University of Bonn
CAISA lab

Natural Language Processing Lab

Summer Semester 2025

Project Proposal of Team # 3

SciREX: Scientific Relation Extraction

Team Members:

name: Zyad Altahan
name: Sayantak Karar
name: Sulaeman Aloradi
name: Abdelwahab Elshennawy

June 24, 2025

1 Research Topic

As the volume of scientific literature continues to grow exponentially, researchers face increasing difficulty in staying informed about emerging findings and complex interconnections across disciplines. This project explores the application of large language models (LLMs) for automated scientific information extraction, specifically relation extraction within biomedical texts. Our primary goal is to systematically evaluate and compare various relation extraction techniques using the BioRED dataset, a challenging benchmark with rich entity and relation annotations in biomedical research. By leveraging both traditional supervised approaches and newer instruction-tuned or generation-based LLMs, we aim to assess how well these models can extract structured scientific knowledge, especially in low-resource and semantically complex settings. This study contributes to the broader goal of enabling scalable, automated curation of scientific knowledge and improving downstream tasks such as biomedical search, question answering, and hypothesis generation.

2 Motivation and Objectives

The scientific community is inundated with an overwhelming volume of published literature, particularly in fields like biomedicine, where new discoveries are made at an accelerating pace. As a result, manually identifying and organizing key entities (e.g., proteins, diseases, drugs) and the relationships between them has become an increasingly impractical task. Automating this process through Natural Language Processing (NLP) techniques, such as named entity recognition (NER) and relation extraction (RE), offers a scalable solution to synthesize knowledge from vast textual corpora.

Relation extraction and entity recognition are not only critical for scientific information retrieval but are also highly sought-after capabilities across a wide range of modern industries. From biomedical research and pharmaceuticals to finance, law, and environmental science, organizations increasingly rely on robust information extraction pipelines to support decision-making, automate workflows, and uncover hidden insights from unstructured data.

The motivation for this project lies in both a scientific and practical gap. While a variety of techniques for relation extraction exist including fine-tuned language models, span-based classification, and generative approaches there is limited comparative research that evaluates these methods systematically on a well-structured biomedical dataset like BioRED. This project aims to bridge that gap.

Our specific objectives are as follows:

- To identify and implement a set of state-of-the-art relation extraction techniques—including classification-based, question answering formulation, and generative methods—tailored for scientific text.
- To turn the BioRED dataset into a benchmark evaluation setup that enables consistent comparison across methods.
- To explore the performance differences of these techniques in terms of both quantitative accuracy and qualitative behavior on domain-specific texts.
- To assess the generalizability of high-performing methods by testing them on secondary datasets from other domains or with differing linguistic complexity.

- To contribute practical insights that could inform the design of future scientific knowledge extraction pipelines, particularly in low-resource or high-stakes environments.

By addressing these objectives, the project not only contributes to the academic understanding of model capabilities in biomedical NLP but also advances the practical utility of LLMs in real-world scientific workflows.

3 Dataset (Data Exploration)

The BioRED dataset [1] consists of 600 PubMed abstracts, each annotated with entities and relations. The key statistics about these abstracts:

3.1 Entity (Annotation) Landscape

- **Gene / Gene Product** and **Disease / Phenotype**
  - Together make up approximately 60% of all mentions.
- **Train Split**
  - Contains roughly 4× as many mentions of every type.
  - Mirrors the raw document ratio.
- **SequenceVariant** and **CellLine**
  - Are comparatively rare.
  - Account for at most 7% and 3% of all mentions, respectively.

