Named Entity Recognition from Synthesis Procedural Text in Materials Science Domain with Attention-Based Approach

Huichen Yang, William Hsu
Computer Science of Kansas State University
Manhattan, Kansas 66502
huichen@ksu.edu, bhsu@ksu.edu

Abstract

We applied attention-based deep learning to the task of Named Entity Recognition (NER) from synthesis procedural text of scientific literature in materials science domain. Unlike conventional machine learning approaches that need hand-crafted features or training with massive data, our attention-based deep learning method enhances contextualized word representations by using a Bidirectional Encoder Representations from Transformers (BERT) pre-trained language model and then associating with Bidirectional Long Short-term Memory (BiLSTM) and Conditional Random Fields (CRF) layer; this is then BERT-BiLSTM-CRF. Our method shows it is feasible to use a limited annotated corpus with a pre-trained language model to extract entities from synthesis procedures in materials science. The experimental result shows that our approach outperforms other baseline models with significant improvements based on three corpora.

Introduction

The number of published materials science articles has grown rapidly over the past few decades. Much potentially useful information in these published articles could help the materials design group explore and study new material synthesis. Conventionally, new materials are discovered mainly through published experiments in literature, which, however, are usually stored as unstructured text format. This requires great effort to sort and organize. Furthermore, researchers and scientists in materials science cannot access much more than a fraction of such information because their research time is limited. The inevitable result is, therefore, the need to enhance their ability to identify new technologies and find the appropriate literature (Weston et al., 2019).

Natural Language Processing (NLP) with machine learning technology can accelerate the rate of materials science discoveries. Many materials science areas, thermoelectrics, photovoltaics, batteries, and pharmaceuticals, could use these techniques (Kalidindi and Marc, 2015). The fundamental task, then, of NER in NLP is to recognize named entities in the text of published experimental research and group them into pre-defined categories through classification (Nadeau and Satoshi, 2007). In this paper, we focus on NER in the synthesis of procedural text in materials science. The synthesis procedures are defined as the order of the steps based on "participating tagged entities and ultimately roles and operations" that should be in methods sections of materials science research literature. (Yang et al., 2019). Those tagged entities could be material names, operations, and devices, among others. They are essential to extracting procedural information from materials science literature. Figure 1 shows an example of named entities from synthesis procedure text in a materials science article.

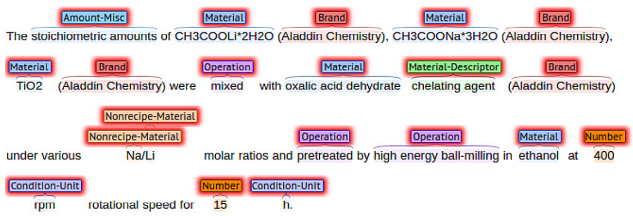

Figure 1: Example of named entities from synthesizing procedural text in materials science literature (Mysore et al., 2019). The highlighted words and phrases indicate entities involved in synthesis procedures.

In materials science, the particular challenge is insufficient annotated corpora; domain experts find labeling very expensive and time-consuming. To address the challenge in materials science, we used word embedding (Mikolov et al., 2013) with a BiLSTM (bidirectional LSTM) and a CRF (Conditional Random Fields) layer (Huang, Xu, and Yu, 2015) as our base line model. We used BERT (Devlin et al., 2019) pre-trained language model to compare contextual embedding to word embedding and then fit the output form BERT into a BiLSTM CRF model to learn the appropriate context information that would predict named entities. Our experiment results were based on three corpora of materials science and show the BERT-BiLSTM-CRF model improves significantly on other models.

Related Work

Named entity extraction from published experimental research is an emerging field, attracting attention from many

researchers. The most recently used approaches can be summarized into two types:

The first approach is entity extraction from materials science literature. This approach uses NER for extracting summary-level information from materials science documents. These named entities are broadly pre-defined in materials science as material name, sample descriptors, and material properties, among others (Weston et al., 2019). The common extraction method collects relevant literature, uses unsupervised learning methods like K-means and Word2Vec (Mikolov et al., 2013), extracts word representation features from large unlabeled corpora, and then fits these word representation vectors along with small annotated corpora to machine learning models like CRF, decision tree with a linear classifier, and hierarchical neural networks for named entity extraction (Munkhdalai et al., 2015; Kim et al., 2017; Kim et al., 2017). The extraction results can be stored in a database as structured data for queries.

