ACM/IEEE JCDL 2020 Workshop 4: Organizing Big Data, Information, and Knowledge, Aug 5th, 2020

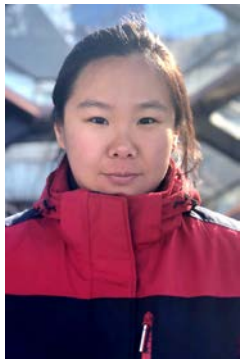
NSF-HDR# 1940239

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

Author(s): Xintong Zhao[1], Jane Greenberg[1], Xiaohua Hu[1], Vanessa Nilsen[2], Eric Toberer[2]

[1]College of Computing and Informatics, Drexel University, USA

[2]Department of Physics, Colorado School of Mines, USA



Xintong Zhao

Jane Greenberg

DREXEL UNIVERSITY Metadata Research Center College of Computing & Informatics

Outline

- Introduction
- Research Questions
- Methods and Procedures
- Results
- Discussion
- Current Progress
- Future Work

Background & Motivation

- The discovery, design and development of new materials are major tasks in the interdisciplinary research of *Materials Science*
  - *From metals and plastics in daily life to advanced materials for collecting clean energy*
- We have huge volume of information available related to materials(Cheung et al., 2009; Ashino, 2010; Weston et al., 2019)
  - Millions of published academic literature
  - Numerical experiment data
  - Computational methods
  - Material properties, processing methods and structures

Current Challenge

The volume of academic articles is too large for researcher to fully read even a portion of papers in their lifetime;

- The use of time becomes inefficient
- Hard to accurately retrieve needed information in short time

Potential Solution: Semantic Labeling for Knowledge

The following could be a sound direction:

The plan is to label most important information for materials researchers from academic literature

Use relation extraction to construct knowledge base

Our goal is to build an "expert system" that helps researchers to locate key information from huge number of text data in short time

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

Ideal outcome

Input:

"Hey system, what are common materials that have thermoelectric property?"

Output:

return integrated information containing the N most frequent materials + related properties/applications + list of papers

Involved Methods

- Named Entity Recognition (NER)
  - NER is a subtask of Information extraction (IE) that can support semantic labeling. NER involves deep learning to detect named entities and their type in a sentence.
  - Since it's supervised learning, a large training set is required

- **Traditional Machine Learning Algorithms for Keyword Extraction**
  - 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.
  - There are multiple algorithms, we take the **RAKE** (Rapid Keyword Extraction) as an example
  - Un-supervised learning, which does not require training set.

Research Questions

- How well those two approaches can be in materials domain?
- Compare to entity recognition which applies deep learning methods, machine learning algorithms for keyword extraction are no longer the hottest topic. Are deep learning methods always better than early machine learning methods? Should we use only one, or we should keep both?
- What are the differences between these two approaches?
- How they can be improved in order to further support materials discovery?

