Technology Review:
Text Retrieval and Mining for Materials Scientific Literature

Ben Yang
CS410 Fall 2021

Applications of natural language processing and text information analysis methods are already ubiquitous and evolving within scientific disciplines. Most notably, physics and bioinformatics are among the most popular fields of text analysis and data science principles in general. In comparison, text information analysis of scientific literature in the domain of materials sciences is still emerging. Yet, the benefits of applying NLP techniques in this field has proven to assist in the acceleration of research and innovation. Analyzing text data has hastened the mundane pace of literature browsing and therefore accelerated understanding and experimentation. Relationships between words and phrases of seemingly unrelated technical lexicons have yielded insight to new chemistries and structure-property relationships. In this review, a high-level summary of the challenges, methodologies, and current applications of natural language processing and text information analysis in materials science literature will be explored.

Challenges of effectively applying text retrieval analysis methods to materials science literature arise from the lack of standardization of publications, highly segmented and heterogeneous data sources, and manipulation of complex and technical vocabularies and word relationships. The lack of a large, centralized database of publications using consistent formatting and language is a major hurdle to the efficiency of text data wrangling and analysis. This inhomogeneous collection of text data is exacerbated by the segmentation of materials science publications. Within the field, there are hundreds of journals that focus on a subset of materials research, and each may have its own formatting standards and practices. Additionally, metadata such as tables and figures that the text refers to may also be non-standardized and up to the authors discretion, therefore further risking missing or incomplete information. Further confusion and handling difficulties arise from the lack of standardization in the vocabulary itself. Through a plethora and frequent use of synonyms, homonyms, and abbreviations in chemistry, aggregation and comprehension of text data across different papers and sub-research areas in materials become an arduous post-processing task. Lastly, the possibility of an author's lack of technical rigor in their experimentation and results could falsely skew text data and conclusions. A significant and current problem in academic research as a whole is the possibility of an author cherry-picking experiments and results to discuss in his or her paper. Thus, the results may not be reproducible or hold up under scrutiny during peer-review and ultimately pose an additional complexity during text information data analysis. However, although these challenges exist in text analysis of materials science literature, methods that have demonstrated promise in this application are currently being used and others are in development.

Text information methods in materials science literature follow a similar framework as other natural language processing techniques. First, text content is acquired often through academic journals, literature, and industry patents relating to the research area of interest. Common text repositories include Springer Nature, Elsevier, and PubMed Central. The content is then preprocessed and tokenized using materials-specific tokenizers such as *OSCAR4* or *ChemDataExtractor* to manage the specialized chemical notation and vocabulary. Entity relations and links are established using algorithms tailored to the syntactic relationships observed in materials literature. These grammatical structures are typically written in the third-person and past tense with no personalization mentioned in the paper, which is different from other bodies of text where NLP is applied. Document segmentation serves to further define the larger structures within a paper to afford better accuracy by subdividing the text into sections such as abstract, method, results, as well as tables and figures. These preprocessing steps allow for named entity recognition analysis through a combination of dictionary look-up, rule-based, and algorithmic methods. Tools that leverage a rule-based approach include *LeadMine* and *ChemicalTagger*. Another technique is ML-based statistical models that represent syntactic and semantic term relationships, also known as word embeddings. Such models are dependent on their respective training set and is represented by the variety of tools, including *Word2Vec* and *FastText*. Lastly, entity relation extraction is the step where discovery and new knowledge typically occurs through identifying features such as clustering of textual similarity or correlations between chemical entities and properties.

Text information models have had several notable contributions to materials innovation in the past decade. One of the most prominent efforts within the polymer field is the Polymer Property Predictor and Database, spearheaded by NIST and a consortium of Chicago-based universities as part of the Materials Genome Project. Using text retrieval tools including *ChemDataExtractor*, the project has procured about 40,000 chemical compounds and their respective phase transition temperatures in a trustworthy and peer-reviewed manner. Another noteworthy effort is described by Kononova et al., where in the field of solid-state chemical synthesis, about 20,000 recipes were extracted from 50,000 paragraphs and then used to further extract a curated procedure of synthesizing germanium-based zeolites. Thus, even though NLP techniques in materials science literature is still in its relative infancy, the accomplishments that have already been achieved is a demonstration of its future potential.

Natural language processing and text information analysis of scientific literature in materials science is still relatively new but has already shown its benefits to the field. The tools and algorithms for tokenization, segmentation, and entity analysis are quickly evolving and specializing to respective subdivisions of research. As NLP and text analysis capabilities grow increasingly robust, so will other tools of data science, including those analyzing quantitative data, extracting graphs and plots, and deciphering images. In combination with these other techniques, NLP and text analysis in materials science has immense potential to exponentially accelerate innovation.

Works Cited

Kononova, Olga, et al. “Opportunities and Challenges of Text Mining in Materials Research.” *IScience*, vol. 24, no. 3, 19 Mar. 2021, https://doi.org/10.1016/j.isci.2021.102155.

Olivetti, Elsa A, et al. *Data-Driven Materials Research Enabled by Natural Language*. Applied Physics Reviews, 21 Dec. 2020, https://ceder.berkeley.edu/publications/2020-data-driven-research-nlp.pdf.