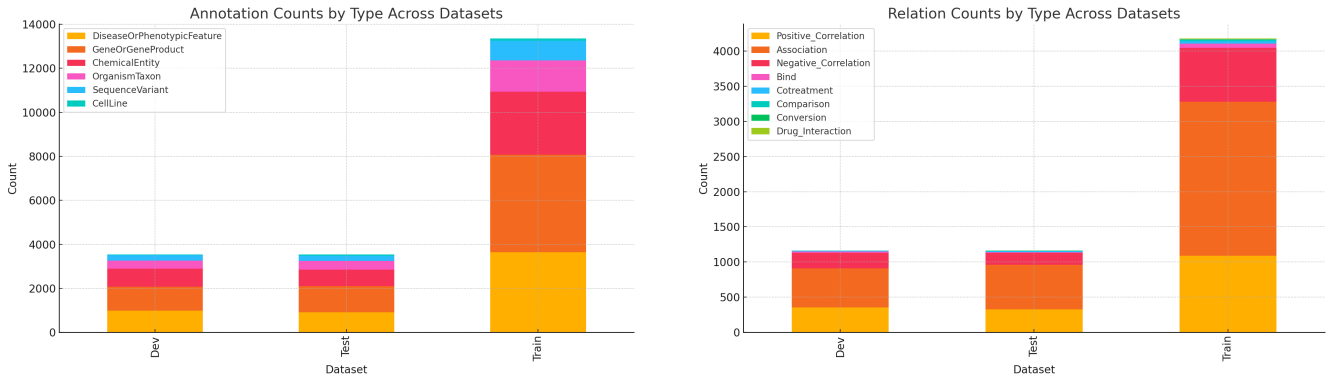
3.2 Relation Landscape

- **Association**
  - Most common relation type.
  - Followed by **Positive-Correlation** and **Negative-Correlation**.
- **Less Frequent Relations**
  - Include **Bind**, **Cotreatment**, **Drug Interaction**, and **Conversion**.
  - Sparse but important for evaluating fine-grained relation extraction.
- **Dev vs Test**
  - Relative proportions of relation types are very stable.
  - Suggests the splits are well-balanced.

3.3 Implications for Downstream Modelling

- **Class Imbalance**
  - Rare entity types (e.g., **CellLine**, **Variant**) and relation types.
  - May need sampling strategies, loss re-weighting, or data augmentation.
- **Plenty of Training Signal**
  - The Train split is large enough.
  - Enables learning robust mention-level representations even with simple neural models.
- **Stable Evaluation**
  - Similar distributions in Dev and Test.
  - Dev can be used reliably for model selection.

As shown in Figure 1, the annotation and relation distributions vary across splits.



(a) Annotation Counts

(b) Relation Counts

Figure 1: Entity and relation distributions across datasets.

4 Methodology

This project investigates the application of large language models and advanced NLP techniques for scientific information extraction in biomedical texts, with a focused emphasis on Relation Extraction (RE). Our methodology centers on evaluating a diverse set of RE approaches—ranging from classification based models to generative capable of identifying semantic relationships between biomedical entities. By leveraging the BioRED dataset, we aim to systematically compare the effectiveness, efficiency, and generalizability of these RE methods in both structured and semantically complex contexts.

We explore three complementary RE techniques, each with varying methodological assumptions and architectural styles.

4.1 BioBERT + Classification Head (Used for NER and RE)

BioBERT [2] is a BERT-based model pretrained on large biomedical corpora such as PubMed abstracts and PMC articles. This method can be used for NER (Named Entity Recognition), which can be required as a pre-process for some other RE methods.
As for relation extraction, a common approach is to fine-tune BioBERT with a classification head that takes the contextualized representation of the sentence (often using special entity markers or the [CLS] token) and predicts a relation label between a pair of entities.

Ease of Implementation: High. Fine-tuning BioBERT using HuggingFace's `AutoModelForSequenceClassification` interface requires minimal effort. Datasets like BioRED can be adapted to this format with light preprocessing.

Performance: BioBERT significantly outperforms general-domain BERT on biomedical RE tasks, with up to 2.8% F1 score improvement, owing to its domain-specific pretraining [2].

4.2 SciFive (T5 for Biomedical RE)

SciFive [3] is a text-to-text transformer based on T5 and pretrained on biomedical corpora. It reformulates relation extraction as a sequence generation task: given a sentence and entity pair, the model generates the appropriate relation label.

Ease of Implementation: Medium. This method requires reformatting inputs and outputs into a text-to-text paradigm. However, using HuggingFace's `T5ForConditionalGeneration` makes integration relatively smooth.

