PHYS ORG

Mining the language of science

November 18 2011

Credit: AI-generated image (disclaimer)

(PhysOrg.com) -- Scientists are developing a computer that can read vast amounts of scientific literature, make connections between facts and develop hypotheses.

Ask any biomedical scientist whether they manage to keep on top of reading all of the publications in their field, let alone an adjacent field, and few will say yes. New publications are appearing at a doubleexponential rate, as measured by MEDLINE – the US National Library of Medicine's biomedical bibliographic database – which now lists over 19 million records and adds up to 4,000 new records daily.

For a prolific field such as cancer research, the number of publications could quickly become unmanageable and important hypothesis-generating evidence may be missed. But what if scientists could instruct a computer to help them?

To be useful, a computer would need to trawl through the literature in the same way that a scientist would: reading the literature to uncover new knowledge, evaluating the quality of the information, looking for patterns and connections between facts, and then generating hypotheses to test. Not only might such a program speed up the progress of scientific discovery but, with the capacity to consider vast numbers of factors, it might even discover information that could be missed by the human brain.

The aim of Dr. Anna Korhonen and researchers in the Natural Language and Information Processing Group in the University of Cambridge's Computer Laboratory is to develop computers that can understand written language in the same way that humans do. One of the projects she is involved in has recently developed a method of 'text mining' one of the most literature-dependent areas of biomedicine: cancer risk assessment of chemicals.

Every year, thousands of new chemicals are developed, any one of which might pose a potential risk to human health. Complex risk assessment procedures are in place to determine the relationship between exposure and the likelihood of developing cancer, but it's a lengthy process, as Royal Society University Research Fellow Dr Korhonen explained: "The first stage of any risk assessment is a literature review. It's a major bottleneck. There could be tens of thousands of articles for a single chemical. Performed manually, it's expensive and, because of the rising number of publications, it's becoming too challenging to manage."

CRAB, the tool her team has developed in collaboration with Professor Ulla Stenius' group at the Institute of Environmental Medicine at Sweden's Karolinska Institutet, is a novel approach to cancer risk assessment that could help risk assessors move beyond manual literature review.

The approach is based on text-mining technology, which has been pioneered by computer scientists, and involves developing programs that can analyse natural language texts, despite their complexity, inconsistency and ambiguity. The tool Dr. Korhonen has developed with her colleagues is the first text-mining tool aimed at aiding literature review in chemical risk assessment.

At the heart of CRAB, the development of which was funded by the Medical Research Council and the Swedish Research Council among others, is a taxonomy that specifies scientific evidence used in cancer risk assessment, including key events that may result in cancer formation. The system takes the textual content of each relevant MEDLINE abstract and classifies it according to the taxonomy. At the press of a button, a profile is rapidly built for any particular chemical using all of the available literature, describing highly specific patterns of connections between chemicals and toxicity.

"Although still under development, the system can be used to make connections that would be difficult to find, even if it had been possible to read all the documents," added Dr. Korhonen. "In a recent experiment, we studied a group of chemicals with unknown mode of action and used the CRAB tool to suggest a new hypothesis that might explain their male-specific carcinogenicity in the pancreas."

The tool will be available for end-users via an online web interface. However, research into improving text mining will continue. One of the biggest current challenges is to develop adaptive technology that can be ported easily between different text types, tasks and scientific fields.

One day, rather than being at the mercy of the flourishing rate of publication, scientists will have at their fingertips a system to work alongside them that will not only point them towards those references that are relevant to their search, but will also tell them why.

Provided by University of Cambridge

Citation: Mining the language of science (2011, November 18) retrieved 1 May 2024 from https://phys.org/news/2011-11-language-science.html