Extracting synthesis procedure from solar cell perovskite based scientific publications using Natural Language Processing.

Sourin Dey

November 20, 2022

1 Abstract

Finding promising materials to produce solar cell devices are a challenging optimization step. This project tries to garner such information as much as possible by intelligently exploring scientific literature using NLP. A pipeline of NLP steps are developed involving database generation, fine-tuning BioBERT model for Chemical Named Entity Recognition (CNER) and finally Relation extraction from scientific articles. Different pre-trained models such as BERT, materials to vector, Bio-BERT have been tested. The BioBERT provided the better performance in detecting CNER and thus paving the way of chemical synthesis information extraction.

2 Introduction

Towards making the tech dominated world pollution free amid the decaying reserve of fossil-fuel, solar cell design has become a rich domain of research. There is abundance of literature published by different labs that produced and reported solar cell experiments. This valuable information has been largely unexplored which has the potential to tell us which materials can be used and which materials will produce bad solar cells. This information can guide solar cell researchers for information rich optimization experiments [1]. This motivated me to work in this domain of exploring scientific articles and extracting useful information from solar cells. NLP has evolved since last 10 years greatly and revolutionized the domain of text analysis. Previously, different descriptors were being used which were replaced by advent of Recurrent Neural Networks (RNN). Due to inability of long range sentence relation learning and vanishing gradient problem, LSTM (Long Short Term Memory) replaced RNN but it still lacked learning context of a sentence because it used directional learning approach and was very slow. Such issues resulted in the development of transformer models and especially BERT (Bidirectional Encoder Representations from Transformers) [2].

3 Problem

Goal: Whether the pipeline is able to capture synthesis information.

Sample Input

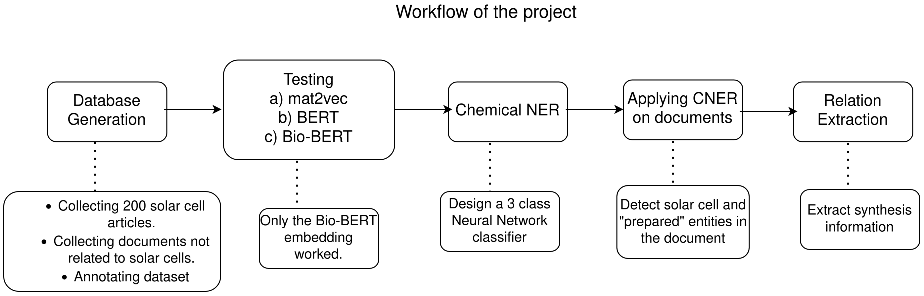
"Replacement of lead in the hybrid organic–inorganic perovskite solar cells invokes the need for non-toxic materials such as Sn. Although solution processed CsSnI3 has been demonstrated as a lead-free halide perovskite which can function as a light absorber with high photocurrent densities, the power conversion efficiencies were bottlenecked by low open circuit voltages. In this work, the open circuit voltages are modulated by chemical doping of CsSnI3 with Br leading to formation of CsSnI3xBrx (0 x 3) perovskites. The beneficial effect of Br incorporation for Voc improvement is evident for CsSnI3 system even without the addition of SnF2. There is an evolution of the crystal structure of CsSnI3 from orthorhombic to cubic for CsSnBr3 accompanied by changes in its optical properties with a blue shift of the absorption and IPCE onset, as the Br– doping is increased. The Voc enhancement is attributed to the decrease in Sn vacancies which is reflected by the lower charge carrier densities of 1015 cm–3 and a high resistance to charge recombination in case of Br rich CsSnI3xBrx perovskite. By the addition of SnF2 to CsSnI3xBrx perovskite, the current densities are improved significantly."

Sample Output *"open circuit voltages modulated chemical doping cssni br leading formation cssni xbrx perovskites."*

This means CsSnI material results in the formation of $\mathrm{CsSnI}_x Br_x$ when open circuit voltage is modulated by chemical doping.