Performance: SciFive achieves state-of-the-art results on multiple biomedical RE benchmarks, outperforming both BioBERT and base T5 on tasks such as CHEMPROT and DDI with substantial F1 score improvements [3].

4.3 QA4RE (Question Answering for Relation Extraction)

QA4RE [4] is a question-answering-based framework that reformulates relation extraction as a QA task. It leverages pre-trained language models to generate and answer natural language questions about relations between entities in text, enabling effective extraction of relational information.

Ease of Implementation: Medium. The official codebase is typically built on transformer-based frameworks like HuggingFace and requires configuring question templates and fine-tuning for specific domains. It supports adaptation to various datasets with appropriate question design.

Performance: QA4RE achieves competitive F1 scores on relation extraction tasks across multiple datasets (e.g., 50.2 on SciERC, 67.1 on WLPC), particularly effective in scenarios requiring precise relation identification and contextual understanding [4].

4.4 Universal Information Extraction (UIE) Methods (Open for Research)

The Universal Information Extraction (UIE) framework aims to unify diverse information extraction tasks into a single, cohesive model, offering significant potential for advancing natural language processing. However, several challenges and opportunities remain open for exploration.

InstructUIE

InstructUIE [5]InstructUIE's approach of fine-tuning large language models with expert-written instructions enables robust multi-task learning. Automating the generation of high-quality, task-specific instructions could reduce the dependency on expert input, improving scalability and accessibility for diverse IE tasks. [5]

Code4UIE

Code4UIE [6]Code4UIE's use of retrieval-augmented code generation offers a structured and precise approach to UIE, yet several research opportunities remain. Developing dynamic schema generation methods that adapt Python class definitions to varying data structures could improve flexibility across heterogeneous datasets. [6]

In summary, our methodology spans simple classification-based models to LLMS' instruction tuning alignment and generative transformers. This allows us to benchmark trade-offs in model complexity, resource usage, and generalizability across tasks.
Priorities: 1. Prepare the data BIORED for later processing by other methods: (Should we use BioBERT NER or directly the NERs from the BIORED dataset?) 2. Use BioBERT to generate the NERs to be used by the other methods for classification (Or use the ones in BIORED) 3. Work on the other methods in parallel to generate Classifications (BioBERT, SciFive, QA4RE) 4. (less priority) Implement the UIE methods (InstructUIE, Code4UIE)

Question: Does the BIORED dataset already have NERs ready for performing the classification by other methods?

5 Expected Results

Through this project, we expect to gain empirical insights into the strengths and limitations of various relation extraction (RE) techniques when applied to biomedical text. Specifically, we anticipate the following results:

- **Baseline Performance**: Classification-based models such as BioBERT with a relation classifier head are expected to perform well out-of-the-box, establishing a strong supervised baseline. These models tend to work reliably for sentence-level relations with clearly defined entities and can serve as a reference point for comparing more advanced approaches.
- **Generative Models**: SciFive, with its sequence-to-sequence formulation, is expected to perform competitively or even surpass classification methods, particularly in complex or ambiguous sentences. Due to its ability to generate context-aware labels, we expect it to handle soft or implicit relations more gracefully, though with a trade-off in interpretability and input/output formatting complexity.
- **LLMs QA alignment**: QA4RE is anticipated to achieve high accuracy on tasks requiring precise relation identification, particularly in datasets with diverse or ambiguous entity relationships, such as the BioRED dataset. Its question-answering formulation, leveraging pre-trained language models, is well-suited for extracting relations from complex biomedical texts by transforming them into targeted natural language queries.

- **Entity Extraction Dependencies**: Techniques like BioBERT and SciFive, which operate on pre-identified entity spans, may exhibit performance variations depending on how accurately these spans are provided. This highlights the importance of consistent and reliable input formatting for fair comparison.
- **Domain Adaptation Benefits**: Across all models, we expect that those pretrained on biomedical corpora (BioBERT, SciFive) will outperform general-domain counterparts (e.g., BERT, T5) due to better handling of domain-specific terminology, abbreviations, and linguistic structures.

