The 39th Annual Conference of the Japanese Society for Artificial Intelligence, 2025

Information Extraction of ORR Catalyst for Fuel Cell from Scientific Literature

Hein Htet[*1] Amgad Ahmed Ali Ibrahim[*1] Yutaka Sasaki[*2] Ryoji Asahi[*1]

[*1] Institute of Innovation for Future Society, Nagoya University, Nagoya, Japan
[*2] Computational Intelligence Laboratory, Toyota Technological Institute, Nagoya, Japan

The development of advanced catalysts for the Oxygen Reduction Reaction (ORR) is critical for improving the performance and efficiency of Polymer Electrolyte Fuel Cells (PEFCs). However, the vast and growing body of scientific literature poses challenges for researchers aiming to identify key insights. This study focuses on the information extraction of ORR catalysts from fuel cell-related literature using a hybrid approach combining manual annotation and automated machine learning techniques. A comprehensive dataset was constructed through the Brat annotation tool, identifying 12 critical entities such as *catalyst*, *support*, and *value*, alongside two relationship types: *equivalent* and *related_to*. The annotated data was used to fine-tune the DyGIE++ framework with the pre-trained BERT models. The model demonstrated effective performance in extracting complex material science concepts and their interrelationships. The finding suggests that this automated framework can accelerate catalyst discovery by providing structured, high-quality data for downstream analysis. This research highlights the potential of Natural Language Processing (NLP) in enabling efficient literature mining and fostering advancements in clean energy techniques.

1. Introduction

The oxygen reduction reaction (ORR) plays a crucial role in the performance of fuel cells, as it directly impacts efficiency and durability. ORR catalysts, particularly those based on platinum-group metals and their alternatives, have been extensively studied to enhance catalytic activity and stability. Given the rapid growth of fuel cell research, a vast amount of scientific literature has been published, presenting valuable insights into catalyst compositions, structures, and performance metrics. However, manually extracting and analyzing this information from numerous articles is time-consuming and inefficient.

One of the main challenges in extracting ORR catalyst information is the complexity and diversity of scientific text. Research papers contain unstructured data, including chemical formulas, experimental conditions, and performance results, which are often scattered across different sections. Additionally, critical information is frequently embedded in tables, figures, and mathematical expressions, making automated extraction even more difficult. Named Entity Recognition (NER) and Relation Extraction (RE) techniques, powered by Natural Language Processing (NLP), provide a promising solution by identifying key entities and their relationships within texts [Yamaguchi 22, Mitsui 23].

In this study, we develop a web-based system that integrates data collection, annotation, and extraction features, as shown in Figure 1, for ORR catalyst information extraction. Specifically, we apply DyGIE++ [Luan 19], a deep learning framework designed for joint NER and RE tasks. Our approach structures the extracted data efficiently, enabling further analysis in catalyst research. Experimental results demonstrate the effectiveness of our methodology in extracting catalyst-related entities and relations, contributing to more accessible and automated knowledge retrieval in fuel cell research.

Contact: Ryoji Asahi, Institutes of Innovation for Future Society, Nagoya University, Green Mobility Collab. Res. Center, 114, Furo-cho, Chikusa-ku, Nagoya, Aichi 464-8063, Japan, 052-747-6869/6736, and asahi.ryoji.d9@f.mail.nagoya-u.ac.jp

2. Proposed Method

In this section, we describe the proposed method for collecting, annotating, integrating, and modeling data related to ORR catalysts for fuel cells.

2.1 Data Collection

This study accessed literature from the *Royal Society of Chemistry (RSC)* through institutional access provided by Nagoya University and followed the applicable terms of use for text mining. We collected full-text articles through a developed web-based platform, which accessed RSC database [Antony 14] from 2010 to 2024. To retrieve relevant articles, we used the following query: *ORR AND Catalyst AND (ECSA OR "mass activity" OR "ORR activity" OR "surface activity")*. Here, ECSA represents Electro Chemical Surface Area.

This search identified a total of 1,259 articles. From these, we focused only on *three* key sections: Abstract, Results & Discussion, and Conclusions, as they contain the most essential information. In this study, the articles were then ranked based on the highest occurrences of mass activity and ORR activity. Through this process, 76 articles were selected as representative studies in ORR catalyst research

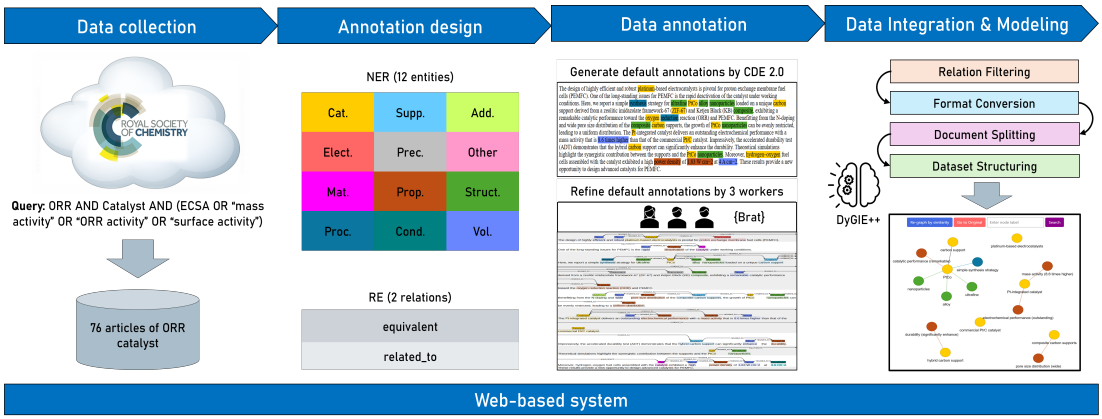

Figure 1: System overview: web-based ORR catalyst data collection and analysis.

for fuel cells and subsequently used for data annotation.

2.2 Annotation Design

For the annotation process, we identified 12 entities grouped into *three* main categories, reflecting the typical components involved in ORR catalyst research.

- **Materials**: This category includes *seven* material types: 1. Catalyst (Cat.), 2. Support (Supp.), 3. Additive (Add.), 4. Electrolyte (Elect.), 5. Precursors (Prec.), 6. Other Material (Other), 7. Material Reference (Mat.).
- **Material Characteristics**: Covering *three* aspects: 8. Property (Prop.), 9. Structure (Struct.), 10. Process (Proc.).
- **Experimental Parameters** Encompassing *two* key parameters: 11. Condition (Cond.), 12. Value (Val.).

In addition, we identified *two* types of relationships based on interactions between these entities:

(1) *equivalent* — links entities that represent the same concept or material

(2) *related_to* — captures connections between entities that share a significant association

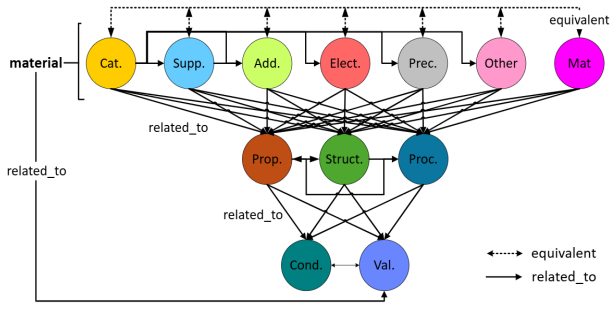

Figure 2: Entity-Relation map.

2.3 Data Annotation

From the 76 articles, *three* key sections— *Abstract*, *Results & Discussion*, and *Conclusions*—were extracted and processed following the defined 12 entities and 2 relations, as shown in Figure 2.

(1) **Automated Pre-Annotation**: To expedite the annotation process, default annotations in Brat format are automatically generated using a combination of *Chemical Data Extractor (CDE)'s parser* [Mavračić 21] and *custom-created parsers*. It assists in identifying key entities and their relations, reducing the manual workload.

(2) **Refine Annotation**: The *three* workers then review, refine, and validate the pre-annotations using the Brat annotation tool [Stenetorp 12]. This tool is hosted on a centralized web-based platform, eliminating the need for local installation. Annotators can seamlessly access and modify annotations through the platform, ensuring consistency and accuracy in data labeling.platform.

2.4 Data Integration & Modeling

The data integration and modeling process involved converting the annotated text into a structured machine-readable format suitable for model training. The steps followed include:

(1) **Relation Filtering**: Invalid cross-relations were identified and removed to ensure consistency and accuracy in tokenization.

(2) **Format Conversion**: Brat annotations were converted into a JSON format, ensuring compatibility with the DyGIE++ framework. This conversion facilitates seamless integration into the model pipeline.

(3) **Document Splitting**: To prevent CUDA out-of-memory errors, large documents were split into smaller segments. This ensures the model can process each document efficiently without exceeding memory limits during training.

(4) **Dataset Structuring**: The annotated and pre-processed data was organized into distinct training, validation, and test sets. These sets were then used to evaluate the performance of the model.

(5) **Modeling**: Several domain-specific pre-trained BERT-based models, such as SciBERT [Beltagy 19], and MatSciBERT [Gupta 22], were fine-tuned on our annotated dataset. Finally, the models' performance were evaluated using precision, recall, and F1-score metrics on the test sets.

$$\text{Precision} = \frac{TP}{TP + FP}, \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

where:

- TP = Number of true positives (correct annotations/extractions)
- FP = Number of false positives (incorrect annotations/extractions)
- FN = Number of false negatives (missed annotations/extractions)

3. Data Extraction

Data extraction involved applying trained models to identify and retrieve relevant information from fuel cell-related scientific literature. The extraction process was facilitated through the developed web platform, which provides an intuitive interface for users. As shown in Figure 3, the user simply needs to select the trained model of choice, then input the text or upload an article file. Once the article was provided, the model performed the extraction process, identifying key entities and relationships.

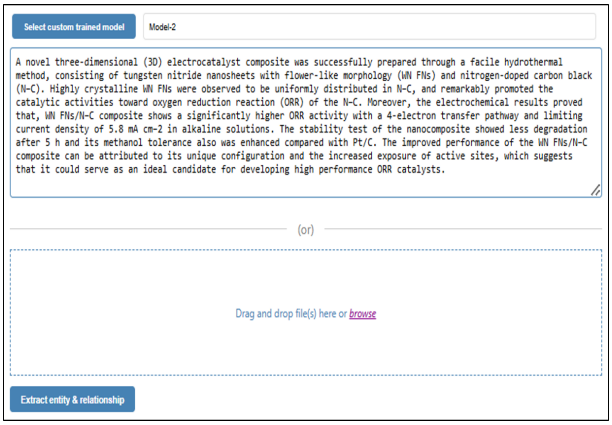

Figure 3: Model selection & input data for extraction.

The extracted results were presented in *two* formats for user convenience: (1) Brat Visualization: (See Figure 4), and (2) Graph Visualization:(See Figure 5).

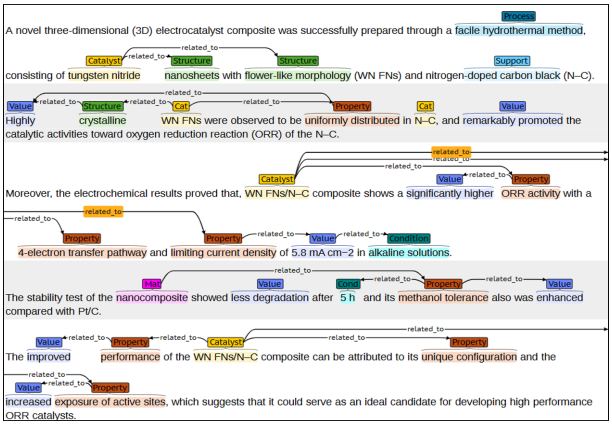

Figure 4: Extracted data in Brat visualization.

4. Evaluations

In this section, we present the performance evaluation of our trained models on ORR catalyst-related scientific literature. A total of 76 articles were annotated, resulting in 554

Figure 5: Extracted data in graph visualization.

documents in the fuel-cell dataset. The dataset was split into training (80%), validation (10%), and testing (10%) sets. In addition to the standard *test* set, we included a *Gold Standard (Gold Std.)* dataset, which was created by three experts in our group as a benchmark dataset.

4.1 Performance of NER and RE Models

We fine-tuned our fuel-cell dataset on *seven* different pre-trained BERT-based models, generating *seven* specialized models: Model-1 (SciBERT), Model-2 (MatSciBERT-1), Model-3 (MatSciBERT-2), Model-4 (MatSciBERT-3), Model-5 (PubMedBERT), Model-6 (BlueBERT), and Model-7 (BioBERT).

As shown in Figure 6, Model-1 and Model-5 achieved the highest NER F1-score of 82.19%, while Model-2 obtained the highest RE F1-score of 66.10% on the Gold Standard test set. Overall, the NER F1-score ranged from 61.66% to 82.19%, while the RE F1-score ranged from 51.27% to 66.10% across the valid, test, and Gold Standard sets.

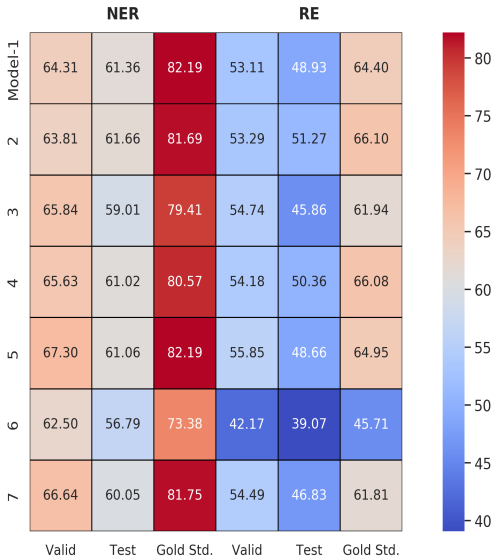

Figure 6: F1-score of valid, test, and gold standard.

4.2 Annotators vs. Model Performance

To establish a benchmark for evaluating model performance, we assessed human annotators using standard metrics, as described in Section 2.4. This evaluation provided a reliable reference for comparison.

For a fair evaluation, we compared both human annotators and model-generated extractions using the *Gold Standard* dataset. This dataset comprises expert-verified annotations, making it the most dependable resource for assessing extraction accuracy between annotators and models.

In the *annotator evaluation*, *three* annotators were given the same article used to construct the Gold Standard dataset, and their performance was measured based on the defined metrics.

For the *model evaluation*, we selected the *three* highest-performing trained models and applied them to the same article. Their performance was assessed using the DyGIE++ evaluation metrics.

By ensuring identical evaluation conditions, we conducted a direct and objective comparison between human annotators and models. Figure 7 presented the NER and RE F1-scores of both annotators and models. The difference in NER performance between annotators and models was minimal, demonstrating strong model reliability. However, for RE, the models exhibited a significant decline in F1-score compared to annotators, indicating that relation extraction remains a more challenging task for models.

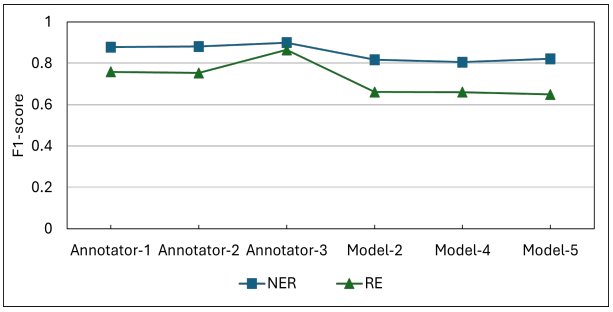

Figure 7: F1-score of annotators and models

5. Conclusion

This study presents a hybrid approach that combines manual annotation with automated machine learning techniques to extract structured information on ORR catalysts from scientific literature. By automating literature mining, researchers can efficiently identify promising materials and synthesis techniques, significantly reducing the time and effort required for manual analysis. This acceleration has the potential to drive faster innovation and facilitate the development of more efficient and sustainable ORR catalysts for fuel cells, contributing to advancements in clean energy technologies.

Our models demonstrated effective performance in fuel-cell literature extraction, particularly for NER. However, improvements are still needed for RE to achieve higher accuracy and reliability.

As future work, we aim to enhance our model's performance, especially in relation extraction, and develop an essential model for ORR catalyst development based on the extracted knowledge.

References

[Stenetorp 12] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, Jun'ichi Tsujii: BRAT: a web-based tool for NLP-assisted text annotation, In Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12), pp. 102–107, Association for Computational Linguistics, (2012).

[Antony 14] Williams Antony, Tkachenko Valery: The royal society of chemistry and the delivery of chemistry data repositories for the community, Journal of Computer - Aided Molecular Design, Vol. 28:1023-1030, Dordrecht, (2014).

[Luan 19] Yi Luan, Dave Wadden, Luheng He, Amy Shah, Mari Ostendorf, Hannaneh Hajishirzi: A general framework for information extraction using dynamic span graphs, Proceedings of NAACL-HLT, pp. 3036–3046, Association for Computational Linguistics, (2019).

[Beltagy 19] Iz Beltagy, Kyle Lo, Arman Cohan: SciBERT: a pretrained language model for scientific text, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620, Association for Computational Linguistics, (2019).

[Mavračić 21] Juraj Mavračić, Callum J. Court, Taketomo Isazawa, Stephen R. Elliott, Jacqueline M. Cole: ChemDataExtractor 2.0: autopopulated ontologies for materials science, Journal of Chemical Information and Modeling, Vol. 61:4280-4289, American Chemical Society, (2021).

[Yamaguchi 22] Kyosuke Yamaguchi, Ryoji Asahi, Yutaka Sasaki: Superconductivity information extraction from the literature: a new corpus and its evaluations, Advanced Engineering Informatics, Vol. 544:101768, Elsevier, (2022).

[Gupta 22] Tanishq Gupta, Mohd Zaki, N. M. Anoop Krishnan, Mausam: MatSciBERT: a materials domain language model for text mining and information extraction, npj Computational Materials, Vol. 8, (2022).

[Mitsui 23] Kento Mitsui, Yutaka Sasaki, Ryoji Asahi: Automatic knowledge acquisition from superconductivity information in literature, Science and Technology of Advanced Materials: Methods, Vol. 3:2206532, Taylor & Francis, (2023).