Design Laboratory

Neural Approaches to Information Extraction in the Materials Science Domain

Under the supervision of

Prof. Manjira Sinha

Group members

Name	Roll No.
Leo Lorence G	19NA3AI29
Ajay Ram Meena	19IE3AI18

Contents

Contents

1 Introduction

Materials science is an area that is growing quickly and has a lot of research writing about it. It is a big task to find useful information and insights in this large group of study papers. Even though generic language models like BERT have been successful in a wide range of natural language processing (NLP) tasks, they often fail when applied to domain-specific tasks because they don't have the specialized knowledge and understanding of the unique context and terminology used in specialized fields. In order to solve this problem, MatSciBERT, a domain-specific language model for materials science, was made. This was done because text mining and information extraction in materials science books needed to be done in a more focused and effective way. MatSciBERT is made to help researchers in materials science understand the specialized words, ideas, and connections they use and get more information from them.

Text mining and information extraction in materials science are usually done with general language models, which might not be able to handle the subtleties and complexity of the field. So, these methods might not give the best results, which could make it harder to find new materials, qualities, and phenomena. MatSciBERT was made to get around these problems by making a language model that is just right for the materials science area. The method used in this study is to make a materials science-specific corpus (MSC), train MatSciBERT on the MSC, and then test its success on different materials science-related tasks. By taking this approach, the main goal is to make text mining and information extraction in materials science more efficient and accurate. This will help the field move forward by making it easier to analyze and understand its works.

2. Problem statement

Domain-specific vocabulary and notations provide considerable obstacles to correctly extracting information from texts in the field of materials science. Traditional manual methods need a great deal of expertise in the subject and take a long time to extract pertinent information from scientific sources.

A potential method for automatically extracting information from text is Natural Language Processing (NLP). However, prior techniques had poor performance due to their lack of domain-specific knowledge and inability to produce contextual embeddings. Further study on the development of NLP methods that incorporate domain-specific knowledge and

produce contextual embeddings is imperative to address these flaws and make it easier to extract pertinent information from materials science literature accurately and efficiently.

Researchers may develop more efficient information extraction tools by developing NLP approaches in this way, thereby accelerating discoveries and breakthroughs in the field of materials science.

3. Objective

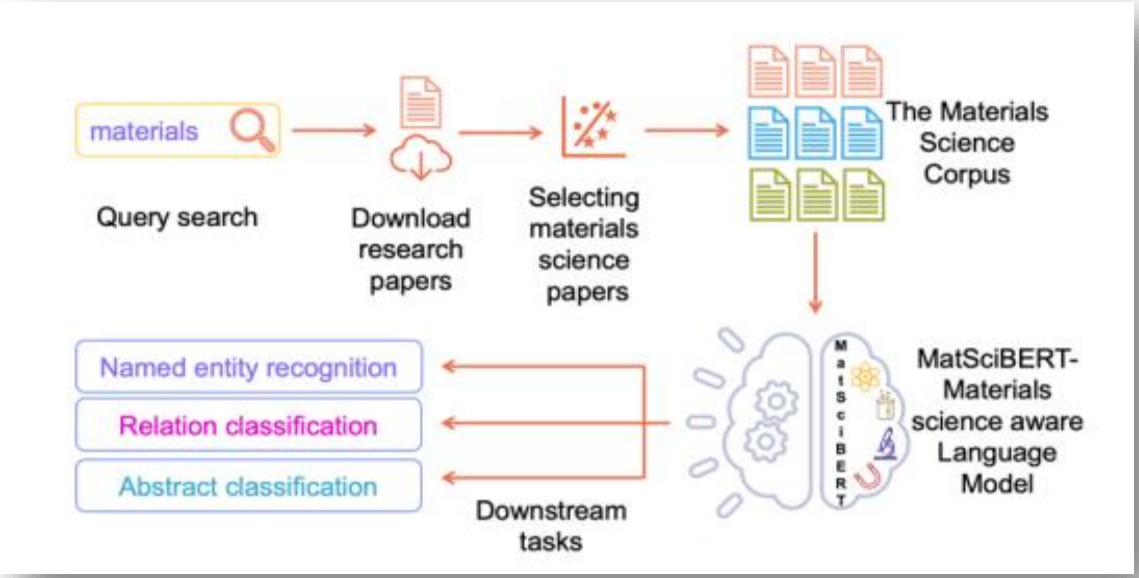
The creation of MatSciBERT, a language model created especially for materials science and trained on the Materials Science Corpus (MSC), is the suggested remedy. By successfully adjusting to numerous downstream tasks pertinent to materials science, this domain-specific approach seeks to considerably expedite research in the area. MatSciBERT is better prepared to address the particular difficulties provided by the literature on materials science by utilizing the specialized knowledge and contextual understanding acquired from the MSC. MatSciBERT has been improved to handle several important downstream jobs, including:

- Named Entity Recognition (NER):
- Relation Classification:
- Glass vs. Non-Glass Abstracts Classification:

MatSciBERT offers scholars a potent tool for gleaning important insights from the massive amount of materials science literature by taking care of these crucial downstream chores. Thus, new materials, properties, and phenomena are made possible, which quickens the field's advancement.

4. Methodology

Through query searches and the selection of pertinent research papers in the field of materials science, the Materials Science Corpus (MSC) was created. The usefulness of MatSciBERT, a specialized language model, for information extraction and text mining in the materials science sector, was demonstrated after it was pre-trained on the MSC and assessed on several downstream tasks. Details information is provided below



4.1 Materials Science Corpus (MSC)

A sizable dataset is needed in order to train a language model (LM) in a generalizable way. For instance, BERT received pre-training on 3.3 billion words from the English Wikipedia and Book Corpus. SciBERT, an LM trained on scientific literature, used a corpus including 82% publications in the biomedical field and 18% in the field of computer science. None of these LMs, however, include any content pertaining to the materials domain.

In this analysis, it was determined that the materials domain was adequately covered by four major categories of materials science literature: inorganic glasses and ceramics, metallic glasses, cement and concrete, and alloys. The Crossref metadata database was queried in the first stage, producing a list of more than 1 million articles. Using their authorized API, papers were obtained from the Elsevier Science Direct database. Because the research articles were in XML format, text extraction required a unique XML parser. The corpus contained both full-text articles and abstracts.

500 papers were personally annotated based on their abstracts to verify the downloads were relevant. After then, these labelled abstracts were used to fine-tune SciBERT classifiers to find pertinent publications from the 1 million downloaded articles. The language model was trained using the papers that were chosen from each category of resources. To evaluate the model's performance on unread text, the Materials Science Corpus (MSC) was split into training (85%) and validation (15%) sets.

There may be a variety of symbols and random characters in scientific writing, and some semantic symbols have several Unicode surface forms. Unicode normalization was carried out on MSC to remove random Unicode characters and map various Unicode characters with similar meanings and appearances to single standard characters or sequences of standard characters in order to remedy these discrepancies. Hugging Face's BertNormalizer from the tokenizers package was used to first normalize the corpus. The Unicode characters found in the MSC were then mapped to a list, which was then generated. To avoid interference during pre-training, random characters were mapped to space. Each dataset underwent this normalization process as well before being sent into the MatSciBERT tokenizer.

This thorough method of curating and analyzing the MSC made it possible to create MatSciBERT, a specialized language model designed specifically for the materials science field.

4.2 Pre-training of MatSciBERT

Pre-training a language model from the ground up requires a lot of computer power and data. To solve this problem, the SciBERT uncased vocabulary is used to tokenize MatSciBERT and initialize it using weights from SciBERT. By using this method, current models that rely on SciBERT may be utilized interchangeably with MatSciBERT, and the vocabulary already used in scientific publications can now be used to appropriately describe new terms in the materials domain.

The optimized RoBERTa training recipe is used to pre-train MatSciBERT and has been found to enhance the performance of the original BERT. Pre-training for MatSciBERT adopts the following modifications. Masking takes place at the word level rather than the wordpiece level in dynamic full word masking. The model predicts each masked wordpiece token independently after randomly masking 15% of the words.

