#4 - Semi-Automated Data Extraction Workbench for Environmental Health

B. Howard, A. Maharana, A. Tandon, and *Ruchir Shah*
Sciome, LLC

Systematic review, already a cornerstone of evidence-based medicine, has recently gained significant popularity in several other disciplines including environmental health and evidence-based toxicology. One critical and time-consuming process that must occur during systematic review is the extraction of relevant qualitative and quantitative raw data from the free text of scientific documents. The specific data types extracted differ among disciplines, but within a given scientific domain, certain data points are extracted repeatedly for each review that is conducted. To that end, Sciome has begun research and development of a semi-automated data extraction workbench for use in this context. We are focusing our research on three specific goals. First, we are using deep learning to build novel data extraction models to extract data elements of interest to those conducting systematic reviews in the domain of environmental health. Second, we are building a web-based data extraction software platform specifically designed for usage in the domain of systematic review. And finally, we plan to introduce new protocols to standardize the inputs and outputs for data extraction software components. Here we report our results so far, including the performance of more than 20 novel data extraction components of relevance to environmental toxicology, created and tested on an annotated dataset from NTP. Performance varied widely among data types with some tasks inherently more difficult than others. For certain simple data types, like sex of the experimental animal, we achieved F-scores in excess of 95%; for more difficult entities, we were still often able to achieve an F-score of 65% or more, given sufficient training data. Because accurate data extraction can be a challenging problem, and given that current methods rarely achieve 100% accuracy, we are integrating our methods into a “human-in-the-loop” system that combines machine and human intelligence in a manner that is superior to using either in isolation. The system will: highlight extracted terms in a pdf; automatically populate extraction forms with extracted data; allow humans to intervene and correct the results; and learn from the corrections to continually update the model. The resulting system will make systematic reviews both more efficient to produce and less expensive to maintain, greatly accelerating the process by which scientific consensus is obtained in a variety of health-related disciplines having great public significance.