Methods and Procedures

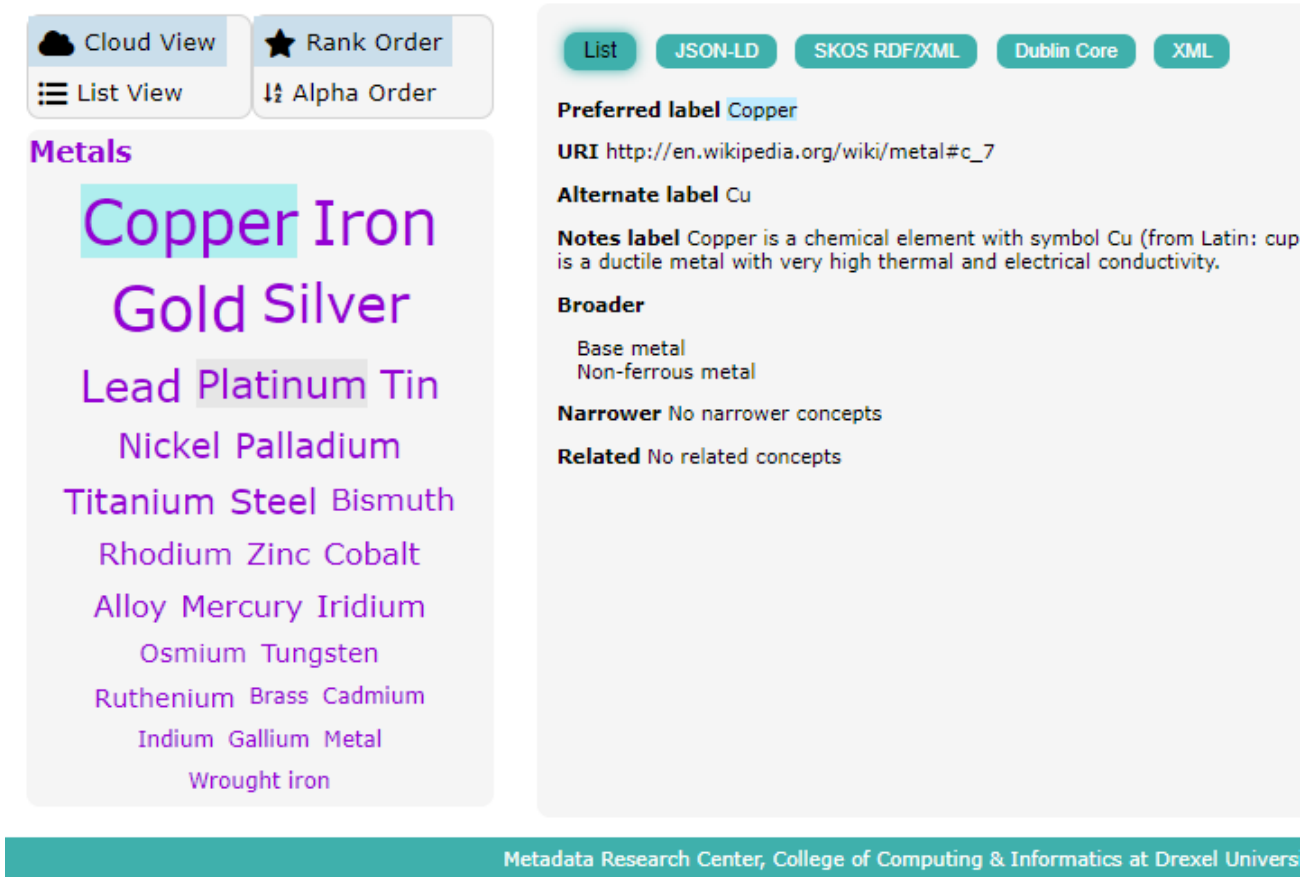
- We conducted an early-phase comparative analysis to explore two approaches supporting semantic labeling.
- We selected two applications for comparison: *HIVE-4-MAT* and the *MatScholar (Weston et al., 2019)*, where *HIVE-4-MAT* carries *RAKE* and *MatScholar* has NER model deployed.

Methods and Procedures (Cont'd)

- We randomly selected 9 abstracts in the database of MatScholar *(Weston et al., 2019)*
  - Abstracts are collected from sources such as Scopus and ScienceDirect APIs, the Springer-Nature API
  - Only English articles are selected
  - Only articles about inorganic materials are selected
  - Selected papers should at least contain one from "structure, property and processing".
- We take the 9 abstracts as input to both applications, and compare the outputs

Methods and Procedures (Cont'd)

- HIVE-4-MAT: we design it as a linked data automatic indexing application, and it is still under construction; it builds off the original HIVE (Zhang et al., 2015) system developed earlier at Metadata Research Center of Drexel University.
  - Original HIVE: http://hive2.cci.drexel.edu:8080/



(Zhang et al., 2015)

Methods and Procedures (Cont'd)

Extracted Entity Tags:

In this study , we attempted to reduce **firing** voltage of **ac - PDPs** by **alloying** **MgO** **electron emission** material with **OZn** . this approach was aimed to reduce **band gap energy** of **MgO** by the **alloying** and thereby promote the auger neutralization reaction of xe+ ions on **MgO** **surface** . **pellets** were prepared by **sintering** **MgO** and **OZn** **powder** mixture at 1300 Â° C for 8 h under nitrogen atmosphere . test panels with such **alloyed** **MgO** **films** showed significantly reduced **firing** voltages , especially when the discharge gas is of high **Xe content** . these results represent a new way of approaching in the development of **electron emission** materials for **ac - PDPs** .

Labels:

material | **application** | **property** | **phase** | **synthesis** | **characterization**

descriptor | **property value** | **property unit**

(Weston et al., 2019)

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

Results

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)

Result (Cont'd)

