Scholarly Big Data: Computational Approaches to Semantic Labeling in Materials Science

Xintong Zhao, Jane Greenberg, Xiaohua Hu
College of Computing and Informatics, Drexel University, Philadelphia, PA USA
{xz485, jg324, xh92} @drexel.edu

Vanessa Meschke, Eric Toberer
Department of Physics, Colorado School of Mines, Golden, CO, USA
{vnilsen, etoberer} @ mines.edu

ABSTRACT

This paper explores computational, semantic labeling for scholarly big data in materials science. We report on a baseline comparative analysis involving ontology-based automatic indexing with the Helping Interdisciplinary Vocabulary Engineering (HIVE-4-MAT) application, using the RAKE algorithm, and the MATScholar system, which uses named entity recognition (NER), supported by an RNN (Recursive Neural Network). Results demonstrate that ontology-based automatic indexing requires less preparation time and provides useful output supporting recall; while NER/RNN requires greater preparation, but produces more precise labels that are likely better for deep learning.

KEYWORDS

Scholarly big data, Computational semantic labeling, Materials science, Ontology-based automatic indexing, Named Entity Recognition, Recursive Neural Network

1 Introduction

Scholarly big data continues to grow exponentially across all disciplines, with estimates predicting 35 million open access documents by 2022 (Wu and Giles, 2020). Materials science researchers recognizes this growth, and that it is impossible for a human to extract knowledge from the vast stores of published research (Weston et al., 2019). This challenge underscores the need for computational approaches to semantic labeling, which is a goal of the NSF-Harnessing the Data Revolution (HDR) initiative, Accelerating the Discovery of Electronic Materials through Human-Computer Active Search. This paper reports on baseline research in this area, specifically a comparative analysis exploring ontology-based automatic indexing and named entity recognition (NER), with two applications accommodating materials science.

2 Materials Science

Materials are an essential part of our everyday lives, from the metals and plastics that comprise smart phones and their encasings, to plastic, glass, and paper for food packaging. Materials science is an interdisciplinary field bringing together chemistry, engineering, physics, and other disciplines to study properties and discover cheaper, more functional, and less harmful materials. Materials science, like every other discipline, has increasingly embraced computation; however, the massive store of unstructured, scholarly big remains untapped. Research in computational semantic labeling can help address this challenge.

3 Computational Semantic Labeling

Semantic labeling is a broad area covering the assignment of descriptors representing *topicality* and the identification of *named entities*. Semantic labels are important for resource discovery and deep learning. We describe two common, computational approaches below.

3.1 Ontology-based automatic indexing

Ontology-based automatic indexing can support semantic labeling. The process involves automatic indexing to extract key terms from a document; followed by matching these initial results to terms encoded in a knowledge structure, such as an ontology. Automatic indexing sequence generally involves term frequency counts; the most popular method is term frequency–inverse document frequency (tf–idf) to determine the significance of a 'term' or 'phrase' in a document, and in comparison to the full corpus. Automatic indexing can also involve more sophisticated information retrieval methods, such as term weighting, similarity and probabilistic measures, and clustering (Melucci and Baeza-Yates, 2011).

3.2 Named Entity Recognition (NER)

NER is a subtask of Information extraction (IE) that can support semantic labeling. NER involves natural language processing to detect named entities (*e1* and *e2*) and their type in a sentence (*S*). Entity include personal or organization names, a geographic area, chemicals and so forth. Today's state-of-the-art NER involves Neural Network models (Jia et al., 2019), which have been extremely effective supporting drug discovery (Wan and Poon, 2018); and this approach has recently attracted attention material science community (Weston et al., 2019)

Overall, both approaches reviewed here motivate the research goals and objectives posited below.

4 Research Goals and Objectives

The goal of this research is to explore computational approaches for semantic labeling of unstructured, scholarly big data. Specific objectives are to:

1) Evaluate the performance of *ontologies-based automatic indexing* and *NER* for semantic labeling.
2) Consider how the two approaches may be improved to enhance knowledge discovery in materials science.

5 Method and Procedures

We conducted a baseline comparative analysis to explore two approaches supporting semantic labeling.

Test Applications and Algorithms

We selected *HIVE-4-MAT* and the *MatScholar* given that each application supports computational semantic labeling in materials science.

