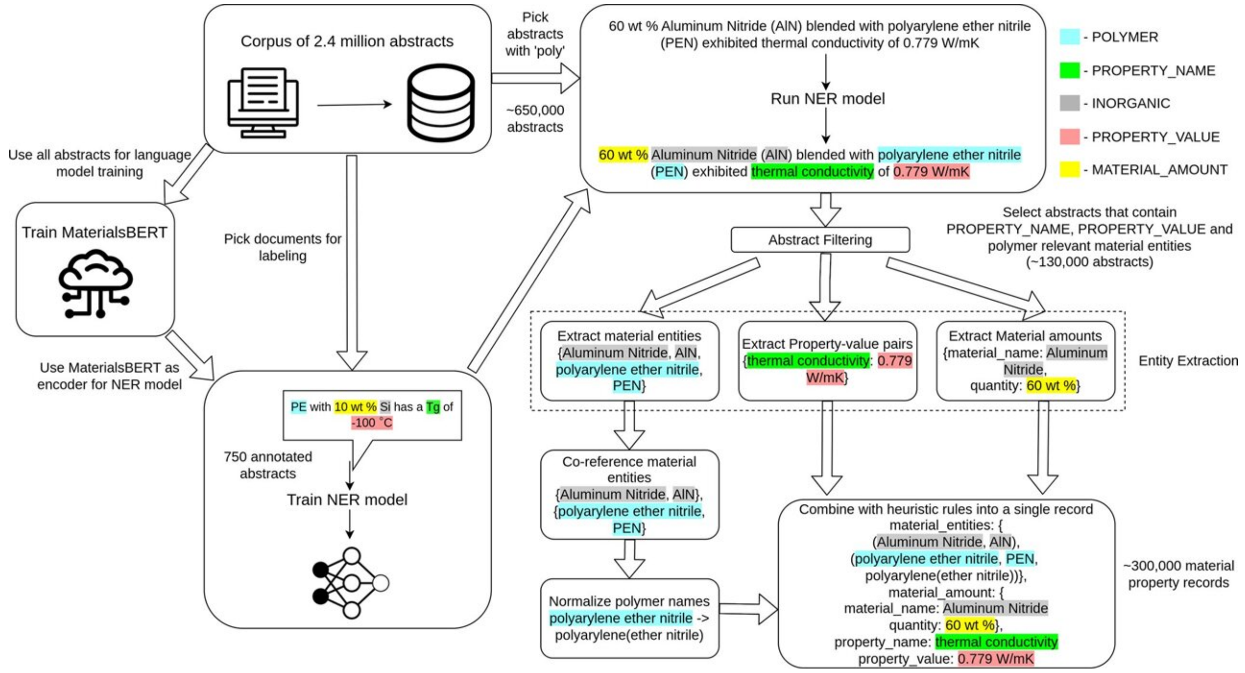
PHYS.ORG

Data extraction tool may lead to discovery of new polymers

July 13 2023


Corpus of 2.4 million abstracts
Pick abstracts with 'poly'
~650,000 abstracts
60 wt % Aluminum Nitride (AlN) blended with polyarylene ether nitrile (PEN) exhibited thermal conductivity of 0.779 W/mK
Run NER model
60 wt % Aluminum Nitride (AlN) blended with polyarylene ether nitrile (PEN) exhibited thermal conductivity of 0.779 W/mK
- POLYMER
- PROPERTY_NAME
- INORGANIC
- PROPERTY_VALUE
- MATERIAL_AMOUNT
Use all abstracts for language model training
Train MaterialsBERT
Pick documents for labeling
Abstract Filtering
Select abstracts that contain PROPERTY_NAME, PROPERTY_VALUE and polymer relevant material entities (~130,000 abstracts)
Use MaterialsBERT as encoder for NER model
PE with 10 wt % Si has a Tg of -100 °C
750 annotated abstracts
Train NER model
Extract material entities {Aluminum Nitride, AlN, polyarylene ether nitrile, PEN}
Extract Property-value pairs {thermal conductivity: 0.779 W/mK}
Extract Material amounts {material_name: Aluminum Nitride, quantity: 60 wt %}
Entity Extraction
Co-reference material entities {Aluminum Nitride, AlN}, {polyarylene ether nitrile, PEN}
Normalize polymer names polyarylene ether nitrile -> polyarylene(ether nitrile)
Combine with heuristic rules into a single record
material_entities: {
(Aluminum Nitride, AlN),
(polyarylene ether nitrile, PEN, polyarylene(ether nitrile))},
material_amount: {
material_name: Aluminum Nitride
quantity: 60 wt %},
property_name: thermal conductivity
property_value: 0.779 W/mK
~300,000 material property records

