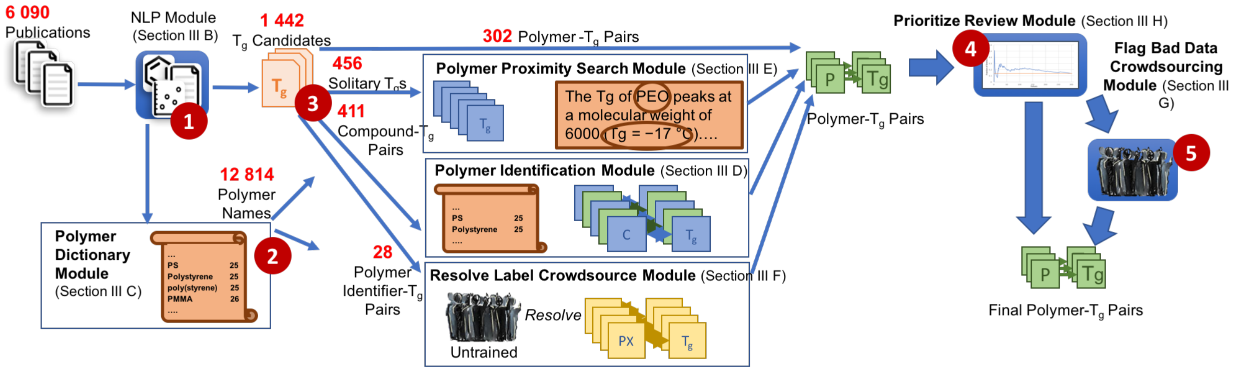
CHiMaD
Center for
Hierarchical
Materials Design
chimad.northwestern.edu

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


6 090
Publications
NLP Module
(Section III B)
1
1 442
T_g Candidates
302 Polymer -T_g Pairs
3
456
Solitary T_gs
411
Compound-T_g
Pairs
Polymer Proximity Search Module (Section III E)
T_g
The Tg of PEO peaks at a molecular weight of 6000 (Tg = −17 °C)....
12 814
Polymer
Names
Polymer Dictionary Module (Section III C)
...
PS 25
Polystyrene 25
poly(styrene) 25
PMMA 26
....
2
28
Polymer
Identifier-T_g
Pairs
Polymer Identification Module (Section III D)
...
PS 25
Polystyrene 25
....
C
T_g
Resolve Label Crowdsource Module (Section III F)
Resolve
Untrained
PX
T_g
P
Tg
Polymer-T_g Pairs
Prioritize Review Module (Section III H)
4
Flag Bad Data Crowdsourcing Module (Section III G)
5
P
Tg
Final Polymer-T_g Pairs

Caption: *1: The six-stage hybrid IE pipeline, showing (1) the NLP Module, which identifies T_g 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–T_g 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

Roselyne B. Tchoua, Kyle Chard, Debra J. Audus, Logan Ward, Joshua Lequieu, Juan de Pablo, and Ian Foster. Towards a hybrid human-computer scientific information extraction pipeline. In 2017 IEEE 13th International Conference on e-Science, Oct. 2017.