A Lithium-Ion Battery Cathode Material Literature Entity Recognition Method Based on Deep Learning

Ziyu Yin[1], Gang Lei[1], Jianmao Xiao[1,4](✉), Xinji Qiu[2], Qian Zhang[3], Lei Chang[2], Haohui Chen[1], Musheng Wu[2], and Xushan Zhao[3]

[1] School of Software, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
jm_xiao@jxnu.edu.cn
[2] Department of Physics, Jiangxi Normal University, Nanchang, Jiangxi 330022, China
[3] Innovation Laboratory, Contemporary Amperex Technology Co., Limited, Ningde, Fujian 352100, China
[4] Jiangxi Provincial Engineering Research Center of Blockchain Data Security and Governance, Nanchang, Jiangxi 330022, China

Abstract. With the continuous deepening of research on lithium-ion battery materials, the number of literature in related fields has shown explosive growth, posing a challenge to researchers in terms of literature reading and information screening. Named entity recognition technology can automatically mark the pre-defined entity content in the literature, helping researchers to quickly locate the key information of the articles, and improve the reading rate of the literature. In this paper, the classification system of lithium-ion battery materials literature is constructed through the combination of automation and manual analysis, and the named entity recognition operation is conducted. We adopt the named entity recognition model based on BiLSTM-CRF, combined with Bert word embedding technology, to automatically identify and extract entities in the literature, and develope an automatic document annotation tool. Our work provides strong support for researchers to quickly understand the literature content and key information, and lays a foundation for subsequent relationship extraction and the construction of a Knowledge graph.

Keywords: Lithium-ion Battery Materials · Information Extraction · Classification System · Named Entity Identification

1 Introduction

Lithium-ion batteries, as one of the key research directions in the field of battery materials, are receiving increasing attention from researchers. To conduct in-depth research and exploration of lithium-ion batteries, it is essential to read various literature on battery materials. Reading a large amount of literature often

Y. Cao and X. Shao (Eds.): DIONE 2023, LNICST 515, pp. 105–119, 2025.
https://doi.org/10.1007/978-3-031-80713-8_8

consumes a lot of time, and it is difficult to grasp the key points of the articles. However, reading annotated literature can help researchers quickly locate the key knowledge points within the text, thereby improving the efficiency of reading the literature. Named Entity Recognition (NER) is a natural language processing technique used to automatically identify specific types of entities, such as names, locations, organizations, and more, from text. In the field of materials science, NER is widely employed to extract and annotate entity information related to battery materials in literature, thereby assisting researchers in reading and analyzing a vast amount of literature more efficiently. Additionally, it provides robust support for subsequent data mining and knowledge management tasks.

NER plays a crucial role in materials data mining. The field of materials science involves a vast amount of literature, and NER technology can accurately identify key entities within the text, such as material and properties. This assists researchers in quickly capturing the essential content of articles without having to go through the entire text, saving a significant amount of reading time. In practical applications, researchers can use existing NER tools or develop their own NER models to perform entity recognition in the battery materials domain. These recognized entities can be used to construct databases, search engines, and support research tasks in other domains as well. For example, Hassna El-Bousiydy et al. [1] used NER technology to construct the Lithium-Ion Battery Annotated Corpus (LIBAC) for the experimental and results sections of journal corpora. This corpus contains 28 relevant entities in the field of lithium-ion battery materials, which can assist researchers in training automated systems for information extraction tasks. Leigh Weston et al. [2] applied NER text mining and entity normalization to perform large-scale information extraction from published abstracts in the field of inorganic materials science, creating a document search engine where users can query literature based on entity content. Additionally, combining active learning with pre-trained language models can enhance the performance of NER, enabling more accurate identification of material entities. For example, Niu Chen et al. [3] utilized active learning and pre-trained language models to achieve entity recognition for English-language corpora of alloy materials.