The second approach is named entity extraction from synthesis procedural text of materials science literature. This approach uses NER to synthesize procedural text (or experimental methods) in the methodology sections of materials science publications. Compared with summary-level NER in materials science, this approach centered on details of entities involved in the experiment itself, including material names and operations in the experiment steps. Some previous research focused on this approach. Mysore et al. (2017) extracted procedural information with action graphs, and Huo et al. (2019) used semi-supervised learning methods with latent Dirichlet allocation (LDA), and random forests to classify inorganic materials from methodology information. We chose to use NER to synthesize procedural text as our main methodology.

Methodology

We treated NER as a sequence labeling problem. BERT-BiLSTM-CRF, the attention-based, deep learning, end-to-end model, was used to solve this problem. Figure 2 shows the structure of BERT-BiLSTM-CRF model. The pre-trained BERT model (Devlin et al., 2019), as the embedding layer, received the raw input sentences. Then the BERT model output the contextual embedding vectors for each word as input to the BiLSTM layer for syntactic and semantic feature representation learning. The final CRF layer output possible tag sequences based on their conditional probability.

BERT Pre-trained language model

BERT (Devlin et al., 2019) is a pre-trained language model based on a deep transformer encoder (Vaswani et al., 2017). It introduced a masked language model (MLM) and next sentence prediction (NSP) to optimize the training process. These mechanisms allowed BERT to use an attention-based, multi-layer, bidirectional transformer mechanism and a normal nonlinear layer to learn contextual information from large unlabeled corpora. Moreover, the pre-trained BERT language model can be easily fine-tuned for a particular downstream task. It is precise because BERT can use contextual information learnability and transferability instead of context-independent word embedding like

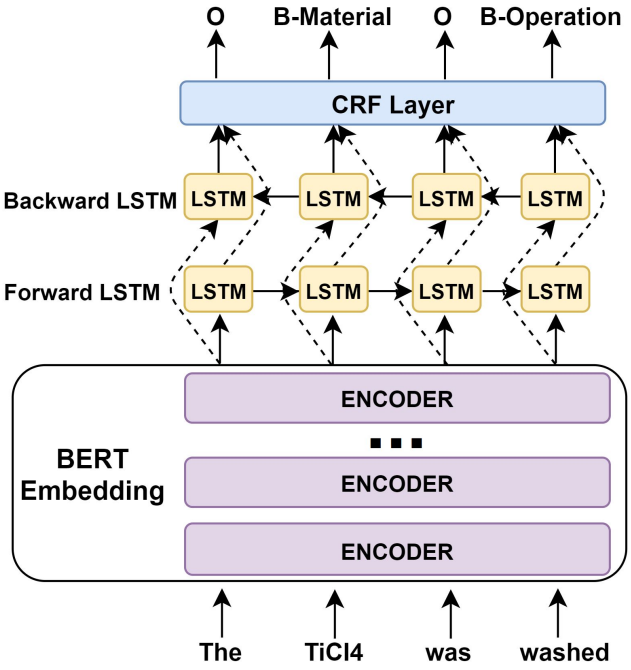

Figure 2: The architecture of the BERT-BiLSTM-CRF model.

Word2Vec (Mikolov et al., 2013). This meant we could use BERT as the embedding layer. After fine-tuning, BERT performed well, even though it was pre-trained with corpora irrelevant to materials science; our NER task demonstrated this ability in our experimental results.

Bidirectional LSTM layer

The bidirectional LSTM is an extension of LSTM that applies a forward and backward LSTM network to sequence processing and links the network to the output layer (Huang et al., 2015). The BiLSTM structure enables the output layer to gather contextual information simultaneously from past (backward) and future (forward). In addition, the BiLSTM has LSTM characteristics that avoid gradient vanishing and exploding that occur in RNN. Both forward and backward LSTM networks use the same equations in LSTM.