Using Masked-LM and Next-Sentence Prediction (NSP), BERT was pre-trained with the goal of removing the NSP loss. It has been discovered that eliminating the NSP loss matches or slightly improves downstream task performance. BERT was previously trained with full-length sequences of varied durations. The RoBERTa writers improved performance by restricting training to full-length sequences. The [SEP] token is used to denote the separation of segments from different documents in input sequences.

4.3 Downstream Tasks

The language model underwent pre-training before being refined for use in particular downstream tasks. The model was trained to recognize domain-specific named entities in a phrase using the BIO scheme for the Named Entity Recognition (NER) problem, where each label denotes the start, middle, or end of a certain entity type. The model was trained to predict the directed connection between two entities in a given sentence for the Relation Classification challenge, where the input consists of a phrase and two entity spans. The Paper Abstract Classification job entailed determining whether or not a research paper's abstract is pertinent to a certain topic. The abstract served as the input, while the binary classification label was the output.

4.3.1 Named Entity Recognition (NER)

This study examines three different language model (LM) designs, each intended to enhance the model's handling of domain-specific information in materials science:

- **LM-Linear:** This architecture makes advantage of a BERT-based transformer model's output embedding of the first word portion of each token. To determine the entity, type of each token, the embeddings are routed through a linear layer with SoftMax activation.
- **LM-CRF:** This design improves on the LM-Linear by adding a Conditional Random Field (CRF) layer in place of the final SoftMax activation. The addition of the CRF layer improves the model's overall performance by allowing it to learn to label tokens that belong to the same entity and learn transition scores between other entity kinds.
- **LMBiLSTM-CRF:** This architecture introduces a stacked bidirectional Long Short-Term Memory (BiLSTM) layer between the LM and CRF layers. The inclusion of the BiLSTM layer enhances the model's ability to handle sequential information, making it more effective in capturing the context of materials science literature.

The models leverage BERT, SciBERT, and MatSciBERT as the underlying language models to implement these architectures. For the LM-Linear architecture, the transformers library's BERT Token Classifier implementation is used. The PyTorch-CRF library's CRF implementation is used for the LM-CRF architecture. The input embeddings are processed using a stacked BiLSTM before being sent to the CRF layer in the LMBiLSTM-CRF architecture.

4.3.2 Relation Classification (RC)

The Entity Markers-Entity Start design suggested by Soares et al. (2019) is used in the study to describe an architecture for relation categorization. The [E1] and [E2] special wordpieces are used in this manner to encapsulate entity spans. To forecast entity relationships, the concatenated output embeddings of these markers are fed via a linear layer with SoftMax activation.

MaxPool and MaxAtt models by Maini et al. (2020) are taken into consideration as two baselines. Both use the same unique tokens as the Entity Markers-Entity Start architecture to encapsulate input entities. GloVe embeddings of the words in the input phrase are processed through a BiLSTM layer, then a MaxPool or MaxAtt model-specific aggregation mechanism, and finally a linear layer with SoftMax activation for relationship prediction.

These methods show how specialized structures may help literature on materials science to better classify relationships. Researchers can find the best techniques for gleaning associations from domain-specific texts by integrating entity identifiers and contrasting aggregation strategies.

4.3.3 Paper Abstract Classification

The output embedding of the CLS token from a BERT-based transformer model is used in the study to propose a text classification architecture. This technique encrypts the full text or abstract and uses the embedding to feed predictions to a classifier.

The BERT Sentence Classifier implementation of the transformers library is used as a starting point, which is comparable to relation classification but without input entities. By contrasting different methods, we may learn about their advantages and disadvantages, pinpoint best practices, and raise the precision and effectiveness of our models. Researchers may create more useful models for extracting and categorizing data in the field of materials science because to this comparison.