This study focuses on the construction of a knowledge classification system and entity recognition for literature on lithium-ion battery materials. The aim is to utilize named entity recognition technology to automatically identify entities and their categories in articles, assisting researchers in quickly grasping the key points of the literature and accelerating the rate of extracting valuable information from a large volume of literature. The main contributions of this paper are as follows:

(1) The combination of automation and manual analysis is adopted to determine the knowledge classification system of lithium-ion battery materials, which improves the efficiency of traditional manual analysis. Further manual adjustment of automatic classification results can improve the accuracy of classification and reduce the influence of subjective factors.

(2) Combined with Bert word embedding, we constructed the BiLSTM-CRF entity recognition model, and our entity recognition performance achieved an F1 score of approximately 83%.
(3) A literature annotation tool based on lithium-ion battery materials is designed to automatically identify and highlight various kinds of entities in the input literature, to help researchers quickly capture research priorities and improve the efficiency of literature reading.

2 Related Work

The combination of artificial intelligence and the materials field has made it possible to extract information from massive literature using machine learning techniques. Currently, most information mining in the materials domain is based on traditional relational databases, which involve statistical analysis of the structure, properties, and functionalities of different materials. For example, Shu Huang et al. [4] modified the Natural language processing tool ChemDataExtractor [5] for the battery field. They automatically extracted textual information from over 220000 academic papers on battery materials, establishing a large-scale database on battery materials and their related properties. They focused on studying functional characteristics such as capacity, conductivity, and coulombic efficiency of battery materials. Liu Yue et al. [6] proposed a data augmentation model that incorporates knowledge from the materials field to construct a high-quality dataset for material science text mining. This approach effectively enhances the performance of downstream entity recognition models. However, none of them have addressed the mining of relationships between different entities, and they cannot uncover the potential associations between materials. Shan Bin et al. [7] summarized the general process and tools of material text mining, discussed the application of material data in material performance prediction, and proposed the potential and prospects of material information extraction using a large language model. Nie et al. [8,9] used knowledge graph technology to conduct deeper information mining on the knowledge correlation of cathode materials for lithium-ion batteries, and found new materials that can be used as cathodes. But this graph was only based on the similarity between entities of materials, and it didn't encompass the storage of relational information such as structural properties, preparation methods, and applications of each material. Therefore, the fusion of machine learning with the field of materials can accelerate the pace of materials information mining. NER technology, as the most fundamental step in information extraction, can effectively facilitate the in-depth analysis and mining of subsequent literature information, providing guidance and inspiration for the advancement of materials research. However, current materials information mining methods are largely confined to storing data at the entity level or solely exploring potential new materials based on entity similarity. It is believed that future researchers, leveraging the rich attribute relationships between entities, can utilize more efficient machine learning algorithms, even in combination with rapidly evolving large language models, to unearth more potential new materials and effectively predict various material properties, thereby advancing the field of materials science.

3 Data Collection and Preprocessing

The abstract of an article serves as a quick pathway for readers to access the content of the paper. It typically provides a concise overview of the main objectives, methods, and results of the article, allowing readers to quickly grasp the core content and contributions of the paper. In this study, we focus on using article abstracts as the primary dataset for Named Entity Recognition (NER) operations. This helps researchers automatically annotate key term knowledge related to the literature, thus improving the efficiency of literature reading. For this study, we retrieved literature from databases such as CNKI and Web of Science. We collaborated with domain experts to determine the keywords and search rules for retrieving literature related to lithium-ion battery materials. Our goal was to include as many relevant articles from the materials field as possible while considering the depth of the literature [10]. We batch-exported article titles, abstracts, keywords, and other relevant content, storing them in XLS format for further analysis. In total, we exported 40227 articles for analysis.

The exported literature data contains a lot of irrelevant or repetitive information, which needs to be cleaned. Firstly, duplicate documents are removed based on the document titles. Secondly, data with empty abstracts, keywords, or titles are excluded. Then, irrelevant tags and dirty data in the text are removed using regular expressions. After data cleaning, we retained a total of 30879 article abstracts. To further screen out the most relevant literature in this field and improve the effectiveness of subsequent experiments, a classifier is trained to select the most useful literature on lithium-ion battery materials. We manually annotate 500 literature abstracts and divide them into training, validation, and testing sets in an 8:1:1 ratio. To achieve the best literature classification effect, we compare the effects of four different binary classification models, as shown in Table 1. Ultimately, the MatBertCNN model is chosen for binary classification of all the collected literature, resulting in a total of 13899 retained articles.