The BiLSTM takes the embedding result from BERT as an input vector for extracting sentence features. The output of the hidden state of BiLSTM will concatenate the forward LSTM H_f and backward LSTM H_b networks as final output $[H_l, H_r]$.

CRF layer

CRF is discriminative probabilistic method subject to a certain correlation constraint among tags. Using CRF as the last layer can help models learn the joint relationship between tags, as well as learn the constraints that ensure the sequences are valid. For instance, in BOI tagging format, the label of the first word in a sentence should start with the tag of "B" or "O", but not "I". These constraints are learned automatically using the training dataset created by the CRF layer during the training process.

Label prediction of the CRF layer combines the output P

from the BiLSTM layer, which represents the score of the i_{th} word in the sentence where y_i is the tag of the i_{th} word, and the transition matrix T represents the transition probability from tag y_i to tag y_{i+1}. We used the following equation to calculate the score of the labels sequence:

$$Score(X, Y) = \sum_{i=1}^{n} P_{i,\, y_i} + \sum_{i=0}^{n} T_{y,\, y_{i+1}} \quad (1)$$

Our goal is to minimize the loss function by maximizing the total score of the probability of sequence $s_{(X,\, y_i)}$. The log loss function is given as follows:

$$\mathcal{L} = Score(X, Y) - log \sum_{y_i \in Y} e^{S(X,\, y_i)} \quad (2)$$

Experiment and Results

In this section, we describe the experiment and the results from three corpora.

Corpora

We used three corpora grouped into two categories to evaluate the model.

- Corpus 1 is a materials synthesis procedural (**MSP**) annotated corpus that was published in 2019 (Mysore et al., 2019). This corpus is annotated by domain-experts from 230 experiment paragraphs describing synthesis procedures in materials science domain.
  - ***MSP***: Contains the operations and their arguments in synthesis experiments, such as material name, operation descriptor, synthesis apparatus, which have 21 different named entities.
- Corpus 2 is an annotated corpus in solid oxide fuel cells (**SOFC**) that is a sub-area of materials science published in 2020 (Friedrich et al., 2020). This corpus is annotated using four annotation schemes based on 45 open-access scholarly articles by domain-experts. We use two of the four corpora, both related to the NRE task; the other two corpora are not related to the NRE task:
  - ***SOFC***: Major entity mention types in experiment-describing sentences that include three different named entities.
  - ***SOFC_Slot***: Experiment slot types in experiment-describing sentences that include 16 different named entities.

All of the corpora are annotated using the BOI format, where B is the word beginning entity, I is words inside the entity, and O is outside of the entity. The BOI labels should be predicted by the NER model; they were then transformed to pre-defined named entities.

Implementation details

We chose two different embedding layers for comparison. The Word2Vec was used as the word embedding layer for the BiLSTM-CRF model. For the BERT-CRF and BERT-BiLSTM-CRF models, we considered BERT as the embedding layer. We used a BERT-based-cased language model, which was pre-trained on cased English text. We chose SciBERT, a BERT model trained on scientific text (Beltagy et al., 2019), for comparison. Both pre-trained models have 12 attention heads, 12 layers and 768 hidden dimensions. We set maximum sequence length at 512, batch size at 16, initial learning rate at 0.05, warm up proportion rate at 0.1, and the dropout rate at 0.2. We used 10 epochs in the BERT-related fine-tuning models: BERT, SciBERT, BERT-CRF, and SciBERT-CRF. We used 100 epochs for training in the BiLSTM related models. In addition, the BERT language models were tuned as BERT embedding during the training process for BiLSTM-related models.

