Acceleration of Quantum Materials Discovery Driven by Machine Learning and Text Data Mining

Xianjun Yang[1], Julia Zuo[2], Linda Petzold[1], Stephen Wilson[2]

University of California, Santa Barbara Department of Computer Science[1] Department of Materials[2]

UC SANTA BARBARA Quantum Foundry DMR-1906325

Abstract

1. Traditional materials discovery methods involve massive expense of calculation or experiments due to the complexity of experimental condition combinations[1]
2. Machine learning(ML) has recently demonstrated powerful capabilities for finding useful hidden information from huge and noisy datasets[2]
3. Currently there are not sufficient experimental records to harness the power of machine learning to speed up materials discovery[3]
4. Thus we are building an infrastructure for data mining from the literature, using Natural Language Processing (NLP)[4]

NLP

1. **Named Entity Recognition:** the first step towards information extraction that seeks to locate and classify named entities in text into pre-defined categories
2. **Relation Extraction:** the task of predicting attributes and relations for entities in a sentence

Previous study

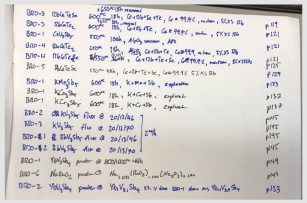
1. **Text mining** has been widely used in biomedical domain literature, but few applications exist in materials domain
2. **State-Of-The-Art** results:
   1. Solid Oxide Fuel Cells[5]: competitive baseline work for information extraction
   2. Titania nanotubes[6]: a pipeline for text mining in materials domain

Our focus:

1. **Material:** quantum material
2. **Method**: explore better NLP architecture

Data

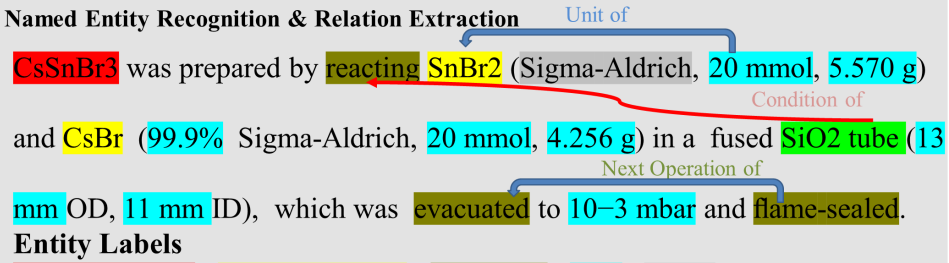
1. One specific material: 63 GaNb4Se8 synthesis records at Wilson Lab



Lab Notes

2. Text corpus: over 20k published papers related to quantum materials downloaded from different journal websites

Named Entity Recognition & Relation Extraction



Entity Labels

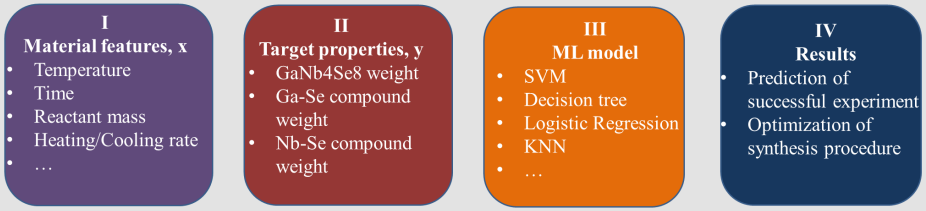
Target-Material Recipe-Material Operation Unit Brand More…

Relation Labels

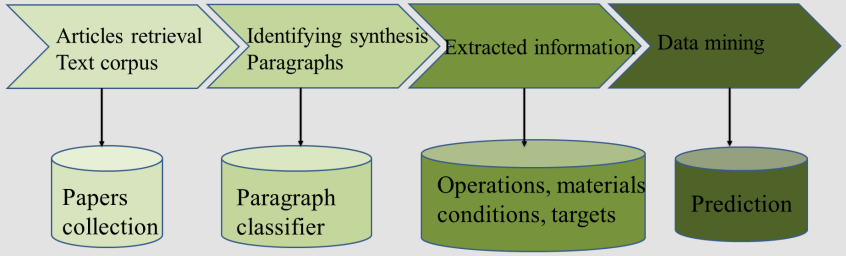
Condition of Unit of Next operation of More…

Methodology

1. ML for synthesis of GaNb4Se8:



2. NLP for text mining



Step 1: Web scraping with approved API
Step 2: Word embeddings + LR, RNN, BiLSTM, SciBERT
Step 3: Named Entity Extraction, Relation Extraction, Graph Neural Network
Step 4: Text mining and interpretation

Results

1. GaNb4Se8 synthesis prediction: 85% accuracy but not reliable due to small dataset
2. Text mining:
   - Step 1: ~20k corpus
   - Step 2: ~ 90% accuracy
   - Step 3: in progress
   - Step 4: coming soon

Future Plan

1. GaNb4Se8 synthesis prediction: improve dataset size
2. Text mining: improve NER and RE accuracy
3. Interpret the results and provide guidance for future synthesis
4. Text mining for literature abstracts
5. Transfer learning for text mining in other materials domains

References

[1] Raccuglia *et al.*, (2016)
[2] Frey *et al.*, (2019)
[3] Weston *et al.*, (2019)
[4] Kononova *et al.* (2019)
[5] Friedrich *et al.*(2020)
[6] Kim *et al.* (2017)