1Н

УДК 519.4

Нао, Н.
Samara National Research University named after Academician S. P. Korolev
34 Moskovskoye shosse, Samara, 443086, Russia

NAMED ENTITY RECOGNITION IN REFRACTORY HIGH-ENTROPY ALLOYS USING DEEP LEARNING

Хао Х.
Самарский национальный исследовательский университет имени академика С. П. Королёва
Россия, 443086, г. Самара, Московское шоссе, 34

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В ТУГОПЛАВКИХ ВЫСОКОЭНТРОПИЙНЫХ СПЛАВАХ С ИСПОЛЬЗОВАНИЕМ ГЛУБИННОГО ОБУЧЕНИЯ

To address the challenges posed by the rapid growth of literature in the field of refractory high-entropy alloys (RHEAs) and the low efficiency of key information extraction, this paper proposes a semi-automated information extraction workflow. The method leverages a large language model for initial annotation, combined with manual review to construct a high-quality corpus. Based on this, a BERT-BiLSTM-CRF named entity recognition (NER) model is trained to automatically identify and extract information related to materials, processing, structure, and properties. The final results show that the NER model achieves an F1 score of 77% on the test set, significantly reducing manual curation costs and providing support for the construction of materials knowledge bases and data-driven research on new materials.
Key words: Refractory high-entropy alloys, named entity recognition, large language model, BERT-BiLSTM-CRF.

В ответ на быстрый рост объема литературы в области тугоплавких высокоэнтропийных сплавов (RHEAs) и низкую эффективность извлечения ключевой информации в данной статье предлагается полуавтоматический рабочий процесс извлечения информации. Метод использует большую языковую модель для первоначальной разметки в сочетании с ручной проверкой для построения высококачественного корпуса. На этой основе обучена модель распознавания именованных сущностей BERT-BiLSTM-CRF для автоматического распознавания и извлечения информации о материалах, процессах, структуре и свойствах. Окончательные результаты показывают, что модель NER достигла оценки F1 в 77% на тестовом наборе, что значительно снизило затраты на ручную обработку и обеспечило поддержку для построения базы знаний о материалах и исследований новых материалов на основе данных.
Ключевые слова: тугоплавкие высокоэнтропийные сплавы, распознавание именованных сущностей, большая языковая модель, BERT-BiLSTM-CRF.

Introduction

Refractory high-entropy alloys (RHEAs) are regarded as potential key materials in fields such as aerospace and nuclear energy due to their exceptional performance under extreme conditions [1]. Compared to traditional nickel-based superalloys, RHEAs exhibit higher melting points, superior high-temperature stability, and outstanding corrosion and oxidation resistance. With the rapid advancement of research, the number of related scientific achievements and academic publications has grown exponentially. Since 2023 alone, thousands of relevant research articles have been added to international databases such as Web of Science and ScienceDirect.

However, a significant amount of critical information in these publications—such as alloy compositions, processing conditions, microstructures, and performance parameters—often exists in the form of unstructured text. Relying solely on manual reading and organization is not only inefficient but also prone to omissions, which has become a bottleneck limiting knowledge utilization and data-driven research.

The rapid development of natural language processing (NLP) technologies has provided new tools for knowledge extraction in materials science. Among these, named entity recognition (NER), as a core task, can convert proper nouns and key parameters in text into structured data, thereby supporting the construction of materials knowledge graphs and databases [2-4]. Previous studies have shown that deep learning-based NER methods (e.g., BiLSTM-CRF, BERT) have achieved remarkable progress in general text processing, and some domain-specific models (e.g., MatSciBERT, MatBERT) have also demonstrated advantages in processing materials science corpora [5], [6]. Nevertheless, tailored research for the specific system of RHEAs remains insufficient. The complexity of professional terminology, ambiguous entity boundaries, and diverse expression forms make general models difficult to apply directly, and there is a lack of efficient and systematic solutions.