Corpora	Model	Precision	Recall	F1
MSP	BiLSTM-CRF (Word2Vec)	78.51	74.84	76.63
	BERT	78.94	80.76	79.84
	BERT-CRF	79.75	80.60	80.60
	BERT-BiLSTM-CRF	85.25	83.53	84.38
	SciBERT	79.25	82.84	81.01
	SciBERT-CRF	80.48	82.96	81.70
	SciBERT-BiLSTM-CRF	**86.38**	**85.15**	**85.62**
SOFC	BiLSTM-CRF (Word2Vec)	75.33	74.35	74.84
	BERT	93.01	88.46	90.67
	BERT-CRF	93.32	88.54	91.10
	BERT-BiLSTM-CRF	93.38	90.09	91.43
	SciBERT	93.98	88.77	91.30
	SciBERT-CRF	**94.11**	89.28	**91.62**
	SciBERT-BiLSTM-CRF	93.14	**91.17**	91.57
SOFC_Slot	BiLSTM-CRF (Word2Vec)	63.24	56.29	59.56
	BERT	78.41	71.85	74.99
	BERT-CRF	80.00	72.49	76.06
	BERT-BiLSTM-CRF	89.31	82.08	86.16
	SciBERT	77.35	71.80	74.47
	SciBERT-CRF	78.45	70.46	74.24
	SciBERT-BiLSTM-CRF	**90.31**	**84.25**	**87.17**

Table 1: Evaluation results for three different corpora.

Evaluation methods

We used micro precision, recall, and F1 to evaluate the models because the corpora have a potential class imbalance issue. For example, the sample tagged as Material, Operation, Number, and Amount-Unit dominate the ***MSP*** corpus and reflect most synthesis procedures, but some named entities are not as important. However, macro precision, recall, and F1 treat all classes equally, which could have affected the accuracy of extraction results. The corresponding equations are presented below:

$$micro - Precision = \sum_{i=1}^{N} \frac{set_{\text{pre}} \cap set_{\text{true}}}{set_{\text{pre}} \cap set_{\text{true}} + set_{\text{pre}} \setminus set_{\text{true}}} \quad (3)$$

$$micro - Recall = \sum_{i=1}^{N} \frac{set_{\text{pre}} \cap set_{\text{true}}}{set_{\text{pre}} \cap set_{\text{true}} + set_{\text{true}} \setminus set_{\text{pre}}} \quad (4)$$

$$microF1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (5)$$

In these equations, the set_{pre} represents the prediction set, and the set_{true} represents the true labels set.

Results and analysis

We ran three different corpora using the same models. We used word embedding with BiLSTM-CRF as the baseline

model and connected BERT embedding layer with CRF or BiLSTM-CRF. The results showed that the BERT-BiLSTM-CRF model achieved the best performance in most cases. Table 1 shows the results of our evaluation.

From the results, the pre-trained BERT language model used as embedding layer instead of Word2Vec showed significant improvement over the baseline model. That means the contextual feature of sentence was very helpful in the NER task in synthesis procedural text of materials science literature. In addition, the pre-trained BERT model worked better in the scientific text than in general English text. The results also showed that a fine-tuned, pre-trained language model with small corpora in a domain specific NER task got decent results in general. In addition, the corpus of ***SOFC*** had the best performance because it had only three different named entities with more balanced numbers.

To the best of our knowledge, the ***MSP*** (Mysore et al., 2019) corpus has not been evaluated in any other publication. We compared our results with the evaluations in Friedrich et al. (2020) based on ***SOFC*** and ***SOFC_Slot*** corpora. Table 2 provides a comparison of evaluation results.

Corpora	Model	macro F1
	SciBERT (Friedrich et al,2020)	81.50
SOFC	SciBERT-BiLSTM-CRF (**ours**)	**85.61**
	BiLSTM SciBERT (Friedrich et al,2020)	62.60
SOFC_Slot	SciBERT-BiLSTM-CRF (**ours**)	**64.59**

Table 2: Comparison of evaluation results with SOFC corpora.

Table 2 shows that our SciBERT-BiLSTM-CRF model outperforms both ***SOFC*** and ***SOFC_Slot*** corpora. Please note we chose the macro F1 in our evaluations to remain consistent with Friedrich et al.'s (2020) evaluation methodology.

Conclusion

In this paper, we introduce a promising attention-based deep learning approach, BERT-BiLSTM-CRF, for the NER task for synthesis procedural text of materials science. We evaluated our approach using three synthesis procedural text relevant corpora. The results showed that our BERT-BiLSTM-CRF model improved significantly over the baseline model. We have presented several models that got better results with the pre-trained language model BERT as the embedding layer compared than with word embedding models like Word2Vec. We also compared our model (using the SOFC corpora) to Friedrich et al.'s (2020) model (using the SOFC-Slot corpora). Our model was the better one based on the comparison results. Our work contributes to the community of materials science by demonstrating success in applying an attention-based, deep learning approach to NER of synthesis procedural text. Moreover, our work provides a competitive benchmark with these three corpora.

