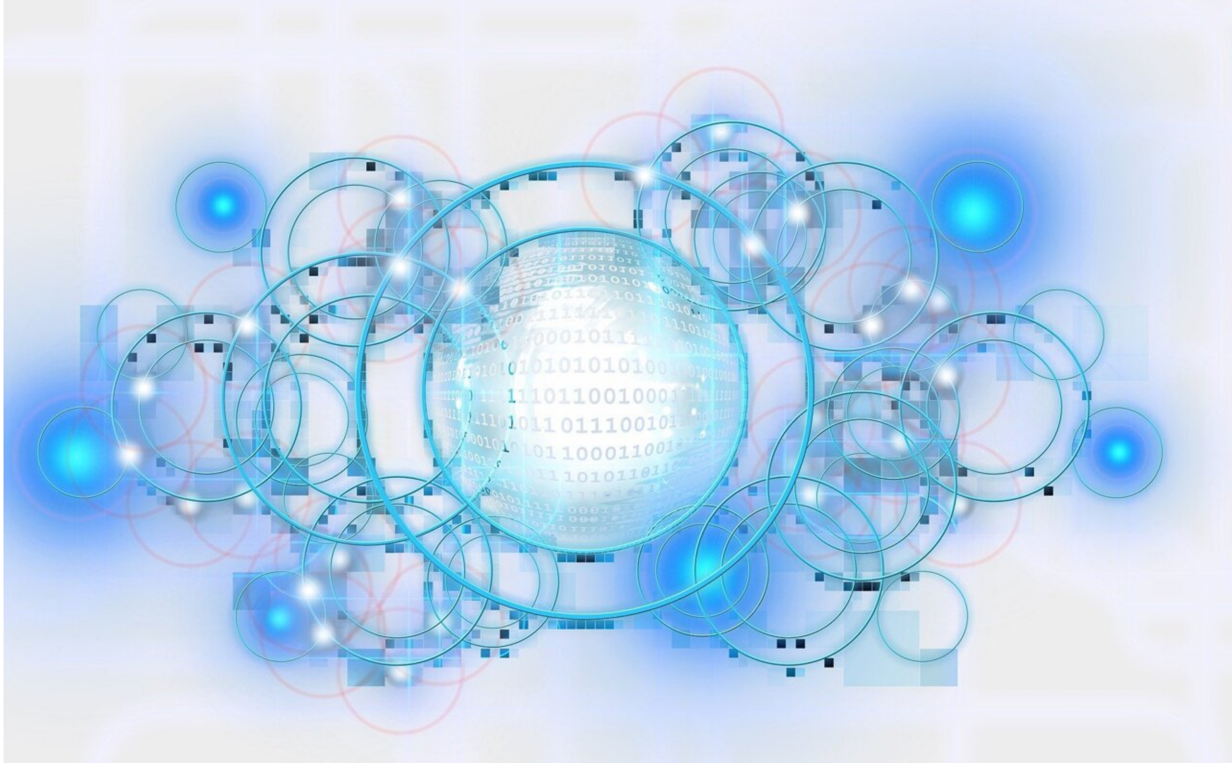
Tech Xplore

Sorting out smart data

January 19 2022, by David Bradley



Credit: Pixabay/CC0 Public Domain

Might scoring the contents of scientific papers based on semantics and lexicon allow a representation of textual experimental data from scientific publications to be extracted? That is the question a team from France hope to answer in the *International Journal of Intelligent Information and Database Systems*.

Martin Lentschat of the University of Montpellier and colleagues there and at the University of Paris-Saclay explain how their approach uses the scientific publication representation (SciPuRe) to describe extracted data through ontological, lexical, and structural features based on the segments in a scientific document. The scientific literature is vast and in many ways readily accessible to experts. However, a substantial amount of the information contained in this enormous space can only be mined, or harvested, for use by those experts, inclusion in meta-analyses or fed into advanced decision-support tools, if it is somehow processed and the data, information, and knowledge extracted into a form that can be used by the available tools.

The team points out that in the biomedical research domain there has been a lot of focus on how knowledge can be extracted automatically from the published literature because of the nature of the often date-rich experimental outputs. However, in other areas, there has been a lack of tools that can home in on useful information without the need to take prior knowledge and expertise into account. Where biomedical research pivots on big data other areas of research require smart data.

Big data needs no assessment, no scoring based on content and context, it can be pulled from a publication and processed because the prior knowledge about what the data mean is intrinsic to the data in a sense. To work with smart data, on the other, hand requires it to be assessed so that irrelevant data in a publication can be discarded, the new work points to how this very process might be automated to allow tools related to those used to handle big data in biomedical research to be used with smart data from other less data-intensive areas of research.

The team's success with the specialist topic discussed suggests that future studies might open up the same approach to other research domains, although whether those are equally as successful will remain to be seen.

"Experiments were carried out on a corpus of fifty English language scientific papers in the food packaging field," the team reports. "They revealed that article segments are an effective criterion for filtering out the majority of the quantitative entity false positives using lexical scores."

More information: Martin Lentschat et al, Towards combined semantic and lexical scores based on a new representation of textual data to extract experimental data from scientific publications, *International Journal of Intelligent Information and Database Systems* (2022). DOI: 10.1504/IJIIDS.2022.120146

Provided by Inderscience

Citation: Sorting out smart data (2022, January 19) retrieved 3 May 2024 from https://techxplore.com/news/2022-01-smart.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.