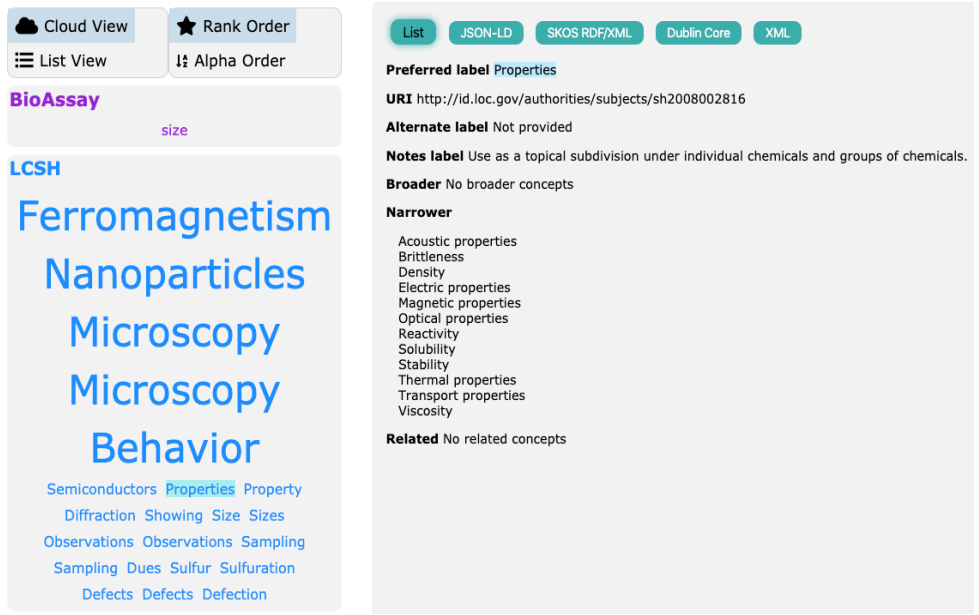
Extracted Entity Tags:

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 ≤ x ≤ 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 SZn nanoparticles of ~ 3 – 5 nm size . an intrinsic weak ferromagnetic behavior was observed in pure SZn 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 .

Labels:

material | application | property | phase | synthesis | characterization | descriptor | property value | property unit

(Weston et al., 2019)



(Zhang et al., 2015)

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

Discussions

Algorithms	Extracted Terms
RAKE (Keyword Extraction)	Ferromagnetism, Nanoparticles, Microoscopy,Behavior, Semiconductors, Properties, Property, Diffraction, Showing, Size, Sizes, Observations, Sampling, Dues, Sulfur, Sulfuration, Defects, Defection
RNN (Entity Extraction)	ferromagnetism(RTFM), optical bandgap, semiconductors, RTFM, energy bandgap, chemically synthesized, Zn1-xCuxS(0≤x≤0.04)DMS, nanoparticles, structural properties, x-ray diffraction (XRD), scanning electron microscopy (SEM), transmission electron microscopy (TEM), cubic, Cu, SZn, ferromagnetic behavior, doped, UV-vis measurement, PL, zinc vacacies, surface defects, FM behavoir

- RAKE algorithm delivers general-relevant labels
- As expected, LSTM-CRF generates much more precise labels

Discussion (Cont'd)

- The two approaches studied have different requirements and costs.
- Developing a neural network model requires the acquisition and labeling of large amounts of data, which can be expensive.

Current Progress

- Building our own Entity Recognition Model with LSTM+CRF structure

Label	precision	recall	f1-score	support	
I-DSC	0.78	0.79	**0.78**	98	
B-PRO	0.8	0.69	**0.74**	772	
O	0.92	0.96	**0.94**	9605	
I-PRO	0.83	0.63	**0.72**	774	
B-DSC	0.87	0.84	**0.85**	437	
I-CMT	0.79	0.77	**0.78**	248	
B-CMT	0.76	0.82	**0.79**	195	
B-SMT	0.83	0.73	**0.78**	171	
I-MAT	0.49	0.62	**0.55**	397	
B-MAT	0.82	0.71	**0.76**	682	
B-SPL	0.75	0.51	**0.6**	75	
I-SMT	0.8	0.75	**0.77**	219	
I-APL	0.67	0.65	**0.66**	135	
B-APL	0.81	0.46	**0.59**	170	
I-SPL	0	0	**0**	15	
Total				13993	
Weighted Avg	0.875628529	0.876159508	**0.8737026**		

Current Progress

- We are expanding the text data from inorganic materials only to both organic and inorganic materials.
- We are creating a dataset for relation extraction

Future Work

- Apply relation extraction to construct a comprehensive knowledge base on materials.

Thank you for watching!

DREXEL UNIVERSITY
Metadata
Research Center
College of Computing & Informatics

Reference

- Ashino, T. (2010). Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. Data Science Journal, 9, 54-61. doi:10.2481/dsj.008-041
- K. Cheung, J. Hunter and J. Drennan, "MatSeek: An Ontology-Based Federated Search Interface for Materials Scientists," in IEEE Intelligent Systems, vol. 24, no. 1, pp. 47-56, Jan.-Feb. 2009, doi: 10.1109/MIS.2009.13.
- 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
- Zhang, Y., Greenberg, J., Ogletree, A., and Tucker, G. (2015). Advancing Materials Science Semantic Metadata via HIVE. International Conference on Dublin Core and Metadata Applications, p. 209-211: https://dcpapers.dublincore.org/pubs/article/view/3783