4 Related Works

To extract chemical experiments, ChemTagger in Java programming platform was developed. However, that module requires the exact sentence containing synthesis information as input. Sometimes, the synthesis procedure can be spanned through more than one sentence. These are shortcomings of ChemTagger. For that reason, this project aimed at implementing Chemical NER or CNER first. CNER has been previously used to detect organic and inorganic materials using different BERT models [3]. NER models have also been developed for text analysis in biomedical domain where the classifier was able to recognize different drugs and diseases related to specific treatment [4]. These CNER models are usually trained by fine-tuning the BioBERT or SciBERT models because the vanilla BERT fails to provide rich information about contexts of diseases and drugs while the earlier to can efficiently provide [5]. Besides the wide array of application of NER classifiers, it has one major application in finding sentences of specific interests and then extracting relation of the sentence which is called relation extractions. These works have been implemented in different domains for text analysis. One similar example is the research at [6] where the authors implemented the NER and relation extraction to find interaction between drugs and diseases. Pipeline of this project follows the identical path although I have tried different approach of fine-tuning initially which did not provide promising outcome.


Workflow of the project
Database Generation
Testing
a) mat2vec
b) BERT
c) Bio-BERT
Chemical NER
Applying CNER on documents
Relation Extraction
• Collecting 200 solar cell articles.
• Collecting documents not related to solar cells.
• Annotating dataset
Only the Bio-BERT embedding worked.
Design a 3 class Neural Network classifier
Detect solar cell and "prepared" entities in the document
Extract synthesis information

4.1 Approach

In my project, the first thing to detect is chemical entities recognition [7] i.e. whether the entity is a new or already manufactured solar cell, whether it is produced. Thus for the CNER task, I had to design a 3 class classifier which should classify solar cells, manufacture synonymous words and all other words or materials. After developing this model, the model becomes able to detect the sentence containing 2 entities: (a) solar cell and (b) words similar to added or grown or manufactured. This step helps to narrow down the sentence relation extraction job quite easier and reduces redundancies. It is because not every sentence contains synthesis information. Some sentences may contain synthesis information about other materials which are not solar cell and the entity recognizing classifier should be capable in detecting that as well. After the sentences containing synthesis extracting information are detected, sentence relation extraction is implemented which produces the solar cell synthesis summary.

4.1.1 Database Generation and Annotation

I have collected 200 solar cell articles mentioned by the research counducted in [1]. Due to the timing issue of project implementation, the abstracts of these articles were extracted. Then basic NLP preprocessing steps were followed. One very important task in this step is stemming words. This enables to reduce variation of same words in embedding generation. The words containing solar materials like MaPbI_3, perovskites, halide perovskites are annotated as label B while words like “grown”,“produced”,“fabricated” are annotated as label I and all other words are annotated as O which is neither B nor I class. Thus, the CNER turned into a multi class classification problem.

4.1.2 Chemical Named Entity Recognition

4.1.3 mat2vec model [8]

For the implementation, I first tried the embedding provided by materials to vector (mat2vec) [8] model which similar to word2vec model [9]. This mat2vec model was exposed to chemical scientific literature during training and it is capable of mapping similar chemical terms or molecules or any similar entities to closer area in embedding spaces. This motivated me to use this embedding space to detect solar cells and the another entity from the literature. However, this model performed very poor in doing the task.

4.1.4 vanila BERT model

Similar to mat2vec, vanilla BERT model was also unable to perform well and gave very poor result. This is because BERT was mostly exposed to non-chemical or non-biomedical domain data and thus it was difficult for this model to make distinction among chemical entities. The results are mentioned in Table 1.

4.1.5 BioBERT model [10]

This model outperformed the other models in terms of performance of the CNER task because this was trained using biomedical texts of different biomedical domains. The authors mentioned that it outperformed other language models significantly in biomedical NER, relation extraction as well question answering tasks. This motivated me to take their embedding i.e. finetuning BioBERT model.

4.1.6 Sentence Summarization

This step comes after applying CNER on test cases. This can be exemplified with the following worked out instance. Let’s pick an abstract from an solar cell article as follows:

“Replacement of lead in the hybrid organic–inorganic perovskite solar cells invokes the need for non-toxic materials such as Sn. Although solution processed CsSnI3 has been demonstrated as a lead-free halide perovskite which can function as a light absorber with high photocurrent densities, the power conversion efficiencies were bottlenecked by low open circuit voltages. In this work, the open circuit voltages are modulated by chemical doping of CsSnI3 with Br leading to formation of CsSnI3-xBrx ($0 \leq x \leq 3$) *perovskites. The beneficial effect of Br*

incorporation for Voc improvement is evident for CsSnI3 system even without the addition of SnF2. There is an evolution of the crystal structure of CsSnI3 from orthorhombic to cubic for CsSnBr3 accompanied by changes in its optical properties with a blue shift of the absorption and IPCE onset, as the Br– doping is increased. The Voc enhancement is attributed to the decrease in Sn vacancies which is reflected by the lower charge carrier densities of 1015 cm–3 and a high resistance to charge recombination in case of Br rich CsSnI3-xBrx perovskite. By the addition of SnF2 to CsSnI3-xBrx perovskite, the current densities are improved significantly."

