clues to their meaning. For example, relationships are often expressed through verbs (such as invented, married or elected) or verbs followed by prepositions (such as invented by, married to or elected in). It is often quite straightforward for a computer to locate the verbs in a sentence, identify entities related by the verb, and use these to create statements of fact. Of course this doesn't always go perfectly. Such a system might infer, for example, that 'Kentucky Fried Chicken' means that the state of Kentucky fried some chicken. But massive bodies of text such as the corpus of web pages are highly redundant: many assertions are expressed multiple times in different ways. When a system extracts the same assertion many times from distinct, independently authored sentences, the chance that the inferred meaning is sound goes up exponentially.

Open information extraction obviates topic-specific collections of example sentences, and instead relies on its general model of how information is expressed in English sentences to cover the broad, and unanticipated, universe of topics on the Internet.

> ***"The main obstacle to question answering seems to be a curious lack of ambition and imagination."***

Other approaches to information extraction are also yielding important results. There are projects at Carnegie Mellon University in Pittsburgh, Pennsylvania, Stanford University in Palo Alto, California, and New York University, at companies such as Google and Microsoft, and at numerous start-ups. In contrast to open information extraction, however, these methods cannot automatically operate at the scale of the web. For instance, some projects have come up with ultrafast systems to provide in-depth understanding of sentences, but they only work in specific domains such as finance, or in particular genres such as Wikipedia articles.

Some scientists are experimenting with 'power tools' that delve into the content of scientific articles to suggest novel connections and potential hypotheses (see *Nature* **463,** 416–418; 2010). Many of these tools are not fully automated, however, which immediately leads to challenges in extending them beyond carefully circumscribed arenas such as gene names in PubMed abstracts. The use of open information extraction would substantially broaden their scope. The open-source code for our system is available at go.nature.com/ei3p4f.

Much more research has to be done to improve information-extraction systems — including our own. Their abilities need to be extended from being able to infer relations expressed by verbs to those expressed by nouns and adjectives. Information is often qualified by its source, intent and the context of previous sentences. The systems need to be able to detect those, and other, subtleties. Finally, automated methods have to be mapped to a broad set of languages, many of which pose their own idiosyncratic challenges.

PROGRESS IN JEOPARDY

The main obstacle to the paradigm shift from information retrieval to question answering seems to be a curious lack of ambition and imagination. Much of the research on natural language processing is focused on limited tasks, such as recovering the syntactic structure of sentences rather than trying to uncover their meaning, or on methods that do not scale to massive corpora and arbitrary topics because of their reliance on manually-annotated data, or on algorithms whose computation grows explosively with the amount of text involved.

In 2009, the US Defense Advanced Research Projects Agency began a 'Machine Reading' programme that has focused attention on this area, with tens of millions of dollars awarded in contracts to a handful of teams — including mine. An order of magnitude more research funding is necessary, as is a focus on scaling up current methods to the size and heterogeneity of the web.

One exceptional system — IBM's Watson — utilizes a combination of information extracted from a corpus of text equivalent to more than 1 million books combined with databases of facts and massive computational power. Watson won a televised game of Jeopardy against two world-class human players in February this year. The multi-billion dollar question that IBM is now investigating is 'can Watson be generalized beyond the game of Jeopardy?'

General-purpose question-answering systems will be a boon to scientists searching the literature, and to the increasing number of us who access the web's richness through a mobile phone with a tiny screen that necessitates concise responses. Without it, we risk drowning in the growing sea of information. ■

Oren Etzioni *is the director of the Turing Center at the University of Washington, Seattle, Washington 98195, USA.*
e-mail: etzioni@cs.washington.edu