A few challenges in using NER in materials science will be further investigated in future work. For example, material name acronyms or abbreviations are a source of ambiguity; named entity detection of mention boundaries is also worth attention. The other concern is the entity label imbalance. For instance, there are 4843 named entities of materials in the ***MSP*** corpus, but only 122 named entities of Condition-Type. Future work should improve the application of our model in materials science domain.

References

Devlin, J., Ming-Wei Chang, Kenton Lee and Kristina Toutanova."BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding" *NAACL-HLT* (2019).

Friedrich, Annemarie, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruscyk, and Lukas Lange. "The SOFC-Exp Corpus and Neural Approaches to Information Extraction in the Materials Science Domain." *arXiv preprint arXiv:2006.03039* (2020).

Huang, Zhiheng, Wei Xu, and Kai Yu."Bidirectional LSTM-CRF models for sequence tagging." *arXiv preprint arXiv:1508.01991* (2015).

Huo, Haoyan, Ziqin Rong, O. Kononova, W. Sun, T. Botari, Tanjin He, V. Tshitoyan and G. Ceder."Semi-supervised machine-learning classification of materials synthesis procedures." *npj Computational Materials* 5 (2019): 1-7.

Iz Beltagy, Kyle Lo, and Arman Cohan. "SciBERT: A pretrained language model for scientific text." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620 (2019)

Kalidindi, Surya R., and Marc De Graef."Materials data science: current status and future outlook" *Annual Review of Materials Research* 45 (2015): 171-193.

Kim, Edward, Kevin Huang, Adam Saunders, Andrew McCallum, Gerbrand Ceder, and Elsa Olivetti. "Materials synthesis insights from scientific literature via text extraction and machine learning" *Chemistry of Materials* 29, no. 21 (2017): 9436-9444.

Kim, E., Kevin Huang, A. Tomala, Sara Matthews, Emma Strubell, A. Saunders, A. McCallum and Elsa Olivetti."Machine-learned and codified synthesis parameters of oxide materials." *Scientific Data* 4 (2017): n. pag.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean."Distributed representations of words and phrases and their compositionality" In *Advances in neural information processing systems* 26 (2013): 3111-3119.

Munkhdalai, Tsendsuren, Meijing Li, Khuyagbaatar Batsuren, Hyeon Ah Park, Nak Hyeon Choi, and Keun Ho Ryu."Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations" *Journal of cheminformatics* 7, no. S1 (2015): S9.

Mysore, Sheshera, E. Kim, Emma Strubell, A. Liu, Haw-Shiuan Chang, Srikrishna Kompella, Kevin Huang, A. McCallum and Elsa Olivetti. "Automatically Extracting Action Graphs from Materials Science Synthesis Procedures." *ArXiv* abs/1711.06872 (2017).

Mysore, Sheshera, Zach Jensen, Edward Kim, Kevin Huang,

Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti.“The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures” *arXiv preprint arXiv:1905.06939* (2019).

Nadeau, David, and Satoshi Sekine.“A survey of named entity recognition and classification” *ingvisticae Investigationes 30*, no. 1 (2007): 3-26.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin.“Attention is all you need” In *Advances in neural information processing systems*, pp. 5998-6008. 2017.

Weston, Leigh, Vahe Tshitoyan, John Dagdelen, Olga Kononova, Amalie Trewartha, Kristin A. Persson, Gerbrand Ceder, and Anubhav Jain. “Named entity recognition and normalization applied to large-scale information extraction from the materials science literature” *Journal of chemical information and modeling* 59, no. 9 (2019): 3692-3702.

Yang, Huichen, Carlos A. Aguirre, F. Maria, Derek Christensen, Luis Bobadilla, Emily Davich, Jordan Roth et al.“Pipelines for Procedural Information Extraction from Scientific Literature: Towards Recipes using Machine Learning and Data Science” In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 2, pp. 41-46. IEEE, 2019.