> *HIVE-4-MAT*[1] is a linked data automatic indexing application. HIVE-4-MAT builds off the HIVE system incorporated into the DataNet Federation Consortium's iRODS system (Conway, et al, 2013). The HIVE-4-MAT prototype includes the following four ontologies: 1) BioAssay Ontology, 2) Library of Congress Subject Headings (LCSH), 3) Smart Appliances REFerence ontology (SAREF) and 4) US Geological Survey (USGS) terminology, and supports automatic indexing with the RAKE (Rapid Automatic Keyword Extraction) algorithm. RAKE is an unsupervised algorithm that processes and parses text into a set of candidate keywords based on co-occurrence (Rose et al., 2010). Once the list of candidate keywords is selected, the HIVE system matches candidate keywords with terms from ontologies.
>
> *MATScholar*[2] is a NER web-accessible application supporting entity extraction and classification, and uses RNN/LSTM (Recurrent Neural Network-Long Short Term Memory). RNN-LSTM is a classic type of neural network that is widely applied in natural language processing tasks (Jia et al., 2019; Miwa and Basal, 2016). MATScholar's NER algorithm is supported by a training set of 800 hand-annotated abstracts and uses color and codes to identify seven entity classes: 1) inorganic material (MAT), 2) symmetry/phase label (SPL), 3)

[1] HIVE: http://hive2.cci.drexel.edu:8080/

[2] MATScholar: https://www.matscholar.com

sample descriptor (DSC), 4) material property (PRO), 5) material application (APL), 6) synthesis method (SMT), and 7) characterization method (CMT).

Sample and Evaluation

The sample included a set of nine randomly selected abstracts, drawn from MATScholar, which includes a collection of over 3 million abstracts drawn from the from the Scopus API.

The sample size of nine, while small, was considered sufficient for this baseline study, given human evaluation requirements, and the goals to assess the performance and value of each algorithm for semantic labeling. The evaluation was conducted by information science and materials science experts included 3-tier scale of relevant (R), partially relevant (PR), and non-relevant (NR) HIVE-4-MAT outputs; while the F score was calculated for results with MatScholar.

6 Results

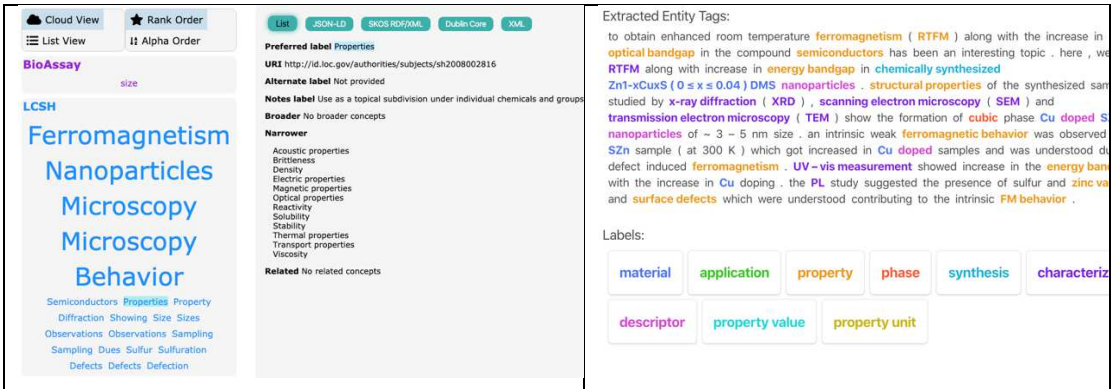
The results include an examination of both the output and performance evaluation. *Figure* 1 incudes an example of unstructured, scholarly big data form the sample.

Example: Input Abstract

To obtain enhanced room temperature ferromagnetism (RTFM) along with the increase in optical bandgap in the compound semiconductors has been an interesting topic. Here, we report RTFM along with increase in energy bandgap in chemically synthesized Zn1−xCuxS ($0 \leq x \leq 0.04$) DMS nanoparticles. Structural properties of the synthesized samples studied by X-ray diffraction (XRD), scanning electron microscopy (SEM) and transmission electron microscopy (TEM) show the formation of cubic phase Cu doped ZnS nanoparticles of ~3–5 nm size. An intrinsic weak ferromagnetic behavior was observed in pure ZnS sample (at 300 K) which got increased in Cu doped samples and was understood due to defect induced ferromagnetism. UV–vis measurement showed increase in the energy bandgap with the increase in Cu doping. The PL study suggested the presence of sulfur and zinc vacancies and surface defects which were understood contributing to the intrinsic FM behavior. (Patel et al., 2017, Effect of impurity concentration on optical and magnetic properties in ZnS:Cu nanoparticles)

Figure 1. Example Input

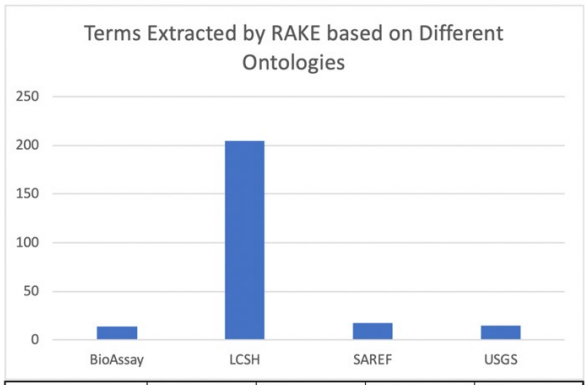
Figure 2 includes output from each application and their algorithm. HIVE-4-MAT/RAKE (left-hand side) presents a list of terms drawn from the ontologies. The hierarchical structure helps determine contextual meaning. MATScholar/RNN-LSTM results (right-hand side), presents the color encoded, labeled entities.



