NER for Specialized Scientific Domains: Fine-Tuning on Patents for Plasma Technology and Battery Materials

Farag Saad[1], Hidir Aras[1], Markus M. Becker[2] and Carsten Becker-Willinger[3]

[1]Leibniz Institute for Information Infrastructure (FIZ Karlsruhe), Hermann-von-Helmholtz Platz 1, 76344 Eggenstein-Leopoldshafen, Germany
[2]Leibniz Institute for Plasma Science and Technology (INP), Felix-Hausdorff-Str. 2, 17489 Greifswald, Germany
[3]Leibniz Institute for New Materials (INM), Campus D2 2, 66123 Saarbrücken, Germany

Abstract

Domain-specific Named Entity Recognition (NER) allows to identify and extract specific types of entities from text. In particular for technical domains such as plasma technology and battery materials extracting and aligning such entities with complex (structured) semantic information such as in Knowledge Graphs (KG) plays a crucial role. In this work, we fine-tuned SciBERT, BERT-for-Patents, and BatteryBERT for domain-specific NER based on systematically constructed annotated datasets specific to the regarded domain. Despite the relatively limited size of the training data, particularly for battery materials, the models achieved strong overall performance. By leveraging the linguistic knowledge encoded in the pretrained models, combined with domain-specific patterns learned from the training datasets, the developed models effectively identified and classified entities based on their contextual usage. Our evaluation demonstrated that fine-tuning domain-adapted pretrained models significantly enhance NER effectiveness in specialized scientific and technological domains.

Keywords

Named Entity Recognition (NER), NLP, Deep learning, Plasma Technology, Battery Materials

1. Introduction

Named Entity Recognition (NER) is a key task in Natural Language Processing (NLP) that involves identifying and classifying entities within unstructured text. In scientific and technical domains such as Plasma Technology (PT) and Battery Materials (BM), NER plays a critical role in extracting structured information from complex texts and aligning it with explicit semantic models such as in Knowledge Graphs. However, applying NER to these specialized fields presents unique challenges. Patent literature, in particular, is filled with ambiguous terminology, unconventional phrasing, and rapidly evolving terminology, making entity recognition difficult [1]. Additionally, there is a limited availability of annotated corpora, further complicating the task.

Effective NER models must therefore be tailored to handle domain-specific complexities and trained on high-quality annotated datasets. Accurate identification of entities is essential for enabling advanced downstream tasks such as knowledge graph construction, literature mining, and patent analysis. Consequently, improving NER in PT and BM domains supports the broader goal of making scientific information more accessible and interoperable.

Recent adaptations of BERT (Bidirectional Encoder Representations from Transformers) model [2], such as SciBERT [3], BatteryBERT [4], and BERT-for-Patents [5], have shown promise in specialized scientific and technical domains. However, few studies have focused on the application of these models to PT and BM, particularly in the context of patent literature. Existing approaches lack the customization needed to address the specific challenges of these domains, such as the frequent emergence of new terms and the abstract writing style common in patents.

In this paper, we developed an approach, to capture the unique terminology and context of PT and BM within patent texts by leveraging domain-specific BERT variants, such as SciBERT, BERT-for-Patents,

6th Workshop on Patent Text Mining and Semantic Technologies (PatentSemTech) 2025
*Corresponding author. Email: farag.saad@fiz-karlsruhe.de
farag.saad@fiz-karlsruhe.de (F. Saad); hidir.aras@fiz-karlsruhe.de (H. Aras); markus.becker@inp-greifswald.de (M. M. Becker); Carsten.Becker-Willinger@leibniz-inm.de (C. Becker-Willinger)

and BatteryBERT. To achieve this, we have curated high-quality, domain-specific training datasets (See Section 3) that accurately reflect the specific nature of these technologies. By fine-tuning these BERT variants, we train several NER models to effectively extract and classify entities relevant to these specialized fields.

This methodology allows us to address the unique challenges posed by the specialized language and rapidly evolving terminologies present in patent documents, ultimately contributing to more accurate and efficient information extraction in these critical areas of research.

2. Related Work

Traditional NER approaches, including rule-based, statistical, and hybrid methods, have been used in technical domains but often struggle with the complexity of domain-specific language. These methods typically rely on predefined rules or feature engineering, which makes them less adaptable to the dynamic and evolving vocabulary found in technical domains [6]. While dictionary-based approaches and pattern-matching methods have shown some success in specific domains, they often lack of generalizability needed for broader applications [7]. The emergence of deep learning models, particularly transformer-based architectures such as BERT, has significantly advanced NER performance. These models are capable of learning contextual dependencies and handling subword-level semantics, which are critical for understanding the complex material compositions and experimental parameters found in engineering and scientific texts. However, pretrained models such as BERT often struggle with domain-specific terminology and specialized language used in patent documents, where terminology may be newly invented or ambiguous.

The emergence of deep learning has significantly improved NER performance, particularly with the use of deep learning approaches, which alleviate the need for extensive feature engineering by learning contextual dependencies [8]. However, these models still face limitations in capturing long-range dependencies and incorporating subword-level semantics, both of which are critical for accurately parsing complex material structure [9]. Recent advances in transformer-based architectures, in particular BERT model, have revolutionized NER tasks by offering contextualized embeddings and enabling fine-tuning on downstream applications. Domain-specific BERT variants such as SciBERT [3], BatteryBERT [4] and BERT-for-Patents [5] have been developed to capture the specificity of scientific and materials-specific terminology, yielding significant improvements over general-purpose models.

SciBERT is a pretrained language model specifically designed for scientific texts, demonstrating superior performance over general-purpose models like BERT when applied to scientific content. Trained on a vast collection of scientific papers, SciBERT effectively captures domain-specific language, terminology, and syntactic structures. The model has shown significant improvements in various scientific tasks, including scientific paper classification, named entity recognition (NER), and relation extraction. Its ability to understand complex scientific language makes it particularly valuable for domains with specialized vocabularies, such as biomedical research and chemistry. SciBERT's architecture allows it to generalize well across different scientific disciplines while maintaining high accuracy in domain-specific tasks. As a result, SciBERT is widely adopted in natural language processing pipelines tailored to scientific literature analysis.

BatteryBERT, developed by Huang and Cole (2022), is a domain-specific variant of the BERT model that has been fine-tuned specifically for the field of battery research, capturing the unique terminology and concepts within this domain. As a result, BatteryBERT significantly outperforms general-purpose models in extracting crucial information from battery-related texts. Its training on a specialized corpus enables it to better recognize complex technical entities and fine-grained semantic distinctions commonly found in battery materials literature. This makes it particularly effective for tasks such as material property extraction and electrode classification. Moreover, BatteryBERT's focused pretraining helps reduce errors caused by ambiguous terms and enhances its ability to interpret context-specific complexities. This specialization ultimately leads to more reliable and accurate finetuning for information extraction in battery research applications.

BERT-for-Patents is a specialized variant of the BERT model, designed specifically for patent documents. By leveraging BERT's capabilities, this model has been fine-tuned to understand the unique terminology and structure of patent texts, offering significant improvements over general-purpose language models. BERT-for-Patents has been demonstrated to enhance tasks such as patent classification, text mining, which are essential for efficient patent analysis and retrieval. Its training on a large corpus of patent literature allows it to capture domain-specific language patterns, legal jargon, and complex sentence structures that are typical in patents. This specialization allows for more accurate fine-tuning in tasks like entity recognition and relationship extraction, outperforming generic language models in the patent domain. As a result, the model can be effectively adapted to various technical fields within patents, making it a powerful and flexible tool for understanding intellectual property.

Building on specialized variants of these BERT model, recent research has focused on fine-tuning models such as SciBERT [3], BioBERT [10], and BlueBERT [11] on domain-specific corpora. These fine-tuned models have demonstrated significant improvements in performance across various low-resource scientific domains. For instance, Rostam and Kertész (2024) fine-tuned BERT-based models for scientific text classification tasks and found that domain-specific models like SciBERT consistently outperformed general-purpose models [12]. Despite significant advancements in BERT-based NER, their application in highly specialized domains such as PT and BM remains largely underexplored. Fine-tuning pretrained scientific language models on domain-specific corpora presents a promising strategy to bridge this gap.

In this work, we investigate the effectiveness of fine-tuning transformer-based models for NER tasks using patent documents from the PT and BM domains. To support this effort, we introduce a domain-specific annotated corpus and fine-tune multiple BERT-based model variants to enhance entity recognition performance. Our approach offers a comprehensive solution for information extraction in these complex and rapidly advancing scientific and technological fields. To the best of our knowledge, this is the first study to investigate Named NER in the domains of PT and BM within the context of patent literature.

3. Training Data Construction

Developing a robust NER model for specialized domains requires high-quality, domain-specific training data. Given the lack of publicly available annotated corpora for the regarded domains, particularly within patent literature, we constructed our own annotated dataset by systematically labeling the titles and abstracts of selected patent documents from each domain.

To initiate the annotation process, we first prepared an initial list of seed entities relevant to PT and BM. These seed entities served as the basis for pre-annotating the corpus, which was subsequently reviewed and corrected by domain experts using a Prodigy annotation tool[1] (See Figure 1) , thereby reducing the manual workload for human annotators. Candidate entity lists were automatically extracted from two domain-specific Wikipedia categories: Plasma Physics[2] and Battery (Electricity)[3]. While these categories provided a broad range of potential entities central to PT and BM, not all extracted entities were relevant. Therefore, a filtering step was introduced to ensure only relevant entities were selected.

The filtering process involved matching the extracted entities against a corpus of patent texts related to PT and BM domains. Entities were ranked by their frequency of occurrence to prioritize entities that were more likely to be relevant to patent literature. Furthermore, domain-specific relevance was verified by checking that each entity's context aligned with the scope of PT and BM technologies, respectively. A shortlist of entities was manually reviewed by domain experts to ensure accuracy. During this review, irrelevant entities were excluded, and additional entities were incorporated using expert knowledge and external resources, including relevant ontologies that offered a comprehensive view of the domain-specific vocabulary [13] and [14].

This procedure and ontology-based selection of core entity types also supported a high inter-annotator

[1] https://prodi.gy/
[2] https://en.wikipedia.org/wiki/Category:Plasma_physics
[3] https://en.wikipedia.org/wiki/Category:Electric_battery

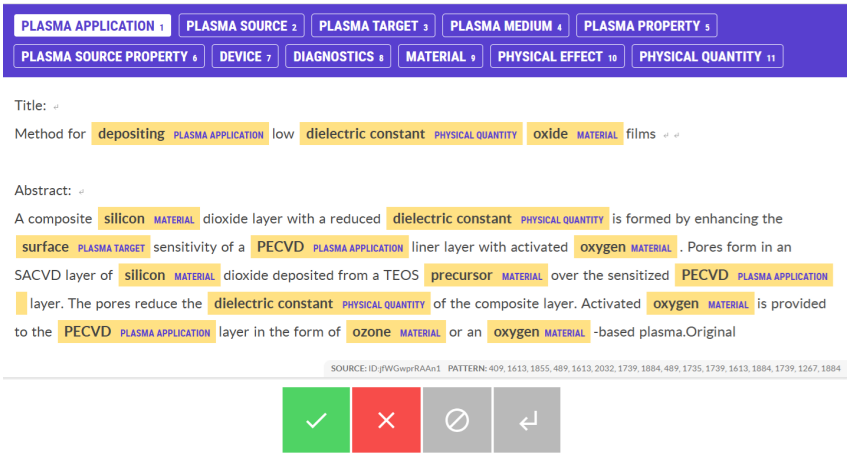

Figure 1: Example of visually pre-annotated data

agreement (IAA), which was evaluated by adopting a practical and widely accepted approach based on a representative sample of the training data. This strategy aligns with best practices used in large-scale annotation projects, such as the methodology followed by Google Healthcare, where IAA is typically calculated on a subset (usually 5–20%) to assess annotator consistency without duplicating annotation efforts across the full dataset[4]. We began with a preliminary session where annotators collaboratively reviewed the annotation guidelines, discussed ambiguous cases, and resolved potential points of confusion. Since the core entity types were selected based on well accepted concepts of the respective research domain, a high IAA of >80% was achieved from the beginning and this phase mostly resolved technical questions related to the annotation process. It should be noted that this resource-efficient process, which provided a common understanding of the task, did not allow us to measure the IAA for the entire dataset. Nevertheless, it ensured the consistency and quality of the annotations, similar to the phased training and review approach used in Google Healthcare's annotation framework.

The annotation process was iterative. In each round, the newly annotated entities from the previous iteration were incorporated into the pre-annotation pipeline, allowing for continuous refinement. This iterative approach not only improved the accuracy and coverage of the dataset but also enabled the corpus to expand with new entities that emerged during the annotation process. Over time, this iterative refinement helped capture emerging trends and terminology, further enhancing the relevance and quality of the dataset.

The final set of entity types was defined in collaboration with domain experts. For PT, the core entity types are: *Plasma Application, Plasma Target, Plasma Source, Plasma Medium, Plasma Property, Plasma Source Property, Device, Diagnostics, Material, Physical Effect, and Physical Quantity.* For BM, the core entity types are: *Property, Cathode, Anode, Technology, Additive, Component, and Electrolyte.* These entity types represent fundamental concepts critical for understanding the fundamental principles within PT [15, 16] and BM. Table 1 and Table 2 show the core entity types along with their frequencies in the annotated data for the Plasma Technology and Battery Materials domains, receptively.

4. Approach based on Fine-Tuning BERT Model variants

BERT is an exceptionally powerful, general-purpose language model that can be fine-tuned for a wide range of text-based machine learning tasks. Instead of training models from scratch, one can leverage pre-trained BERT variants, such as SciBERT, BERT-for-Patents, and BatteryBERT, which have been specifically trained on domain-specific corpora. These models can then be fine-tuned for various NLP tasks, including NER, classification, sentiment analysis, etc.

[4]https://github.com/google/healthcare-text-annotation/blob/master/methodology/annotation-methodology.md

Table 1
Entity type distribution in the annotated Plasma Technology corpus

Type	Distinct Entity Count
Device	1341
Material	820
Physical Quantity	668
Plasma Application	591
Plasma Target	490
Physical Effect	437
Plasma Source Property	375
Plasma Source	327
Plasma Medium	224
Plasma Property	214
Diagnostics	79

Table 2
Core Entity type distribution in the annotated Battery Materials data

Type	Distinct Entity Count
Technology	545
Property	510
Component	429
Anode	295
Cathode	247
Electrolyte	240
Additive	183

Fine-tuning involves adapting a pre-trained model, like BatteryBERT, which has already learned a rich representation of language from large-scale unsupervised training, to a specific downstream task. This is achieved by training the model on a smaller, task-specific labeled dataset. During fine-tuning, the weights of the pre-trained model are adjusted to optimize performance for the target task, such as NER. By leveraging the general language knowledge embedded in the pre-trained model, fine-tuning enables the model to quickly adapt to specialized tasks, even with limited task-specific training data.

Figure 2 illustrates the high-level components involved in fine-tuning a selected pre-trained BERT models for the NER task. We treat the identification of entities in patent text as a sequence labeling task, where a label is assigned to each word or token within the identified entity or phrase, based on its context. Each sequence of tokens are tokenized $token_1, token_2, .., token_n$ using the appropriate tokenizer.

The tokenizer usually splits tokens into sub-tokens (or sub-words) where some special tokens ($[CLS]$ and $[PAD]$) are added. The $[CLS]$ token refers to the beginning of the sentence or sequence, as in BERT the sequence length is fixed (max 512 token), the $[PAD]$ token is responsible for unifying the length of each sentence to the longest one in that all sentences fed to the BERT model must have the same length. As figure 2 shows, based on the input sequence three different BERT contextual embeddings for each tokenized token $v_1, v_2, .., v_n$, capturing each token's context through many of attention heads are generated.

- The token embeddings is calculated based on token, however, if any token is not present in the selected BERT model vocabulary, BERT tries to generate its embeddings based on the sub-words level.
- The Segment/Sentence embeddings is calculated based on a single segment or two segments. As two segments are present in the same sequence they are separated by the [SEP] token where each segment has its own embeddings.
- The position embeddings represents the token's position within a sentence e.g., to identify which is the first, second, third, etc., token of the sentence.

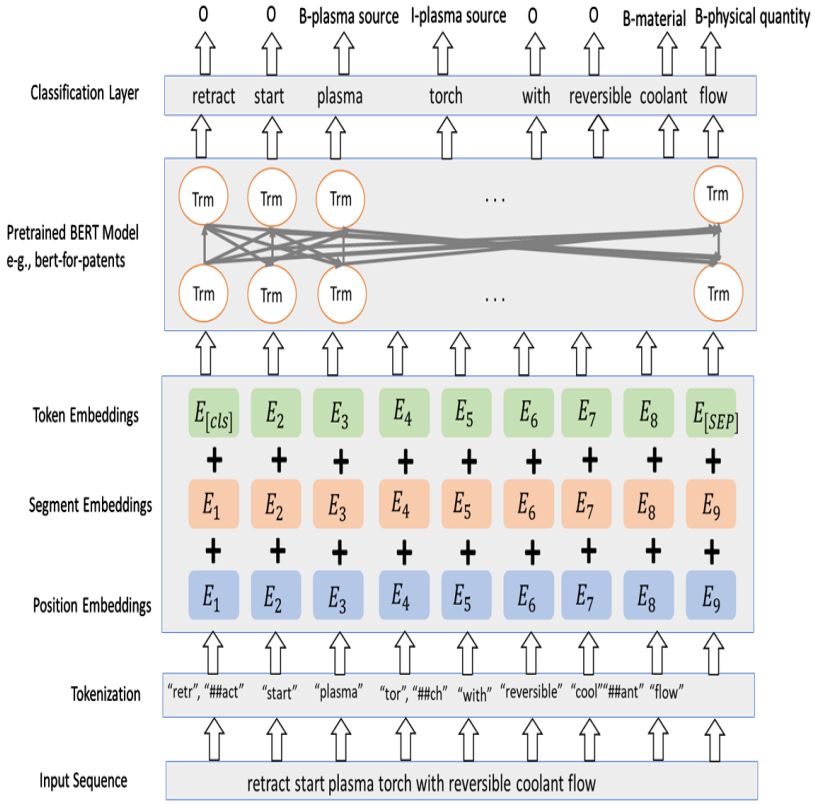

Figure 2: Overall fine-tuned NER model architecture

Once the embeddings are generated they will be summed and fed to the output layer that is the classification layer to classify each sequence token to its relevant label $y_1, y_2, .., y_n$.

Specifically, the fine-tuning process begins by extending the base BERT architecture with a token-level classification layer, transforming it into a model suitable for sequence labeling tasks. This additional layer enables the model to predict entity labels for individual tokens within a given input sequence, leveraging the rich contextual embeddings produced by the underlying BERT layers. Once the architecture is adapted, the model is fine-tuned on a curated, manually annotated dataset specific to the target domain. Each token is paired with its corresponding entity label, enabling the model to learn explicit associations between contextual usage and semantic categories. The fine-tuning is conducted in a supervised manner using gradient-based optimization techniques, during which the parameters of the pre-trained model are updated to reflect domain-specific linguistic and semantic patterns. Importantly, the knowledge embedded in the pre-trained model provides a strong basis of general language understanding, while the domain-specific labeled data serves as a corrective signal that guides the model toward recognizing entities unique to the specialized context. This synergy allows the model to achieve robust performance even when fine-tuning data is limited, by efficiently integrating prior knowledge with task-specific supervision.

In Section 5, we present the evaluation methodology used to assess the performance of the developed NER models for the plasma technology and battery materials domains within patent texts.

5. Evaluation

For the training and testing of the developed NER models, the dataset was divided into two parts: 80% for training and 20% for testing. To evaluate the effectiveness of the developed models in low-resource scientific domains, we conducted experiments on two distinct patent-related use cases: PT (see Table 3) and BM (see Table 4). The models evaluated include fine-tuned SciBERT and BERT-for-

Patents. Additionally, for the Battery Materials use case, another domain-specific pre-trained model, BatteryBERT, was fine-tuned and evaluated. Model performance was assessed at the entity level across relevant scientific core entity types, using standard metrics of precision, recall, and F1-score.

In the PT domain, both SciBERT and BERT-for-Patents demonstrated competitive performance, achieving relatively high F1-scores for most core entity types. For example, *Physical Quantity* (81%), *Plasma Medium* (79%), and *Plasma Source* (78%). These results indicate that both fine-tuned general scientific models (SciBERT) and patent-specific models (BERT-for-Patents) are capable of extracting semantically rich domain entities with reasonable accuracy.

Table 3
Plasma Technology NER models overall scores for precision, recall, and F1-score based on fine-tuning SciBERT and BERT-for-Patents (ranked by F-Score).

Entity Type	SciBERT			BERT-for-Patents		
	Prec	Rec	F1	Prec	Rec	F1
Physical Quantity	0.84	0.78	**0.81**	0.83	0.78	0.80
Plasma Medium	0.87	0.72	**0.79**	0.83	0.72	0.77
Plasma Source	0.85	0.73	**0.78**	0.84	0.71	0.77
Plasma Application	0.76	0.73	0.74	0.76	0.75	**0.76**
Material	0.78	0.71	0.74	0.78	0.73	**0.75**
Device	0.77	0.70	**0.73**	0.74	0.70	0.72
Plasma Source Property	0.69	0.68	0.69	0.70	0.68	**0.70**
Physical Effect	0.72	0.65	**0.68**	0.70	0.64	**0.68**
Plasma Target	0.68	0.55	**0.61**	0.62	0.57	0.59
Plasma Property	0.50	0.48	**0.50**	0.48	0.48	0.48
Diagnostics	0.42	0.42	0.43	0.48	0.50	**0.49**

Table 4
Battery Materials NER models overall scores for precision, recall, and F1-score based on fine-tuning BatteryBERT, SciBERT, and BERT-for-Patents (ranked by F-Score).

Entity Type	BatteryBERT			SciBERT			BERT-for-Patents		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Cathode	0.78	0.80	**0.79**	0.67	0.70	0.68	0.69	0.78	0.73
Anode	0.71	0.70	**0.70**	0.65	0.60	0.62	0.64	0.61	0.63
Additive	0.68	0.68	**0.68**	0.64	0.64	0.64	0.66	0.71	**0.68**
Electrolyte	0.69	0.64	**0.67**	0.70	0.56	0.62	0.69	0.62	0.66
Property	0.68	0.67	0.67	0.69	0.70	0.69	0.69	0.71	**0.70**
Component	0.65	0.67	**0.66**	0.71	0.63	**0.66**	0.64	0.66	0.65
Technology	0.60	0.52	**0.56**	0.57	0.47	0.52	0.56	0.47	0.51

In particular, SciBERT slightly outperformed BERT-for-Patents for core entity types such as *Plasma Medium* and *Device*. This slightly better performance can be attributed to SciBERT's pretraining corpus, which consists of large-scale scientific texts spanning diverse disciplines, including physics, engineering, and biomedical sciences. Scientific publications frequently describe experimental setups, apparatuses, and material environments in detail, providing SciBERT with richer contextual insight to technical entities. This broader scientific context enhances the model's ability to generalize and accurately identify entities in patent-based evaluation sets, even when certain entity terms are underrepresented in the corpus. Conversely, BERT-for-Patents, while trained on a large corpus of patent documents, may face slight challenges in capturing the richer context for certain core entity types. Patent language often focuses on legal or application aspects, which can make technical descriptions unclear. This likely explains why BERT-for-Patents underperformed on entity types that require high technical specificity.

Particularly, both models showed reduced performance on underrepresented or ambiguous entity types such as *Diagnostics* and *Plasma Property*. These performance drops are likely due to the inherent vagueness of these entity types and their relatively low frequency in the annotated training data.

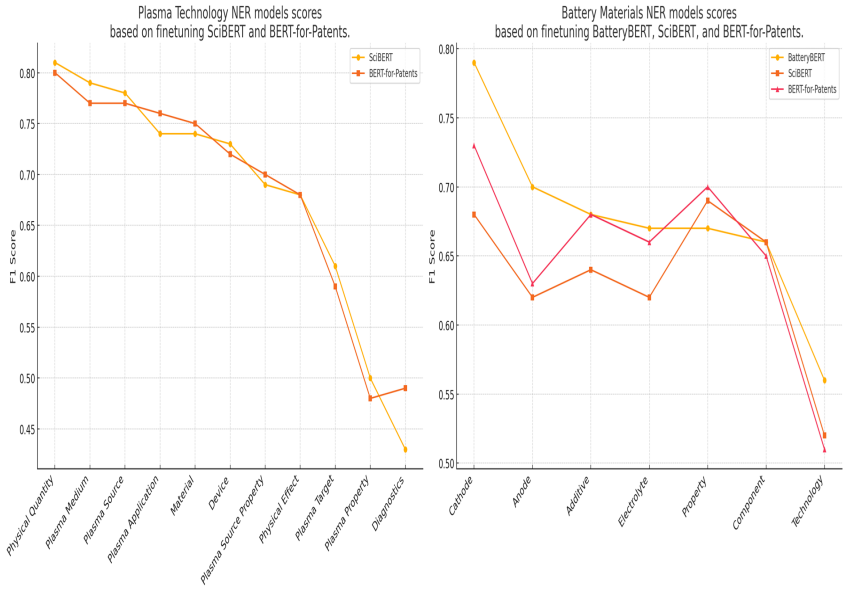

Figure 3: PT and BM NER models overall scores for precision, recall, and F1-score

The BM domain presented greater challenges, primarily because the available annotated dataset was smaller and had a narrower focus (about 59% smaller than the Plasma Technology training data). The complexity of battery materials terminology, including detailed chemical and material properties, made it difficult to achieve high performance across all entity types with the small training dataset. Additionally, the wide variety of battery applications and evolving terminology caused confusion, making it harder for the models to adapt to the changing language without fine-tuning on a larger, balanced, and representative training dataset. This emphasizes the importance of maintaining a balanced distribution of entity types to ensure consistent generalization across all core entity types.

Despite these limitations, the BM NER models achieved promising results for many core entity types. In particular, the BM NER model fine-tuned on BatteryBERT outperformed both SciBERT and BERT-for-Patents on most core entity types, such as *Cathode* and *Anode*. BatteryBERT was specifically pretrained on a domain-specific corpus of battery-related scientific literature, which likely contributed to its strong performance. However, for more general entity types like *Property*, SciBERT and BERT-for-Patents performed better, achieving slightly higher F1-scores (69% and 70%, respectively). This can be attributed to the broader training corpora of these models, which include frequent and varied mentions of general scientific entities across different fields. As a result, they are better able to generalize the meaning of cross-domain concepts like *Property*.

The results from the Battery Materials domain showed a significant performance drop for the *Technology* core entity type across all models, indicating that semantic ambiguity and contextual overlap with general scientific or industrial terminology make this entity type particularly challenging to disambiguate. The *Electrolyte* core entity type, which involves complex and context-dependent chemical formulations, exhibited slight drops in recall across all models. This likely related to insufficient representation of such technical details in the training data, highlighting the need for further refinement in domain-specific data annotation. Addressing this gap is planned for the next iteration of data annotation and model training, as enriching the training corpus with more detailed chemical terminology will likely improve model performance.

Moreover, these findings suggest that while transformer-based models, such as SciBERT, BERT-for-Patents, and BatteryBERT, perform well in the identification of specialized patent-related entities, performance varies significantly based on the extent of domain-specific training. The evaluation also highlights the importance of training dataset size and quality in fine-tuning domain-specific models. The relatively small annotated dataset in the Battery Materials domain highlights a potential

limitation of transformer models in low-resource domains, where models struggle to learn from limited training examples. The model's performance on low-frequency entity types indicates that significant improvements could be achieved with a larger, more comprehensive training dataset.

Overall, the evaluation demonstrates that even with relatively small, manually curated training datasets, fine-tuned transformer models can achieve strong performance in identifying technical concepts within complex patent texts. Domain-specific pretraining plays a critical role: SciBERT provides robust general scientific grounding across a broad range of entity types, while BatteryBERT offers fine-grained precision for battery-specific entity types. Meanwhile, patent-specific models like BERT-for-Patents, although broadly relevant, may not adequately distinguish battery-related entities such as *Anode* and *Cathode*. These observations suggest that further exploration into domain-specific architectures could enhance the performance of NER models in specialized scientific fields.

Figure 3 shows these trends, illustrating how pretraining objectives and domain alignment affect model performance for specific core entity types.

6. Future Work

The future work aims to address the limitations observed during model evaluation and improve the overall performance of the NER models. Several key areas are currently being developed and refined:

- *Improved Data Annotation*: We are expanding and refining the annotated datasets for the Plasma Technology and Battery Materials domains, with a focus on a new domain, *Additive Manufacturing (AM)*. For example, we are adding more detailed chemical and material property terms to the Battery Materials dataset to improve model performance in categories like *Electrolyte*, where performance gaps were identified. We are also prioritizing a balanced distribution of entity types to enhance model generalization across all core entity types, such as *Plasma Property* and *Diagnostics*.
- *Domain-Specific Fine-Tuning*: To further enhance model performance, we are exploring additional fine-tuning techniques using domain-specific corpora. This involves incorporating specialized texts that reflect the evolving terminology in the focused domains. For instance, we plan to annotate patent texts from other sections, such as the *Description* and *Claims*, which provide richer context for domain-specific entity types. The *Description* section often elaborates on the technical details of the invention, while the *Claims* outline the specific legal and functional aspects, both of which are crucial for improving entity recognition in a more domain-specific context.
- *Addressing Semantic Ambiguities*: We are developing methods to address semantic ambiguities, particularly for entity types like *Technology*, where overlapping terms with general scientific or industrial language make identification more difficult. Our approach involves strategies for better disambiguation, using richer contextual information. This includes leveraging advanced language representations to capture fine differences in meaning and improving the model's capacity to differentiate between closely related concepts. Ultimately, these efforts aim to reduce labeling errors and enhance the precision of entity recognition in complex texts.
- *Model Architecture Enhancements*: Ongoing experiments are exploring potential improvements in model architecture and training strategies. We are considering incorporating domain-specific knowledge, such as integrating domain-specific ontologies into the training pipeline to further improve model generalization across subdomains by better capturing hierarchical relationships between entities and contextual dependencies in complex sentence structures.

7. Conclusion

In this work, we presented the development of domain-specific NER models for plasma technology and battery materials. By fine-tuning SciBERT, BERT-for-Patents, and BatteryBERT on high-quality, domain-specific datasets derived from patent texts, we demonstrated that domain adaptation significantly

improves NER performance, even when training data is relatively small. In the plasma technology domain, SciBERT and BERT-for-Patents achieved competitive F1-scores across a wide range of core entity types. In the battery materials domain, the NER models fine-tuned on BatteryBERT, a model pretrained on battery-related scientific literature, delivered particularly strong performance. In particular, fine-tuning BatteryBERT led to superior results compared to general-purpose models, especially for core entity types such as *Cathode* and *Anode*, despite the small size of the available training data. Notably, the models performed robustly despite the inherent ambiguity and syntactic complexity of patent language, which often lacks clarity and consistency. This underlines their capacity to adapt to challenging textual environments. Such adaptability is crucial for extracting structured knowledge from domains where well-defined language is not always used These results highlight the effectiveness of domain-specific fine-tuning in low-resource settings and demonstrate the critical role of specialized language models in advancing information extraction from highly specialized scientific fields. Future work will focus on further enriching the training datasets, both in terms of quality and quantity, by curating a more balanced and comprehensive raw corpus from patent texts. Particular attention will be given to semantically strengthening the representation of all core entity types, especially those with currently limited training instances.

Acknowledgments

This work was partly funded by the DFG project Patents4Science, Project id: 496963457.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] F. Saad, H. Aras, R. Hackl-Sommer, Improving Named Entity Recognition for Biomedical and Patent Data using Bi-LSTM deep neural network models, in: E. Métais, F. Meziane, H. Horacek, P. Cimiano (Eds.), Natural Language Processing and Information Systems, Springer International Publishing, Cham, 2020, pp. 25–36. doi:10.1007/978-3-030-51310-8_3.

[2] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/V1/N19-1423.

[3] I. Beltagy, K. Lo, A. Cohan, SciBERT: A pretrained language model for scientific text, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. doi:10.18653/v1/D19-1371.

[4] S. Huang, J. M. Cole, BatteryBERT: A pretrained language model for battery database enhancement, Journal of Chemical Information and Modeling 62 (2022) 6365–6377. doi:10.1021/acs.jcim.2c00035.

[5] R. Srebrovic, J. Yonamine, Leveraging the BERT algorithm for patents with TensorFlow and BigQuery, 2020. Accessed: 2025-04-27.

[6] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for Named Entity Recognition, in: K. Knight, A. Nenkova, O. Rambow (Eds.), Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies, Association for Computational Linguistics, San Diego, California, 2016, pp. 260–270. doi:10.18653/v1/N16-1030.

[7] Z. Fu, Y. Su, Z. Meng, N. Collier, Biomedical Named Entity Recognition via Dictionary-based Synonym Generalization, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 14621–14635. doi:10.18653/v1/2023.emnlp-main.903.

[8] Z. Hu, W. Hou, X. Liu, Deep learning for Named Entity Recognition: A survey, Neural Computing and Applications 36 (2024) 8995–9022. doi:10.1007/s00521-024-09646-6.

[9] F. Saad, Named Entity Recognition for Biomedical Patent Text using Bi-LSTM variants, in: Proceedings of the 21st International Conference on Information Integration and Web-Based Applications & Services, iiWAS2019, Association for Computing Machinery, New York, NY, USA, 2020, p. 617–621. doi:10.1145/3366030.3366104.

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.

[11] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets, in: D. Demner-Fushman, K. B. Cohen, S. Ananiadou, J. Tsujii (Eds.), Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 58–65. doi:10.18653/v1/W19-5006.

[12] Z. R. K. Rostam, G. Kertész, Fine-tuning large language models for scientific text classification: A comparative study, in: 2024 IEEE 6th International Symposium on Logistics and Industrial Informatics (LINDI), 2024, pp. 000233–000238. doi:10.1109/LINDI63813.2024.10820432.

[13] M. M. Becker, I. Chaerony Siffa, H. Aras, VIVO-based plasma knowledge graph for improving the discoverability of patent information in plasma science and technology, in: Proceedings of the E-Science-Tage 2025, 2025. doi:10.11588/heidok.00036414, accessed: 2025-04-28.

[14] S. Clark, J. Friis, T. Vegge, BattINFO: The ontology for the battery interface genome - materials acceleration platform (BIG-MAP), in: Proceedings of the 3rd EMMC International Workshop, 2021. Accessed: 2025-04-28.

[15] S. Franke, L. Paulet, J. Schäfer, D. O'Connell, M. M. Becker, Plasma-MDS, a metadata schema for plasma science with examples from plasma technology, Scientific Data 7, 439 (2020). doi:10.1038/s41597-020-00771-0.

[16] I. Chaerony Siffa, R. Wagner, M. M. Becker, Semantic information management in Low-Temperature Plasma Science and Technology with VIVO, J. Phys. D: Appl. Phys 58 (2025) 235204. doi:10.1088/1361-6463/add710.