At the same time, the emergence of large language models (LLM) has brought new opportunities for domain-specific text mining. LLM have demonstrated excellent performance in zero-shot and few-shot learning, yet they still face challenges related to accuracy and controllability in specialized domain applications. Therefore, the academic community has gradually begun to explore hybrid workflows that combine "LLM-assisted annotation, human review, and deep learning training" to enhance result reliability while maintaining efficiency [7-9].

1. Against this background, this study proposes an intelligent information extraction method for RHEAs. The innovations include:
2. Constructing an entity labeling system covering four dimensions: materials, processsing, structure, and properties;
3. Designing an efficient annotation and modeling pipeline based on "LLM pre-annotation, human review, and deep learning training";
4. Building a high-quality corpus based on over 200 RHEA-related publications and training a BERT-BiLSTM-CRF model to achieve automated extraction of key information.

The research outcomes provide a data foundation for the construction of a RHEA knowledge base and the discovery of new materials.

Dataset Construction and Processing Methods

To construct a high-quality named entity recognition dataset tailored for the RHEAs domain, this study proposes a semi-automated workflow comprising: (1) literature preprocessing and candidate sentence construction; (2) entity annotation based on LLM followed by manual verification; (3) data post-processing and format standardization; and (4) NER model training and optimization.

The research data were sourced from over 200 English-language publications on RHEAs obtained from the Web of Science and ScienceDirect databases. The PDF files were first converted to TXT format using pdfminer, and sentence segmentation was performed based on SpaCy. Candidate sentences were then filtered using a predefined domain-specific terminology database for RHEAs, which includes keywords such as alloy names, processing techniques, and properties. Further cleaning operations involved removing noise such as garbled characters, headers and footers, author and journal information, as well as eliminating sentences with fewer than five words. Approximately 30,000 valid text units were ultimately obtained, with each sentence assigned a unique identifier.

To address the knowledge extraction requirements of RHEA-related texts, a four-category entity labeling system was designed, covering materials, processing, structure, and properties. The specific definitions are as follows:

- MATERIALS: Includes alloys, composites, coatings, and elemental symbols, such as $TiZrHfNbAl_{0.5}$, AlCoCrFeNi, Cr_2O_3, Co, and TiN thin film.
- PROCESSING: Encompasses preparation techniques, analytical methods, and related parameter descriptions, such as arc-melting, annealing at 1200 °C for 24 h, sintering, SEM, TEM, and EBSD.
- STRUCTURE: Covers phase composition and microstructural characteristics, such as BCC phase, Laves phase, dendritic structure, and grain boundary.
- PROPERTIES: Includes both qualitative and quantitative performance descriptions, where numerical values must be accompanied by units—e.g., corrosion resistance, 350 HV, 1100 MPa, and 15% elongation.

During the annotation phase, a large language model (deepseek-r1-distill-llama-70b) was employed for preliminary entity recognition. To ensure output quality, the prompts were meticulously designed to clarify annotation rules for each entity category, supplemented with positive and negative examples. The model-generated annotations were subsequently reviewed and refined manually to form a high-confidence corpus. This process effectively balanced efficiency and accuracy while significantly reducing manual effort. Specific annotation guidelines included:

- Material entities should be prioritized for annotation using full names or chemical formulas.
- Processing entities should include both the technique and associated parameter descriptions, while isolated numerical values were not annotated.
- Structure entities required complete and specific terminology, avoiding generic terms appearing in isolation.
- Property entities encompassed both qualitative and quantitative information, with the latter mandatorily including units.

In the data post-processing stage, the annotated results were automatically converted into start and end positions for each entity via scripting and standardized into the JSONL format. These were then imported into Label Studio for manual visual inspection and correction. Additionally, the data were transformed into the BIOES sequence labeling format to meet subsequent model training requirements. Throughout the entire dataset construction process, sentence identifiers were retained to ensure traceability and consistency.