Table 1. Classification effects of the four classifiers

Model	Accuracy	Recall	F1-score
MatBert	0.796	0.796	0.796
MatBertCNN	**0.821**	**0.837**	**0.827**
MatBertRNN	0.788	0.714	0.742
MatBertRCNN	0.774	0.796	0.784

4 Knowledge System Construction

4.1 Extract Candidate Terms

In this study, the abstracts, keywords, and titles are tokenized using the NLTK tokenizer, with whitespace as the delimiter. To prevent redundancy caused by capitalization, all words are converted to lowercase. The tokenized results are then part-of-speech tagged, retaining verbs and nouns which are more meaningful for clustering, while removing words of other parts of speech. Additionally, words underwent lemmatization to normalize their forms.

In the materials domain, there are cases of chemical formula separation errors in the tokenization results. The tokenized results may contain words connected by symbols like "/", """, ".", as well as words starting or ending with special characters, such as "polymer/graphite", "-lithium", and "cycle_". so we split the connected words and remove extra characters before or after the words, then add them to the tokenization list. As materials literature includes various types of numerical content, all words containing numbers are initially removed to determine the main entity categories through clustering, and the numerical content is analyzed and categorized manually in subsequent steps.

We apply the NLTK built-in stopword list, supplemented with 995 stopwords collected manually, to remove stopwords. The processed tokenized results are included in the candidate term set, and we perform word frequency statistics, resulting in 22,086 unique words. However, considering word frequency alone for selecting candidate terms only takes into account quantity characteristics. In some specialized fields, important terms may occur less frequently compared to common words. Therefore, we filter the tokenized results by combining the TF-IDF values of words. Words with TF-IDF weights below the threshold of 7.696E-06 and words with a length less than 4 are removed. Ultimately, we retain 10,385 words for clustering. The top words based on TF-IDF values are shown in Table 2.

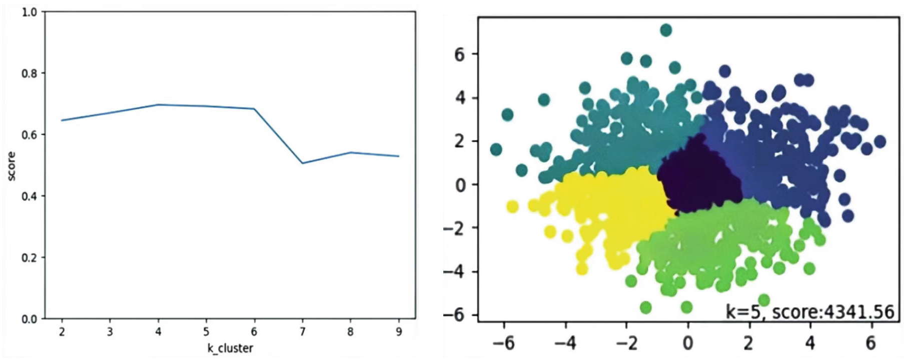
Table 2. Example of candidate terms sorted by TF-IDF values

words	weight	words	weight	words	weight
capacity	0.015594	discharge	0.012292	method	0.010084
carbon	0.014491	surface	0.012066	result	0.009964
cycle	0.013567	property	0.011767	sulfur	0.009964
electrode	0.013108	lithium-ion	0.011701	cycling	0.009793
electrolyte	0.012538	stability	0.011402	performance	0.009731
cell	0.012392	structure	0.010817	storage	0.009683
anode	0.012324	energy	0.010414	temperature	0.009641

4.2 Candidate Term Clustering

After the processing in the previous section, we obtained candidate terms in the field of lithium-ion battery materials. To better determine the research entities involved in the materials domain, we perform clustering on the candidate terms and discuss with domain experts to comprehensively establish the knowledge classification system.