HIVE-4-MAT/RAKE	MATScholar/RNN-LSTM

Figure 2. Outputs from Each Application

Figure 3 shows the number of potentially relevant terms extracted with RAKE varies from 14 to 18 for the three ontologies (BioAssay, USGS, and SAREF), and 205 terms from LCSH.



Ontology Name	BioAssay	LCSH	SAREF	USGS
Num Terms	14	205	18	15

Figure 3. Number of Terms Extracted by Ontologies

Relevance results indicate that 14 BioAssy BioAssay ontology tersm were considered relevant—on some level and slightly over half the LCSH terms (51.71%) were considered "relevant," with 18.54% "partially relevant" Additionally, SAREF and USGS ontology, were promising, with 83.33% and 73.33% of terms considered "relevant" to the abstracts respectively.

MatScholar results are reported in *Table 1,* which calculates the extraction accuracy of each category (entity), followed by a total accuracy including all entities.

Accuracy(f1)	Total	MAT	SPL	DSC	PRO	APL	SMT	CMT
Development Set	87.09	92.58	85.24	91.4	80.19	80.6	81.32	86.52
Test Set	87.04	90.3	82.05	92.13	83.19	80.63	81.37	86.01

Table 1. Output Accuracy f1 score (%) of Entity Extraction Algorithm (Weston et al., 2019)

The f1 score as the indicator for accuracy, a measure that widely used in information retrieval and NER. The accuracy falls into the range from 80.19% to 92.13%.

7 Discussion

The initial results reported on in this paper give insight into ontology-based automatic indexing and entity extraction algorithms for semantic labeling in materials science. Results demonstrate that ontology-driven algorithm provide general-relevant labels, while entity extraction algorithm may provide more precise labels. In assessing if we should use one approach over the other, it depends on the circumstance and the broader goals of leveraging semantic labels.

The two approaches studied have different requirements and costs. Developing a neural network model requires the acquisition and labeling of large amounts of data, incurring a cost; although precise labels generally have greater value for deep learning. Compared to RNN model, RAKE algorithm requires significantly less data: the list of candidate keywords is simply extracted based on the given abstract, regardless of size. Current results are further challenged by limited availability materials science ontologies, particularly compared to biomedicine. More ontologies may vastly improve these results and prospects.

8 Conclusion

Overall, our research suggests that both approaches have value. Ontology-based automatic indexing requires less preparation time and provides useful output supporting a view of the field and recall; and NER/RNN requires greater preparation, but more precise labels will better support deep learning. The results here are part of a baseline study, and more research is underway involving a larger data-set. Given limitations with materials science ontologies, we are also looking toward relation extraction and building robust ontologies to improve computational label and, ultimately accelerate the discovery of new materials.

ACKNOWLEDGMENTS

NSF/OAC-Office of Advanced Cyberinfrastructure #1940239.

REFERENCES

Conway, M. C., Greenberg, J., Moore, R., Whitton, M., & Zhang, L. (2013). Advancing the DFC Semantic Technology Platform via HIVE Innovation. Proceedings of the 7th Metadata and Semantics Research Conference. Thessaloniki, Greece, November 19-22, 2013,pp. 14-21. https://doi.org/0.1007/978-3-319-03437-9_3.

Jia, R., Wong, C., & Poon, H. (2019). Document-Level N-ary Relation Extraction with Multiscale Representation Learning. Proceedings of the 2019 Conference of the North. Presented at the Proceedings of the 2019 Conference of the North. https://doi.org/10.18653/v1/n19-1370

Melucci, M., & Baeza-Yates, R. (Eds.). (2011). Advanced topics in information retrieval (Vol. 33). Springer Science & Business Media.

Miwa, M., & Bansal, M. (2016). End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). https://doi.org/10.18653/v1/p16-1105

Rose, S., Engel, D., Cramer, N., & Cowley, W. (2010). Automatic Keyword Extraction from Individual Documents. In Text Mining (pp. 1–20). https://doi.org/10.1002/9780470689646.ch1

Wang, H., & Poon, H. (2018). Deep Probabilistic Logic: A Unifying Framework for Indirect Supervision. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. https://doi.org/10.18653/v1/d18-1215

Weston, L., Tshitoyan, V., Dagdelen, J., Kononova, O., Trewartha, A., Persson, K. A., Ceder, G., & Jain, A. (2019). Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature. Journal of Chemical Information and Modeling, 59(9), 3692–3702. https://doi.org/10.1021/acs.jcim.9b00470

Wu, J., & Giles, C. L. (2020). Scholarly Very Large Data: Challenges for Digital Libraries (White Paper): https://tigerprints.clemson.edu/cgi/viewcontent.cgi?article=1021&context=hugedata