5 Results

Here we are presenting the results for all the three downstream tasks on different datasets. Multiple runs (seeds) are used to evaluate the consistency of a model's performance. Three seeds were used, each with a different seed. Cross-validation splits are used to assess the performance and generalizability of a model by splitting the data into multiple parts. In each iteration, four parts were used for training and the remaining part was used for validation. This helps to estimate the model's performance on unseen data more reliably

5.1 Results on Named Entity Recognition (NER)

Average Macro-F1 scores on SOFC-Slot and SOFC datasets test sets using 3 seeds and 5-fold cross-validation for reliability

	Architecture	**LM=MatSciERT**		**LM = SciBERT**		**LM = BERT**		**SOTA**
SOFC-Slot dataset		**Test**	**Validation**	**Test**	**Validation**	**Test**	**Validation**	**Test = 62.6** **Validation = 67.8 ± 12.9**
	LM-Linear	**65.82±1.53**	**68.53±3.48**	**57.64±2.49**	**60.58±4.68**	**59.29±1.96**	**63.75±4.24**	
	LM-CRF	**64.42 ±1.78**	**69.45±2.54**	**56.24±3.46**	**69.45±2.88**	**58.26±1.73**	**67.56±2.64**	
	LM-BiLSTM-CRF	**64.86±2.78**	**67.81±4.96**	**60.59±2.45**	**66.49±2.95**	**56.84±2.07**	**66.36±2.48**	
SOFC Dataset	**LM-Linear**	**81.45±2.01**	**80.37±2.00**	**77.05±2.54**	**82.04±1.98**	**77.08±1.75**	**77.11±2.96**	**Test = 81.5** **Validation = 81.7± 4.2**
	LM-CRF	**82.39±1.23**	**81.52±3.09**	**79.19±1.47**	**80.78±1.36**	**79.08±2.52**	**82.26±2.07**	
	LM-BiLSTM-CRF	**82.84±1.92**	**81.81±3.67**	**78.65±2.00**	**81.02±2.29**	**78.15±0.55**	**79.24±1.51**	

Test set Macro-F1 scores for Matscholar averaged across three seeds to estimate model performance reliably

Matscholar NER dataset	Architecture	LM=MatSciERT		LM = SciBERT		LM = BERT		SOTA
		Test	Validation	Test	Validation	Test	Validation	Test = 85.10 Validation = 85.41
	LM-Linear	83.19±1.18	88.57±3.41	83.85±5.31	86.55±5.55	81.15±5.81	81.75±5.15	
	LM-CRF	87.47 ±1.78	89.56±1.80	85.5 ± 5.77	88.57±5.56	81.57±5.15	81.61±5.81	
	LM-BiLSTM-CRF	86.09 ±0.46	89.15±0.57	85.66±5.11	87.66±5.15	83.35±5.15	81.57±5.15	

5.2 Results on Relation Classification (RC)

The results on the test set of the **Materials Synthesis Procedures dataset** were averaged over three different seeds to obtain a more reliable estimate of the performance

	MatSciERT		**SciBERT**		**BERT**		**MaxPool**		**MaxAtt**	
	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
Macro-F1	89.1±1.77	88.33±1.34	87.77±1.38	87.73±1.37	83.41± 3.43	83.93±1.78	83.39 ± 3.34	81.93±1.73	81.39 ± 1.83	83.33± 7.73
Micro-F1	93.94±1.71	93.31±1.71	93.14±1.37	93.13±1.18	91.36 ±1.69	91.44±1.34	86.83 ± l.33	86.68±1.84	87.36 ± 1.61	87.67± 3.34

5.3 Results on Paper Abstract Classification

The test set results for the glass vs. non-glass dataset were averaged over three seeds for improved reliability

	MatSciERT		SciBERT		BERT		MaxPool		MaxAtt	
	Test	Validation	Test	Validation	Test	Validation	Test	Validation	Test	Validation
Accuracy	96.77±5.36	95.33±5.77	93.44±5.57	94.55±5.5	93.89±5.6	93.33±5.98	93.44±5.33	97.77±5.56	93.44±5.68	93.77±5.36

6. Conclusion

In conclusion, MatSciBERT, a language model based on the GPT-3.5 architecture, presents a versatile tool for various materials science applications. By leveraging contextual embeddings, MatSciBERT surpasses traditional approaches, such as TF-IDF or Word2Vec, in document classification and topic modeling. Furthermore, its fine-tuned Named Entity Recognition (NER) model allows for accurate information extraction from image captions, which can be categorized using the Matscholar NER dataset. The insights gained from analyzing top entities in each category contribute to the advancement of materials science research, demonstrating the potential impact of MatSciBERT on this rapidly evolving field