In this study, we use the CBOW model in Word2Vec to convert the tokenized results into 2-dimensional feature vectors. These vectors are then inputted into the clustering model for clustering operations. We utilize the K-means algorithm for cluster analysis of the candidate terms. The K-means algorithm requires the number of desired clusters to be specified. If the initial number of clusters is not appropriate, it can affect the final clustering results. To select the most suitable number of clusters, we compare the clustering effects for different values of k and evaluate them using the contour coefficient. The contour coefficients calculated in this study are shown in Fig. 1(a). From the figure, we can see that the clustering effect is optimal when k equals 4 or 5. Combining this with manual analysis and discussions, we ultimately select k = 5 for clustering. The visualization results are shown in Fig. 1(b).



(a) Contour coefficients of k-means **(b)** Cluster effect (k=5)

Fig. 1. k-means result

4.3 Classification System Determination

Based on the material entity recognition classification system referenced in [2], we analyze the final clustering results from the previous section to construct a classification system for literature on lithium-ion battery materials. In the clustering results, Category 1 contains 419 terms, it includes words such as "battery", "electrolyte" and "libs", which describe the application of the material. Consider defining this category as "Application". Category 2 contains 286 terms, it includes words such as "structure", "microscopy" and "crystallinity", which describe the nature and internal structure of the material. Consider defining this category as "Description". Category 3 contains 632 terms, its contents include words such as "technique", "diffraction" and "spectroscopy", which describe the preparation method or characterization method of material. Consider defining this category as "Method". Category 4 contains 8815 terms, including indicator words such as "capacity", "size" and "reversibility", which are used to evaluate material properties. Consider defining this category as "Attribution". Category 5 contains 233 terms, including words describing the external morphology of materials such as "nanoparticles" and "nanosheets," as well as composition elements of materials like "carbon" and "sulfur." Overall, these terms provide introductions to the materials discussed in the papers. Therefore, we categorize this as "Materials". Secondly, considering that words with numerical values were removed during data preprocessing, we particularly analyze the numerical content in the article. The numerical values in the paper generally describe measurement data of material property indicators and the corresponding experimental conditions. Therefore, two additional entities, "Condition" and "Data", are added to the existing 5 clusters. As a result, we establish 7 entity categories: "Materials", "Description", "Method", "Attribution", "Application", "Condition" and "Data". The definition of each entity category is presented in Table 3.

Table 3. Classification effects of the four classifiers

entity	definition	examples
materials (MAT)	Chemical formula and materials related to the study substance	lithium, LiMn2O4, NMC
description (DSC)	Characteristics and the internal structure of the material	spinel, layer, Mg-dopped
method (MET)	The synthesis and characterization methods of the materials	carbonization, DFT
attribution (ATTR)	Evaluation indicators and attributes with measurable values	length, good capacity, energy density
application (APL)	Application scenarios of the materials	cathode, solution, catalytic
condition (CON)	The prevailing conditions of the experiment	1 C, 300 cycles, 0.1 A g(-1)
data (DATA)	Values of the measured experimental results	167.2 mAh g(-1), 80%, 3 nm

5 Named Entity Recognition

5.1 Corpus Building

As a Supervised learning task, Named Entity Recognition needs to take the labeled data as the input of the model, and train the Supervised learning model to learn the mapping relationship from the input text to the label sequence. The annotated corpus used for model input is the foundation for training model parameters, and its quality directly affects the accuracy of the training model. Therefore, the quality and scale of the corpus are crucial.

Due to the unique nature of literature in the professional field, there are few corpora for literature on lithium-ion battery materials, and they can not fully match our defined classification system. Therefore, we choose to manually construct a knowledge corpus of lithium-ion battery materials, specifically by annotating the abstracts of the literature. In this study, the BIO labeling method is selected, and with the help of the Yedda annotation tool [11], which has a simple and easy-to-understand visual interface, supports customized entity labels and quantity, and can export the annotation results to BIO format with one click. The various entity category labels marked in this article are shown in Table 4.

Table 4. Various entity labels of lithium-ion battery materials

entity category	entity start label	entity intermediate or end label
materials	B-MAT	I-MAT
description	B-DSC	I-DSC
method	B-MET	I-MET
attribution	B-ATTR	I-ATTR
application	B-APL	I-APL
condition	B-CON	I-CON
data	B-DATA	I-DATA
other	O	O

The annotation guidelines for each entity category are as follows: Both the full name and abbreviations of entities should be annotated. Punctuation marks that are unrelated to the entity should not be included in the annotation. Pronouns should not be annotated. The content of MAT category labeling is specific to a certain material, and ions and individual elements should not be annotated. For the ATTR category, the modifying words preceding the attribute should be included in the annotation. For example, words like "good" or "high" should be included when annotating. For the DSC category, words like "structure" or "phase" that exist alone should not be annotated. For the DSC category, words like "structure" or "phase" that exist alone should not be annotated. Units should be included when annotating entities in the DATA category.

We selected 520 abstracts that contain as many entities as possible and are most relevant to lithium-ion battery materials from the collected literature abstracts for annotation, and divided them into training, validation, and testing sets in an 8:1:1 ratio.

5.2 Entity Recognition Model Based on Bert-BiLSTM-CRF

The Bert+BiLSTM+CRF model is a sequence labeling model that combines multiple deep learning components commonly used in natural language processing tasks, particularly named entity recognition and part-of-speech tagging. Bert, a pre-trained deep bidirectional Transformer model, is employed for natural language understanding tasks. It possesses the ability to understand deep bidirectional context, enabling it to better capture the semantic information of words within context. BiLSTM, a bidirectional recurrent neural network (RNN) structure, effectively captures contextual features of words by passing information both forward and backward at each time step. CRF, a probabilistic graphical model typically used in sequence labeling tasks, considers dependencies between labels, assisting the model in better predicting the label sequence for the entire sequence, rather than independently labeling each word. This ensures consistency and coherence in labeling. This model, which combines Bert's contextual understanding, BiLSTM's sequence modeling, and CRF's label dependency modeling, excels in sequence labeling tasks. It has a better grasp of the semantic information in the text while considering relationships between labels, thereby enhancing the performance of tasks like named entity recognition. Therefore, in this paper, we construct a named entity recognition model for the materials domain based on the Bert+BiLSTM+CRF method. The model primarily consists of five components: the input layer, word embedding layer, BiLSTM layer, CRF layer, and model output, as illustrated in Fig. 2.

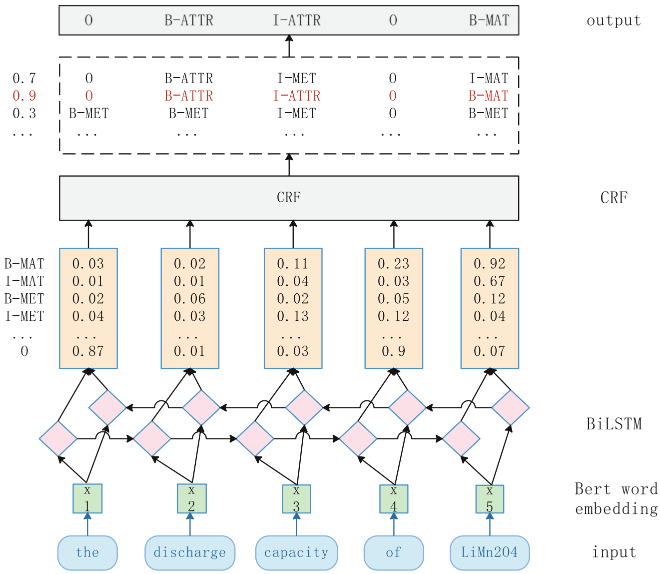


Fig. 2. Entity recognition model architecture

The input layer takes the annotated BIO corpus as the input data for model training and passes it to the word embedding layer. The word embedding layer assigns a low-dimensional vector representation to each word and converts the text data into continuous vectors to capture the semantic and grammatical relationships between words. This study uses the Bert pre-training model [12] for word embedding operation, and inputs the vectorized words to the bidirectional LSTM layer for context feature learning. The output of the BiLSTM layer is the score of each word vector on each label, and the label corresponding to the highest score is the best prediction label. However, further analysis reveals that the predicted label's output by the BiLSTM layer may have entity headers starting with I labels, or the predicted labels of the same entity belong to different categories, which may result in some invalid label predictions. To address this, the output is passed to the CRF layer for further processing. The CRF layer can model the joint probability distribution of the entire tag sequence, thus considering the dependence between labels, which avoids the case of invalid labels by training the transfer score matrix. Finally, the optimal label sequence is selected based on CRF loss computation and used for output.

5.3 Experiments and Results

This research builds the Bert word embedding layer, BiLSTM layer and CRF layer of the model based on the Pytorch framework and Python 3.9, and conducts training and testing of the entity recognition model oriented to the literature in the field of lithium-ion battery materials. The parameter settings of the model are shown in Table 5:

Table 5. Bert-BiLSTM-CRF model parameters

parameter	definition	values
batch_size	batch size	10
max_len	the maximum length of the processed sequence data	50
lstm_hidden	dimensions of the hidden state of the LSTM	768
lr	learning rate	1e-4
epoch	iterations	30
weight_decay	regularized coefficient of the weight decay	0.01

Precision, Recall, and F1-score are used as the evaluation indicators of the model results. During training, we save the model with a higher F1-score at each iteration. After 30 epochs, the loss value of the model decreased from 1329 to around 1.7, and the F1-score gradually converges to around 82%. The best-performing model is evaluated on the test set, and the evaluation metrics for each entity are shown in Table 6. We counted the evaluation metrics for each entity label and calculated their average effects using the macro and weighted methods.

The macro-average gives equal weight to each class, focusing on the performance of each class individually, which provides an objective assessment of the overall performance of the multi-class classification problem. The weighted average assigns weights to each class based on the class distribution and computes the weighted average metrics. It can better reflect the impact of class imbalance on the evaluation results, giving more attention and weight to classes with fewer samples and providing more accurate evaluation results.

Table 6. Evaluation of entity identification results

entity label	Precision	Recall	F1-score	Number of entity labels
B-MAT	0.77	0.78	0.77	134
B-MET	0.79	0.89	0.84	75
B-APL	0.62	0.78	0.69	45
B-ATTR	0.72	0.80	0.76	71
B-DSC	0.67	0.78	0.72	46
B-CON	0.88	0.83	0.86	36
B-DATA	0.82	0.90	0.86	31
I-MAT	0.77	0.67	0.71	30
I-MET	0.82	0.94	0.87	63
I-APL	0.46	0.70	0.55	23
I-ATTR	0.72	0.95	0.82	75
I-DSC	0.54	0.47	0.50	15
I-CON	0.84	0.89	0.87	55
I-DATA	0.89	0.94	0.92	53
O	0.97	0.93	0.95	2223
macro avg	0.75	0.82	0.78	2975
weighted avg	0.92	0.91	0.91	2975

The analysis of the experimental results shows that the model has the best identification effect on numerical entities such as CON and DATA. Both the beginning and the middle part of the entity can be accurately identified, and the F1 value can reach more than 85%, indicating that the numerical features are easy to learn to identify. Secondly, the recognition performance of MET and ATTR entities follows closely, as the descriptions of methods and attributions in the text are mostly limited and repetitive, making it easier for the model to recognize the features of such entities. We find that the number of B-MAT tags is the highest except for O tags, while the number of I-MAT tags is only 30, indicating that most research materials are based on a single word as an entity without an intermediate part. Moreover, due to the diversity of battery materials in the professional field, the recognition performance of the model for such entities still needs to be improved. Finally are the DSC and APL entities.

Our analysis find that their recognition performance is the lowest, but it can also reach around 70%, which is within an acceptable range. This may be due to the limited occurrence of these entities in the text, resulting in insufficient training for the model to recognize them effectively.

Considering the overall evaluation of all labels, the weighted average F1-score reaches around 91%, taking into account the label imbalance, such as the excessive occurrence of the O label. This indicates that our trained model performs well overall and can accurately recognize various entity labels in an abstract. In terms of the macro-average F1-score, which assigns equal weight to all labels, the result is slightly lower at around 80% due to some labels having lower individual scores. Therefore, after comprehensive consideration, the effectiveness of the entity recognition model in this study can be proven.

In order to more accurately measure the effectiveness of our model in fully recognizing entities in the text, we used stricter seqeval metrics to evaluate the model. It is a Python library specifically used for sequence annotation task evaluation, which calculates accuracy, recall, and other metrics on an entity basis, rather than on a token basis. Only when each token of an entity is correctly recognized can a positive sample be counted. At the same time, in order to prove the effectiveness of our entity recognition model, we compared the effect of the model without introducing LSTM and CRF layer, and replaced BiLSTM with the classic BiGRU model for comparison. The precision, recall and F1 value results of each model are shown in Table 7. It can be found that the introduction of LSTM and CRF layer effectively improves the effect of entity recognition, indicating that LSTM can better extract the context characteristics of the text. Compare our model with the model replaced by GRU, and observe the change of F1 value in the training process of the two models, as shown in Fig. 3, we can find that the BiLSTM model has a more stable convergence result and higher evaluation index value than BiGRU.

Table 7. Results of the different entity recognition model

entity identification model	Precision	Recall	F1-score
Bert	0.82	0.76	0.79
Bert-CRF	0.82	0.77	0.80
Bert-BiGRU-CRF	0.83	0.80	0.81
Bert-BiLSTM-CRF	**0.84**	**0.82**	**0.83**

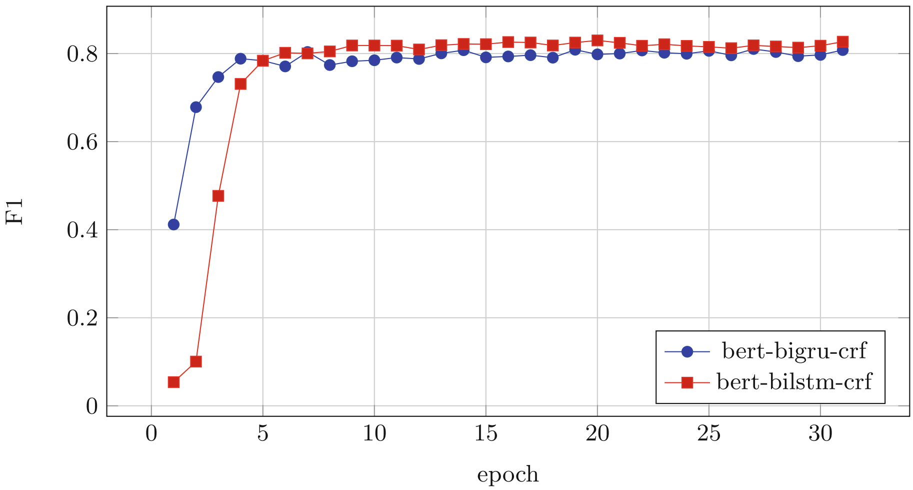


Fig. 3. The effect of LSTM and GRU entity identification model

6 Application

We use the entity recognition model to design an automatic document annotation tool to help researchers quickly understand the key knowledge entities involved in an article and master the core content. By inputting an abstract, the system automatically marks 7 kinds of entity contents including materials, attributions, methods and data results involved in the article, and displays them in different colors on the interface. Users can quickly capture the relevant information of a material and related research knowledge points. For the annotation result returned by the system, researchers can annotate it again, correct or supplement the entity content that is not correctly identified, and then return it to the system for update and improvement.