An example of a single data entry is provided below:

{"data": {"text": "NbMoTaWV and TaNbHfZrTi RHEAs with bcc structures were created for high-temperature applications, demonstrating great temperature strength and outstanding phase stability at 1473 K."}, "annotations": [{"result": [{"type": "labels", "value": {"start": 35, "end": 48, "labels": ["STRUCTURE"]}}, {"type": "labels", "value": {"start": 118, "end": 138, "labels": ["PROPERTIES"]}}, {"type": "labels", "value": {"start": 155, "end": 170, "labels": ["PROPERTIES"]}}, {"type": "labels", "value": {"start": 174, "end": 180, "labels": ["PROCESSING"]}}, {"type": "labels", "value": {"start": 0, "end": 8, "labels": ["MATERIALS"]}}, {"type": "labels", "value": {"start": 13, "end": 23, "labels": ["MATERIALS"]}}]}]}

Experimental Design and Results

To ensure the effectiveness of model training, this study developed a preprocessing pipeline for the JSONL data. The key challenge involved resolving the alignment issue between subword tokenization and character-level annotations: the text and character offset annotations were parsed and converted into BIOES tag sequences. Using the BERT tokenizer's offset_mapping, a mapping was established to assign labels based on the first character of each token. Sequences were uniformly truncated or padded to a length of 256, converted into tensors, and then fed into the model.

For the model architecture, we adopted the well-established BERT-BiLSTM-CRF as the baseline NER model, leveraging pretrained semantic representations, bidirectional sequence modeling, and label dependency constraints to achieve efficient entity recognition. Training was conducted on approximately 17,000 annotated instances with the following hyperparameters: BERT-base, batch size = 16, learning rate = 5e-5, and maximum sequence length = 256. The training process showed stable convergence, and early stopping was triggered at the 8th epoch. The model achieved a macro-average F1 score of 77% on an independent test set. Detailed performance metrics for each category are presented in Table 1. The training and validation loss curves (Figure 2) decreased smoothly and stabilized, indicating an effective training process without overfitting.

To further validate the practicality of the system, we selected two sentences from the literature that were not included in the training data. Here we present one example: "The $V_{2.5}Nb_1Mo_{0.5}Zr$ alloy after annealing at 1000 °C exhibited a predominant BCC phase and a yield strength of 1250 MPa."

The system output was:

"The $V_{2.5}Nb_1Mo_{0.5}Zr$ (MATERIALS) alloy after annealing at 1000 °C (PROCESSING) exhibited a predominant BCC phase (STRUCTURE) and a yield strength of 1250 MPa (PROPERTIES)."

The further extracted knowledge tuple was:

(MATERIALS, PROCESSING, STRUCTURE, PROPERTIES) = ($V_{2.5}Nb_1Mo_{0.5}Zr$, annealing@1000 °C, BCC, 1250 MPa)

This example demonstrates that the system can effectively automatically extract key information such as composition, processing, structure, and properties from complex technical texts and preserve it in a structured form, showing potential for direct application in knowledge base construction and scientific research analysis. The second test case is provided in Figure 3 for reference.

Entity	Precision	Recall	F1-score
MATERIALS	0.89	0.87	0.88
PROCESSING	0.73	0.72	0.72
STRUCTURE	0.67	0.78	0.72
PROPERTIES	0.69	0.63	0.66
Macro Avg	0.76	0.77	0.77
Weighted Avg	0.90	0.90	0.90

Table 1. Entity-level evaluation results

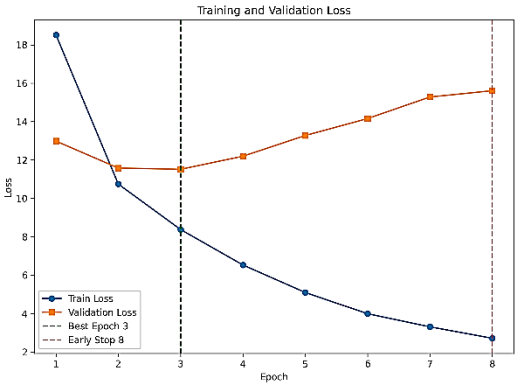

Figure 2 – Training and validation loss curves

```
=== Entity Recognition Result for Sentence 1 ===
Entity                              | Label
--------------------------------------------------
v2.5nb1mo0.5zralloy                 | MATERIALS
annealing                           | PROCESSING
1000°c                              | PROCESSING
bccphase                            | STRUCTURE
1250mpa                             | PROPERTIES
--------------------------------------------------
Original sentence: The V2.5Nb1Mo0.5Zr alloy after annealing at 1000°C exhibited a predominant BCC phase and a yield strength of 1250 MP


=== Entity Recognition Result for Sentence 2 ===
Entity                              | Label
--------------------------------------------------
ti6al4valloy                        | MATERIALS
heat-treated                        | PROCESSING
800°c                               | PROCESSING
950mpa                              | PROPERTIES
--------------------------------------------------
Original sentence: The Ti6Al4V alloy was heat-treated at 800°C, resulting in a tensile strength of 950 MPa.
```

Figure 3 – Entity recognition results for two test sentences from RHEA literature

Discussion

This study not only demonstrates the feasibility of semi-automated named entity recognition in the RHEA domain, but also highlights broader implications for scientific text mining and materials informatics. Compared with purely manual annotation, the proposed workflow substantially reduces the effort required from domain experts while maintaining high accuracy, thereby addressing the urgent need for efficient knowledge integration in rapidly growing scientific literature. By combining carefully designed prompts for large language models with rigorous human verification, the approach mitigates ambiguities associated with complex expressions such as alloy compositions, parameterized processing conditions, and microstructural descriptors. The resulting corpus of more than 17,000 annotated sentences is among the most comprehensive resources currently available for refractory alloys and can serve as a benchmark for future work in this field.

Equally important is the interdisciplinary nature of the workflow. Natural language processing specialists provide methodological expertise in model design and optimization, while materials scientists contribute domain knowledge to refine entity definitions and validate results. This synergy ensures that extracted information is both technically accurate and contextually meaningful for downstream applications, and exemplifies a practical model of human–AI cooperation in data-intensive research.

The structured information derived from the workflow also holds potential for applications beyond named entity recognition. It can serve as the foundation for constructing domain-specific knowledge graphs, support hypothesis generation, and reveal hidden links between composition, processing, structure, and properties. Furthermore, the integration of literature-derived knowledge with experimental and computational databases may ultimately enable closed-loop materials design. Thus, the contribution of this study extends beyond improving NER performance, offering a generalizable framework to accelerate data-driven research and innovation in advanced materials.

Conclusion

This work proposes a semi-automated workflow for information extraction from refractory high-entropy alloy literature, integrating large language models, human verification, and deep learning. The resulting BERT-BiLSTM-CRF model achieved an F1 score of 77%, confirming the effectiveness of the approach. The study contributes a large, high-quality annotated corpus and validates the feasibility of combining LLM-assisted annotation with domain expertise.

Future research will focus on extending the method to nested entity and relation extraction, constructing a comprehensive "composition–processing–structure–property" knowledge graph, and integrating literature mining with performance prediction in an intelligent design platform for advanced materials.

Author's Declaration

No funding was received for the research or authorship of this work.

The author employed DeepSeek-V3, a large language model, to assist in editing and improving the readability of the manuscript, including the refinement of Russian-language content. The ideas, content, and final expression remain the author's own work.

References

1. Miracle, D. B., Senkov, O. N. A critical review of high entropy alloys and related concepts. Acta Materialia. 2017, 122: 448–511.
2. Hearst, M. A. Automatic acquisition of hyponyms from large text corpora. Proceedings of the 14th Conference on Computational Linguistics (COLING). 1992: 539–545.
3. Lafferty, J., McCallum, A., Pereira, F. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning (ICML). 2001: 282–289.
4. Huang, Z., Xu, W., Yu, K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint. arXiv:1508.01991, 2015.
5. Devlin, J., Chang, M. W., Lee, K., Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. 2019: 4171–4186.
6. Kim, E., Huang, K., Jegelka, S., Olivetti, E. Virtual screening of inorganic materials synthesis parameters with deep learning. npj Computational Materials. 2017, 3: 53.
7. Trewartha, A., Dagdelen, J., Huo, H., Cruse, K., Riebesell, J., Jain, A., Ceder, G., Persson, K. A. Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science. Patterns. 2022, 3(4): 100488.
8. Gupta, T., Trewartha, A., Cruse, K., Dagdelen, J., Huo, H., Ceder, G., Jain, A., Persson, K. A. MatSciBERT: A materials domain language model for text mining and information extraction. npj Computational Materials. 2022, 8: 102.
9. Brown, T. B., Mann, B., Ryder, N., et al. Language models are few-shot learners. Advances in Neural Information Processing Systems (NeurIPS). 2020, 33: 1877–1901.

RESUME

Hao, H.

Named Entity Recognition in Refractory High-Entropy Alloys Using Deep Learning

Background: Refractory high-entropy alloys (RHEAs) have attracted increasing attention in aerospace and nuclear engineering due to their exceptional thermal stability and mechanical performance. The rapid growth of publications in this field presents challenges for systematically extracting critical knowledge, as information on alloy compositions, processing routes, microstructures, and properties is often embedded in unstructured text.

Materials and methods: To address this challenge, a semi-automated workflow was developed. Large language models were employed for initial entity annotation, followed by expert verification to ensure data quality. A domain-specific dataset was compiled from over 200 English-language papers, covering four entity types: materials, processing, structures, and properties. Subsequently, a BERT-BiLSTM-CRF model was trained for named entity recognition (NER).

Results: The workflow produced approximately 17,000 annotated sentences. Evaluation demonstrated that the trained model achieved an F1-score of 77% on the test set. The system

effectively identified alloy names, processing parameters, structural characteristics, and performance indicators, confirming its reliability for automatic knowledge extraction.

Conclusion: The proposed method substantially reduces manual annotation effort while ensuring accurate extraction of domain-specific knowledge. It provides a solid foundation for building materials knowledge bases and supports data-driven approaches for the discovery of new alloys.

РЕЗЮМЕ

Хао Х.

Распознавание именованных сущностей в тугоплавких высокоэнтропийных сплавах с использованием глубинного обучения

Тугоплавкие высокоэнтропийные сплавы (RHEAs) привлекают все большее внимание в аэрокосмической и ядерной областях благодаря своей исключительной термической стабильности и механическим характеристикам. Быстрый рост публикаций в этой области создает проблемы для систематического извлечения критически важных знаний, поскольку информация о составе сплавов, методах обработки, микроструктуре и свойствах часто представлена в неструктурированном текстовом виде.

Для решения этой задачи был разработан полуавтоматизированный рабочий процесс. Большие языковые модели использовались для первоначальной разметки сущностей с последующей экспертной верификацией для обеспечения качества данных. Специализированный набор данных был составлен из более чем 200 англоязычных статей и охватил четыре типа сущностей: материалы, обработка, структуры и свойства. Затем была обучена модель BERT-BiLSTM-CRF для распознавания именованных сущностей (NER).

В результате работы процесса было размечено приблизительно 17 000 предложений. Оценка показала, что обученная модель достигла F1-показателя в 77% на тестовом наборе. Система эффективно идентифицировала названия сплавов, параметры обработки, структурные характеристики и показатели свойств, что подтвердило её надежность для автоматического извлечения знаний.

Предложенный метод существенно сокращает усилия по ручной разметке при обеспечении точного извлечения доменно-специфичных знаний. Он обеспечивает прочную основу для создания баз знаний о материалах и поддерживает подходы, основанные на данных, для открытия новых сплавов.

Hao Hu – PhD student, Samara National Research University. Research fields: Machine Learning, Metallic Materials, Metallurgical Engineering. Tel.: +86 18104860834 (China) / +7 980 908 3470 (Russia). E-mail: 641229879@qq.com ORCID: 0009-0004-2902-2593

Статья поступила в редакцию 29.08.2025.