Semi-Automated Information Extraction to Improve Scientific Knowledge Discovery in Environmental Health Science Literature

Authors: Walker VR, Rooney AA, Kleinstreuer NC, Wolfe MS, and Schmitt CP
Division of the National Institute of Environmental Health Sciences, RTP, NC 27709

Introduction: The National Toxicology Program (NTP) at the National Institute of Environmental Health Sciences (NIEHS) conducts literature-based evaluations using systematic review methods to objectively and transparently assess the evidence that environmental substances may be associated with adverse health effects. These reviews collect information on chemicals from published scientific literature such as toxicity and health outcomes assessed, test methods, human populations or animal models, and results using a structured process (NTP 2019). Depending on the research question and the extent of available data, an evaluation may characterize the evidence in scoping reviews with interactive evidence maps (Figure 1) to identify data-poor and data-rich areas for further research, or the evaluation may integrate the evidence to reach conclusions on human health hazards associated with environmental exposures.

Conducting a review is time-consuming and resource intensive—often requiring more than 1,000 hours and $100,000 USD. As such, NTP is actively pursuing automated and semiautomated processes for information extraction (IE) and natural language processing (NLP) in our workflow to reduce time and labor-costs while maintaining quality and reproducibility for our products. The NTP is also applying IE methods to other data needs such as assessing institutional impact through citations and improving document screening and prioritization.

Figure1: Example interactive evidence map (Vinpocetine Outcomes in Animals)

Training Data: The NTP has several datasets available to share with experts to identify, develop, test, and implement new NLP/IE models relevant to systematic reviews and wants to develop environmental health relevant datasets that may inform models.

NIST SRIE: NTP hosted the 2018 Systematic Review Information Extraction (SRIE) track for the National Institute for Standards and Technology (NIST) Text Analysis Conference (TAC). For this challenge, NTP manually curated training and test data sets of 100 articles each by tagging mentions and groups of mentions relevant to IE in published animal studies (Figure 2).

Health Outcomes Data Set: NTP is preparing a data set of 1000 articles manually tagged at the full text level for health outcomes based on content in the title, abstract, and full text.

Systematic Reviews: NTP has access to 664 articles with extracted terms, but without location information that could be used in distant supervision.

Uterotrophic Data set: NTP has developed a data set of 670 articles tagged on six study protocol criteria considered minimal requirements to distinguish high-quality animal studies conducted according to regulatory guidelines (Kleinstreuer et al. 2015).

Challenges with Adoption of NLP

1. Methods advancement requires environmental health relevant datasets that are costly to develop. Approaches are needed for online models that self-improve with use, and generating training data from usage without impacting curation workflows (e.g., without requiring tagging of all negative or positive mentions).

2. Existing methods lack context awareness. Literature assessment workflow needs to identify relevant context of terms not just identify terms somewhere in the paper. For example, a chemical may be used as an anesthetic, part of laboratory procedures, or as an experimental treatment, exposure, or control. Methods to derive context and allow experts to provide context to models are needed.

3. Improvements are needed in handling mentions that are similar to but not specifically found in training sets, as well as for extracting complex concepts (e.g., timing of exposure / endpoints) that are not contiguous but rather have long spans and multiple subphrases.

4. Existing methods are not adept at extracting statistical results and assigning those results to the correct endpoints.

5. Advances in grouping extracted entities are needed. Environmental health assessments gather data from multiple evidence streams - epidemiological studies, non-human animal studies, and in vitro and mechanistic data. These all require data to be sorted and grouped. For example, a single animal study may investigate effects of exposure to one or more chemicals by evaluating multiple endpoints, on several health categories, in one or more species or experimental models. It is critical that semi-automated data extraction approaches consider and maintain the complex structure of the data being extracted.

Acknowledgements

The work presented in this abstract was supported through numerous collaborations with advice and technical assistance from the U.S. Environmental Protection Agency (Michelle Angrish, Kristan Markey, Michele Taylor, Kristina Thayer), ICF International (Yousef Ahmad, Cara Henning, Jessica Wignall Ashley Williams,) and Oak Ridge National Laboratory (Dasha Hermannova, Robert Patton, Christopher Stahl).
References:
