Enhancing Scientific Relation Extraction: A Comparative Study of Machine Learning and Transformer-Based Models on the SciERC Dataset

Sirad Bihi (14151162)
Zhiying Guo (14158989)
Altynai Mambetova (11156561)
Neha Thomas (14159311)

Abstract

Scientific Relation Extraction is essential for structuring knowledge from research literature, yet challenges like long-tail relation distributions and domain-specific terminology hinder accurate extraction. To address this problem, we explore two distinct approaches: XG-RelSci, a machine learning-based pipeline that integrates XGBoost with hand-crafted linguistic and structural features, and SciRelBERT, a Transformer-based model fine-tuned on scientific corpora for entity relationship classification. Extensive experiments on the SciERC dataset demonstrate the superior performance of SciRelBERT in capturing scientific entities, relations, and coreference clusters without manual feature engineering.

1 Introduction

Relation extraction, a sub-task of Information Extraction, has evolved significantly since its emergence in the 1980s. Early methods relied on rule-based systems and manually crafted features, which were later replaced by machine learning approaches. The advent of deep learning, particularly transformer-based models, has further advanced the field, achieving state-of-the-art performance in accuracy and efficiency (Detroja et al., 2023).

While some contemporaneous studies attempt relation extraction by integrating entity recognition, relation extraction, or coreference resolution selectively, most of them lack a unified solution. (Eberts and Ulges, 2019). One of the notable approach by Singh et al. (2013) applies probabilistic graphical models for joint extraction but requires manual feature engineering . Most recent research favors transformer based deep learning models on scientific corpora (e.g. SciBERT) , which has demonstrated superior performance over prior methods (Eberts and Ulges, 2019; Inayah et al., 2023; Beltagy et al., 2019).

In this paper, we compare two approaches for Scientific Relation Extraction: (1) XG-RelSci, a machine learning pipeline using XGBoost with hand-crafted linguistic and structural features, and (2) SciRelBERT, a transformer-based model fine-tuned on scientific corpora for relation classification. We conduct a comparative evaluation on the SciERC dataset, analyzing performance and computational efficiency. Results show that SciRelBERT outperforms feature-based methods in detecting scientific entities, relations, and coreference clusters. Our findings highlight the strengths and trade-offs of each approach, offering insights for future research in scientific information extraction.

1.1 Dataset

The SciERC dataset used in this study consists of 500 scientific abstracts sourced from 12 AI conference and workshop proceedings across four AI research communities within the Semantic Scholar Corpus. It is designed for scientific relation extraction, with structured annotations for entities, relations and coreference links. Originally developed and published in *Multi-Task Identification of Entities, Relations, and Coreference for Scientific Knowledge Graph Construction*, SciERC is publicly available at http://nlp.cs.washington.edu/sciIE/.

1.2 Baseline Model

The SciIE was selected as a baseline model, which employs an end-to-end BiLSTM-based multi-task learning framework. It processes input at the document level, extracts all possible spans (up to 8 words), and encodes them with a one-layer BiLSTM (200-dimensional hidden states). Classification is performed using Feed Forward Neural Networks with two 150-dimensional hidden layers and dropout regularization to improve generalization. The model applies beam pruning to optimize efficiency. Outputs include entity types, relation labels, and coreference clusters (Luan et al., 2018).

Experiments demonstrate that SciIE outperforms traditional pipeline-based methods by leveraging cross-sentence relations through coreference resolution, achieving development and test F1 scores of 68.1% and 64.2% for entity recognition, 39.5% and 39.3% for relation extraction, and 58.0% and 48.2% for coreference resolution, respectively. However, it does not utilize explicit feature engineering or specialized classification enhancements like Focal Loss or self-attention-based classifiers, which are integral to our approach.

2 Model

2.1 XG-RelSci: Feature-driven Machine Learning Model

Our first approach adopts a feature-driven machine learning (ML) framework for scientific relation extraction, leveraging explicit linguistic features to enhance model interpretability and performance.

This approach was selected for several reasons. Firstly, feature engineering is ideal for small datasets, with structured annotations (entities, relations, coreference resolutions) that can be directly encoded into features, while deep learning model may require large datasets to generalize well. Secondly, hand-crafted features improve interpretability, with feature importance analysis providing insights for model refinement. Thirdly, after exploring three different Machine Learning models (i.e., SVM, Random Forest and XGBoost), XGBoost was chosen as the best model for its ability to handle structured and sparse data. It offers computational advantage through parallel processing, tree pruning, and built-in regularization, ensuring robust and scalable relation extraction.

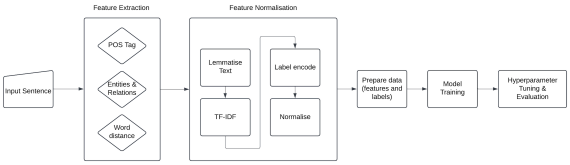

Figure 1: Feature-driven Machine Learning Relation Extraction

As shown in Figure 1, The implementation follows a structured pipeline, beginning with feature extraction to capture key linguistic and relational characteristics. Lexical features (e.g., entity types, text between entities) capture surface-level information essential for identifying relationships between scientific terms. Syntactic features (e.g., dependency paths, POS tags, and word distances) help model structural relationships by capturing grammatical connections between entities. Semantic features, such as word embeddings and span similarity, are used to capture meaning and contextual relevance(spaCy, 2025).

To ensure optimal model performance, categorical features are label-encoded for compatibility, while numerical features are normalized using StandardScaler. TF-IDF features are extracted from text between entities to capture term relevance and incorporated into the dataset. However, we limited vectorisation to the top 100 features to avoid introducing more noise and possible over-fitting.

For consistency, an 80/20 dataset split is applied across all models. The multi-class log loss (mlogloss) function is used as the objective function, with softmax probability outputs facilitating multi-class classification. To optimize model performance, hyperparameter tuning was conducted.

2.2 SciRelBERT: Transformer-Based Deep Learning Model

For our second relation extraction approach, we fine-tune a pretrained BERT-base model on the SciERC dataset. This model was chosen for several reasons. Firstly, unlike traditional sequential models, BERT processes all tokens in parallel using self-attention, enabling it to capture long-range dependencies. This capability is particularly beneficial for scientific relation extraction, where semantically related words often appear far apart within a document. Secondly, it enhances entity-aware encoding through marker-based input formatting, ensuring that entity interactions are contextually well-represented. Focal Loss is integrated to mitigate class imbalance, while stratified sampling and gradient accumulation contribute to training stability and improved generalization. Finally, BERT-based models outperform traditional machine learning approaches by learning rich contextual representations directly from text, eliminating the need for manual feature engineering while offering superior adaptability to complex language structures.

A significant challenge in our task was addressing class imbalance; notably, the 'Used-For' and 'Coreferences' relations account for 38.5% and 25% of the dataset, respectively. To mitigate this, we designed a custom metric function that includes macro-averegaed F1 scores rather than accuracy,

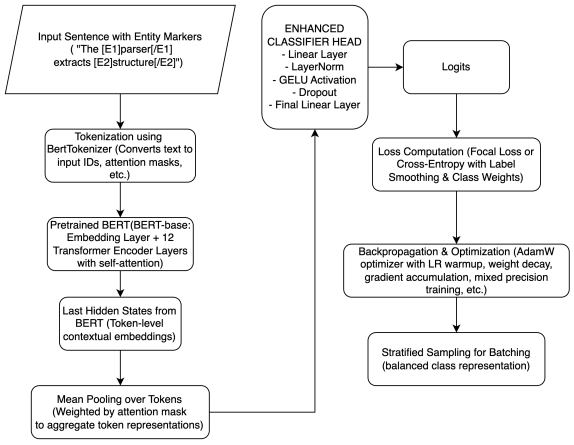

Figure 2: SciRelBERT Model Relation Extraction

ensuring that all classes are equally represented in the evaluation regardless of the frequency.

To further address class imbalance, we used Focal Loss function within our pipeline as shown in Figure 1. The approach modifies the standard cross-entropy loss by down-weighting well-classified examples. The method was introduced in 2017 by Lin et al. and has been extended to NLP techniques with BERT Ostendorff et al.. The Focal Loss function is then wrapped within an additional classifier head (*EnhancedBertClassifier*) to improve upon the model's standard classification layer (Michael Sugimura, 2019).

Within that classifier head, we also define a mean pooling function to use instead of the default CLS pooling strategy. BERT can use either pooling strategies when constructing sentence embeddings from contextualised word embeddings: CLS uses just the first token in a sentence, while mean pooling uses the average of all words (Tsukagoshi et al., 2021), capturing more meaningful information much like span representation in SciIE.

To ensure that each training batch contains a balanced representation of relation classes, we implemented a custom stratified sampler. This sampler groups examples by their labels and constructs balanced mini-batches, thereby ensuring that the model is regularly exposed to underrepresented classes. We integrate this sampler into our training loop using a custom Trainer subclass, which overrides the default data loading mechanism.

Finally, our training process leverages and extends the hyperparameters of Xu et al. with adjustments. We use the same low learning rate of 2e-5, along with a warmup ratio of 0.1, weight decay of 0.01, and gradient accumulation to effectively

Machine Learning Models Comparison		
Model	Accuracy	F1 Score (weighted avg.)
Random Forest	0.7128	0.6839
SVM	0.6649	0.6124
XG-RelSci	0.7388	0.7043

Table 1: XG-RelSci, Random Forest and SVM performance comparison

increase the batch size. Mixed-precision training (fp16) is enabled to accelerate computations and reduce memory usage. These hyperparameter choices facilitated efficient fine-tuning while achieving higher macro F1 scores and better overall performance on the SciERC dataset.

3 Experimental Results

To ensure fair model evaluation, we adopt the weighted average F1 score as the primary metric. By incorporating both precision and recall, the F1 score ensures a fair assessment of how well the model captures rare but important relations in scientific text. Unlike macro F1, the weighted F1 score accounts for class imbalance, which prevents majority classes from skewing the evaluation and provides a more comprehensive assessment across all relation types.

3.1 Evaluation

For the machine learning approach, we explored three models: SVM, Random Forest, and XG-RelSci. As shown in Table 1, XG-RelSci achieved the best overall performance, leading to its selection as the final model. Its gradient boosting mechanism iteratively refines predictions by focusing on harder to classify samples, enhancing classification accuracy. While Random Forest is robust and effective for structured data, it lacks this iterative refinement, which limits its performance improvement. SVM, though powerful in high-dimensional spaces, relies on a single global decision boundary, making it less adaptable to the nuanced linguistic and structural variations in scientific relation extraction(Nattapoj Apichardsilkij, 2024).

SciRelBERT did not yield good results until after introducing fine-tuning measures such as a custom classifier head with a focal loss option, mean pooling, and stratified batch samples. As outlined in Table 3, evaluation results improve significantly from the outset of training until we reach a F1 weighted average score of 0.861, F1 macro, precision and recall macro scores of 0.791, 0.792 and 0.794 re-

SciRelBERT Evaluation	
Metric	Result
F1 Macro	0.791
Precision Macro	0.792
Recall Macro	0.794
F1 Weighted	0.861

Table 2: Evaluation Results for SciRelBERT relation extraction model

Relation Extraction Performance Comparison	
Model	Average F1
SciIE Multiask	39.5
SciIE Single Task Relation Extraction	37.9
XG-RelSci	0.70
SciRelBERT	0.86

Table 3: Comparison of benchmark model and study models on relation extraction task

spectively. The model performs very well on relation classes 'coreference' and 'conjunction, while struggling with 'part-of' and 'feature-of'. Moreover, validation loss gradually decreases until 2000 steps and then slowly increases, possibly suggesting over-fitting (OpenAI, 2025). Stopping training at this point would still result in a F1 weighted average score of 0.835.

3.2 Comparison

SciRelBERT achieves the highest F1 score (0.86), demonstrating its strong ability to capture complex relationships, particularly across sentences. The transformer-based self-attention mechanism enables it to model long-range dependencies, making it highly effective for scientific text. Additionally, the integration of additional strategies improves recall for under-represented relation types by addressing class imbalance. However, BERT-based models require significant computational resources and benefit most from large-scale data, making it less suitable for resource-constrained environments.

XG-RelSci achieves an F1 score of 0.70. While hand-crafted features require manual effort and domain expertise, they provide structured linguistic information that enhances model performance, particularly in small datasets. This model is computationally efficient, interpretable, and well-suited for sentence-level relation extraction. However, it struggles with capturing cross-sentence dependencies, which reduces its effectiveness for more complex relation structures.

SciIE (F1: 0.395 multitask, 0.379 single-task) underperforms compared to both XG-RelSci and SciRelBERT. As an end-to-end model, SciIE eliminates the need for feature engineering, making it more time-efficient. However, its reliance solely on deep neural representations limits interpretability and reduces its ability to capture nuanced linguistic and structural relationships, which XG-RelSci effectively models through hand-crafted features. Furthermore, unlike SciRelBERT, which leverages self-attention to capture long-range dependencies, SciIE relies on BiLSTMs that process tokens sequentially, limiting efficiency and scalability. While SciIE is computationally lightweight and does not require large-scale training data, its inability to effectively model contextual dependencies across sentences constrains its overall performance.

4 Conclusion

This study compares two approaches to scientific relation extraction: the feature-driven XG-RelSci and the transformer-based SciRelBERT, using SciIE as the baseline.

XG-RelSci outperforms SciIE by leveraging hand-crafted linguistic and structural features, enhancing interpretability and efficiency for small datasets. However, its reliance on manual feature engineering makes it time-consuming and limits scalability.

SciRelBERT surpasses both models by employing self-attention to capture long-range dependencies without requiring feature engineering. The integration of Focal Loss further improves recall for underrepresented relations. Despite its superior performance, its high computational cost remains a challenge.

Futher consideration could be leveraging pretrained SciBERT to enhance both performance and efficiency.

5 Use of Generative AI Tools

The authors acknowledge the use of ChatGPT-4o (OpenAI, 2025) for refining the codebase and clarifying methodological implementations, including executing BERT-based relation extraction and interpreting results. Direct outputs are quoted, while paraphrased content is appropriately cited.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Conference on Empirical Methods in Natural Language Processing*.

Kartik Detroja, C.K. Bhensdadia, and Brijesh S. Bhatt. 2023. A survey on relation extraction. *Intelligent Systems with Applications*, 19:200244.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Nur Inayah, Muhaza Liebenlito, Nina Fitriyati, Muhammad Manaqib, and Nabila Aryanti. 2023. Relation classification in scientific article abstracts using scibert with entity marker. *2023 11th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–5.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2018. Focal loss for dense object detection.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Michael Sugimura. 2019. BERT Classifier: Just Another Pytorch Model. Accessed: 2025-03-04.

Nattapoj Apichardsilkij. 2024. Basic Comparison Between RandomForest, SVM, and XGBoost. Accessed: 2025-03-04.

OpenAI. 2025. ChatGPT: A Large Language Model. Accessed: 2025-03-03.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching bert with knowledge graph embeddings for document classification.

Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. 2013. Joint inference of entities, relations, and coreference. In *Conference on Automated Knowledge Base Construction*.

spaCy. 2025. Language processing pipelines. Accessed: 2025-03-01.

Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2021. Defsent: Sentence embeddings using definition sentences.

Yige Xu, Xipeng Qiu, Ligao Zhou, and Xuanjing Huang. 2020. Improving bert fine-tuning via self-ensemble and self-distillation.