Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline

Roselyne B. Tchoua, Kyle Chard, Debra J. Audus, Logan Ward, Joshua Lequieu, Juan de Pablo, and Ian Foster

Caption: 1: The six-stage hybrid IE pipeline, showing (1) the NLP Module, which identifies Tg candidates; (2) the Polymer Dictionary Module, which identifies polymer names in NLP output; (3) the three automated extraction and crowdsourcing modules used to process different forms of candidates; (4) the Flag Bad Data Crowdsource Module, in which crowds flag anomalous results, (5) the Prioritize Review Module, which ranks extracted polymer–Tg pairs to prioritize expert validation, and (6) the Final Expert Review.

Scientific Achievement
The emerging field of materials informatics has the potential to greatly reduce time-to-market and development costs for new materials. The success of such efforts hinges on access to large, high-quality databases of material properties. However, many such data are only to be found encoded in text within esoteric scientific articles, a situation that makes automated extraction difficult and manual extraction time-consuming and error-prone. To address this challenge, we present a hybrid Information Extraction (IE) pipeline to improve the machine-human partnership with respect to extraction quality and person-hours, through a combination of rule-based, machine learning, and crowdsourcing approaches. Our goal is to leverage computer and human strengths to alleviate the burden on human curators by automating initial extraction tasks before prioritizing and assigning specialized curation tasks to humans with different levels of training. To validate our approaches, we focus on the task of extracting the glass transition temperature of polymers from published articles. Applying our approaches to 6,090 articles, we have so far extracted 259 refined data values. We project that this number will grow considerably as we tune our methods and process more articles, to exceed that found in standard, expert-curated polymer data handbooks while also being easier to keep up to date.
Significance
A tremendous amount of scientific information, critical to both research and industry, is available in decades of journal literature, though it is often locked in difficult-to-parse natural language formats. We are exploring how hybrid human-automated information extraction (IE) and curation pipelines can be used to liberate this information. Many of these IE and curation pipelines may be applicable to extraction of other scientific information from the literature in future efforts.

Citation