Now CNER classifies each token of this sentence and then a simple if else program finds out the sentence containing both class B and class I as follows:

"*Although solution processed CsSnI3 has been demonstrated as a lead-free halide perovskite which can function as a light absorber with high photocurrent densities, the power conversion efficiencies were bottlenecked by low open circuit voltages. In this work, the open circuit voltages are modulated by chemical doping of CsSnI3 with Br leading to formation of CsSnI3-xBrx* ($0 \leq x \leq 3$) *perovskites."*

Both of the above two sentences contain the solar cell entity "$CsSnI_3$" (class B) and "formation" entity which is class I. Now summarization outputs the result.

5 Evaluation

For Chemical NER tasks, I added 200 solar cell abstract sentences to the already existed BC4CHEM database [10] which results in a total of 898599 tokens in the training data where the distribution looks like follows:

- Solar materials and other chemicals are assigned label B (total 29681 in training)
- stem(Produce,Fabricated,Manufactured,Processed,Synthesized) are assigned label I (total 38857 in training)
- all other tokens are assigned O (total 830061 tokens in training)

The CNER test result contains only another 200 solar cell articles. The test data numbers are written along with model performances in the following table.

Table 1: mat2vec, BERT and BioBERT pretrained model based performance on CNER task

Fine-Tuned On	Class	#Data points	Precision	Recall	F1
mat2vec	B	203	0.44	0.48	0.46
mat2vec	I	204	0.53	0.58	0.55
mat2vec	O	4507	0.78	0.77	0.774
BERT	B	203	0.49	0.42	0.45
BERT	I	204	0.42	0.45	0.43
BERT	O	4507	0.82	0.78	0.799
Bio-BERT	B	203	0.68	0.67	0.67
Bio-BERT	I	204	0.64	0.65	0.645
Bio-BERT	O	4507	0.88	0.85	0.864

Finally, the BioBERT based NER model is picked to classify the chemical entities and then a simple program identifies the solar cell synthesis sentences. Out of 200 test documents, 180 of them contained synthesis information. The model could correctly identify 115 documents and misidentified 11 non-synthesis documents as synthesis documents.
Precision= $\frac{TP}{TP+FP} = \frac{115}{115+65} = 64\%$
Recall= $\frac{Tp}{TP+FN} = \frac{115}{115+11} = 91\%$
As this similar identification method has not been implemented yet, so I am not able to make comparison to other baseline models as they don't exist. However, there are biomedical disease detecting NER tasks models available fine-tuned on BioBERT which provides 91% precision and 92% recall on BC4CHEM database.

The model could be improved significantly with increase of traning sample as this annotation process was not very efficient and at the same time it consumes high amount of time.

6 Conclusion

The fine tuned Bio-BERT model is capable to detect chemical entities but is not capable to successfully detect solar cells. Besides, more works are needed to make it more usable. For example, the word 'grown' is used when solar cell material is developed and the word 'fabricated' is used when solar cell device is manufactured. This two sentences have very different output and one might be interested in knowing the material processing information and not the device processing information or vice-versa. This project workflow can be further employed to dig deeper into a material synthesis article. One may be need information about the corresponding efficiency of the synthesized material. In that case the material entity and efficiency should be detected by the model. Similarly, it can help finding out catalysis used in the synthesis procedure. Thus, this project has several potential in material science domain.

References

[1] T Jesper Jacobsson, Adam Hultqvist, Alberto García-Fernández, Aman Anand, Amran Al-Ashouri, Anders Hagfeldt, Andrea Crovetto, Antonio Abate, Antonio Gaetano Ricciardulli, Anuja Vijayan, et al. An open-access database and analysis tool for perovskite solar cells based on the fair data principles. *Nature Energy*, 7(1):107–115, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Max E Savery, Willie J Rogers, Malvika Pillai, James G Mork, and Dina Demner-Fushman. Chemical entity recognition for medline indexing. *AMIA Summits on Translational Science Proceedings*, 2020:561, 2020.

[4] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48, 2017.

[5] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

[6] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, page 673, 2020.

[7] David Nadeau and Satoshi Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[8] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[9] Xin Rong. word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.