The interface of the literature annotation tool is shown in Fig. 4. Users can manually enter the content in the text box or click the "Upload Attachment" button to read the content by uploading the file. Select the number and category of entities to be marked in the "Annotation Settings", and then click to start marking, and the system will display the marked results. At the same time, users can make secondary annotations on the results displayed in the system, slide the mouse to select the missing content, and select the entity type to save and submit to the system.

Fig. 4. Functional interface of literature annotation

7 Summary

In order to enable researchers to quickly grasp the key knowledge of the article in reading massive literature information and improve the efficiency of the literature reading, this study focuses on named entity recognition in the field of lithium-ion battery materials literature. Firstly, a combination of automation and manual analysis is used to construct an entity classification system for lithium-ion battery materials. Then, a named entity recognition model based on Bert-BiLSTM-CRF is developed, achieving an F1-score of over 80% in experimental results. This model is applied to the literature annotation function, allowing researchers to read annotated literature with highlighted key knowledge entities, facilitating the extraction of useful information. Our work provides convenience for researchers to mine literature information and plays an important role in promoting the research and application of battery materials.

In the future, we will continue to improve the effectiveness of entity recognition, and on this basis, carry out the relation extraction work between entities, so as to build the knowledge graph in the literature field of lithium-ion battery materials, obtain the correlation knowledge information between different materials, and contribute to knowledge mining and development in the field of battery materials.

Acknowledgements. This work is supported by Jiangxi Provincial Natural Science Foundation under Grant No.20224BAB212015, 20224ACB202007, Science and Technology Project of Jiangxi Provincial Education Department (GJJ160315), the National Natural Science Foundation of China (NSFC) under Grant No.51962010.

References

1. El-Bousiydy, H., Troncoso, J.F., Johansson, P., et al.: LIBAC: an annotated corpus for automated "reading" of the lithium-ion battery research literature. Chemistry Materials **35**(5): 1849–1857 (2023)
2. Weston, L., Tshitoyan, V., Dagdelen, J., et al.: Named entity recognition and normalization applied to large-scale information extraction from the materials science literature. J. Chem. Inform. Modeling **59**(9) (2019). https://doi.org/10.1021/acs.jcim.9b00470.
3. Niu, C.: Research on Solid Identification Method of Alloy Materials Based on Small Annotation [2023-07-04]
4. Huang, S., Cole, J.M.: A database of battery materials auto-generated using ChemDataExtractor. Sci. Data **7**(1), 260 (2020)
5. Swain, M.C., Cole, J.M.: ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature. J. Chem. Inf. Model. **56**, 1894–1904 (2016)
6. Liu, Y., Liu, D.H., Ge, X.Y., et al.: A high-quality dataset construction method for text mining in materials science. J. Phys. **72**(07), 128–141 (2023)
7. Shan, B., Li, H.J., Wen, Y.W., et al.: Material information extraction and data-driven research in the era of large language models. Metallic Funct. Materials **30**(03), 1–16 (2023). https://doi.org/10.13228/j.boyuan.issn1005-8192.20230048.
8. Nie, Z., Liu, Y., Yang, L., Li, S., Pan, F.: Construction and application of materials knowledge graph based on author disambiguation: revisiting the evolution of LiFePO ¡sub¿4¡/sub¿. Adv. Energy Materials **11**(16) (2021). https://doi.org/10.1002/aenm.202003580
9. Nie, Z., Zheng, S., Liu, Y., et al.: Automating materials exploration with a semantic knowledge graph for Li-ion battery cathodes. Adv. Func. Mater. **32**(26), 2201437 (2022)
10. Lü T.T.: Construction and application of knowledge graph oriented to professional literature: taking machine learning literature as an example[D]. Beijing Jiaotong University (2020). https://doi.org/10.26944/d.cnki.gbfju.2020.001085
11. Yang, J., Zhang, Y.: YEDDA: A Lightweight Collaborative Text Span Annotation Tool [J] (2017)
12. Devlin, J., Chang, M.W., Lee, K., et al.: BERT: pre-training of deep bidirectional transformers for language understanding (2018). https://doi.org/10.48550/arXiv.1810.04805