Pipeline used for extracting material property records from a corpus of abstracts. Credit: *npj Computational Materials* (2023). DOI: 10.1038/s41524-023-01003-w

The amount of published materials science research is growing at an exponential rate, too fast for scientists to keep up. To help these scholars, a first-of-its-kind materials science data extraction pipeline is now available to make their research easier and faster.

The pipeline extracts material property records from published papers and populates the data into a new application called Polymer Scholar. The platform works like a browser to search polymers and materials properties by keyword, rather than reading through countless articles.

The application makes materials research more efficient, which could lead to the discovery of new polymers.

"Essentially, we have created an index on materials science literature that is much more granular than ones in a typical index that a search engine would create," said Georgia Tech Ph.D. student Pranav Shetty, the lead designer of the pipeline.

"Our hope is that materials science researchers can make use of this capability in their day-to-day lives and workflows, and therefore, allow their work to have much more usability toward studying polymers and developing new materials."

The group's paper says the number of materials science papers published annually grows at a rate of 6% compounded annually. This amount of content makes for long, difficult work for scientists and in need of a computing solution.

The group's answer is MaterialsBERT, a model they built and trained that powers the pipeline.

MaterialsBERT categorizes words in the text by association with a material property record. After the model associates text with records, the data is fed to Polymer Scholar. Scientists can use Polymer Scholar to study data, searching by either a polymer name or a property, like boiling point or tensile strength.

The group used 2.4 million materials science abstracts to train

MaterialsBERT. In tests, the model outperformed five other models on three of five entity-recognition datasets.

According to the study, the pipeline needed only 60 hours to obtain 300,000 material property records from over 130,000 abstracts.

As a comparison, materials scientists currently use a database called PoLyInfo. This system has over 492,000 material property records, manually curated by hand over the span of many years. Georgia Tech's pipeline can accomplish in hours what took humans years to do in PoLyInfo.

"Polymer Scholar and MaterialsBERT are powered by a large corpus of 2.4 million materials science articles, which took some time and effort to develop the infrastructure to support such a large collection," said Chao Zhang, an assistant professor in the School of Computational Science and Engineering (CSE). "This body of papers made all the difference in training MaterialsBERT because it improved the language model's ability to identify and extract data."

Polymer research is vital because of the roles polymers play in manufacturing, healthcare, electronics, and other industries. Polymers have desirable properties that make them useful for future applications.

When polymer research slows, it inhibits the development of new technologies. These technologies are needed to overcome today's challenges like climate change, faltering infrastructure, and sustainable energy.

In their paper, the group analyzed data using polymer solar cells, fuel cells, and supercapacitors as keywords in Polymer Scholar. This showed that scholars can use the pipeline to infer trends and phenomena in materials science literature. It also used practical examples to demonstrate applicability.

The journal *npj Computational Materials* published the group's paper because of its findings.

The pipeline is the latest work for the group which is committed to applying computational methods to lead innovations in materials science.

"Our long-term vision is to use the extracted data to train models that can predict material properties," Ramprasad said. "Creating a pipeline to extract this data that can seamlessly feed into predictive models will ultimately lead to an extraordinary pace of materials discovery."

More information: Pranav Shetty et al, A general-purpose material property data extraction pipeline from large polymer corpora using natural language processing, *npj Computational Materials* (2023). DOI: 10.1038/s41524-023-01003-w

Provided by Georgia Institute of Technology

Citation: Data extraction tool may lead to discovery of new polymers (2023, July 13) retrieved 29 April 2024 from https://phys.org/news/2023-07-tool-discovery-polymers.html