We also anticipate that the implementation and evaluation process itself will reveal practical considerations regarding compute efficiency, ease of adaptation, and model robustness—factors that are increasingly relevant when deploying RE systems in real-world biomedical pipelines.

6 Evaluation Metrics

To assess the performance of our relation extraction techniques, we will employ both standard and task-specific evaluation metrics. These metrics are chosen to ensure fair, transparent, and comparable evaluation across different models and methods.

6.1 Relation Extraction

For all relation extraction models, we will compute the following metrics:

- **Precision (P)**: The proportion of predicted relations that are correct.
- **Recall (R)**: The proportion of gold-standard relations that are correctly predicted.
- **F1 Score**: The harmonic mean of precision and recall, representing the overall effectiveness of the model in identifying correct relations.

A relation prediction is considered correct if both the entity span boundaries and the relation label match the ground truth. In cases where models output free-form text (e.g., SciFive), a mapping procedure will be applied to compare generated labels with gold labels. We further assume that entity span prediction is assumed to be correct or pre-resolved, depending on the model.

6.2 Computational and Practical Considerations

In addition to accuracy-based metrics, we will also qualitatively track:

- **Ease of integration and reproducibility**: Based on training stability, setup complexity, and codebase modularity.
- **Inference efficiency**: Relative runtime and GPU/memory usage during inference on large biomedical documents.
- **Generalizability**: Performance differences between BioRED and auxiliary datasets (if used) to evaluate domain robustness.

Together, these metrics will allow us to make both quantitative and practical comparisons between extraction techniques and assess their viability for real-world scientific information processing systems.

7 Challenges and Limitations

While the proposed methodology is comprehensive and builds upon state-of-the-art tools, several challenges and limitations must be acknowledged.

7.1 Domain-Specific Complexity

Biomedical texts are characterized by high linguistic density, domain-specific terminology, and frequent use of abbreviations and acronyms. Even pretrained models such as BioBERT and SciFive may struggle with edge cases, such as nested entities, overlapping relations, or implicit knowledge that requires world or clinical understanding.

7.2 Entity-Relation Dependency

Certain relation extraction methods operate on predefined entity spans, meaning their performance is inherently tied to the quality of those spans. If entity boundaries or types are inaccurate, this can lead to misclassification of relations or missed connections. While our project does not address entity recognition directly, we recognize this dependency as a potential source of variation in RE outcomes, especially in pipeline-based models that assume perfect entity inputs. Future work could explore the impact of different span generation strategies on RE robustness.

7.3 Evaluation Ambiguity in Generative Models

Models like SciFive generate relation labels in natural language. This introduces ambiguity in evaluation—generated outputs may be semantically correct but not lexically identical to the gold label. Designing accurate and fair matching functions for evaluation is a non-trivial task and can affect reported performance.

7.4 Implementation Overhead

Not all models offer the same ease of use. Some methods require non-trivial setup, integration and customized preprocessing. Even with public codebases, porting to new datasets like BioRED requires careful alignment with annotation schemes and data formats.

7.5 Compute Requirements

While fine-tuning models like BioBERT or SpanMarker is relatively lightweight, training or decoding with large encoder-decoder models such as SciFive can be resource-intensive. This may restrict experimentation cycles or model ensembling, especially when evaluating across multiple datasets.

7.6 Dataset Limitations

BioRED, while well-annotated, is still limited in size and may not cover the full diversity of biomedical relation types encountered in real-world literature. This limits the generalizability of the conclusions drawn from evaluation, especially for models designed for broader or open-domain applications.

Despite these challenges, we believe that carefully controlled experiments and modular architecture choices will allow us to mitigate many of these risks and still yield valuable insights for the broader NLP and biomedical research communities.

Bibliography

[1] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 2022.

[2] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[3] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: a text-to-text transformer model for biomedical literature, 2021.

[4] Yu Su Kai Zhang, Bernal Jiménez Gutiérrez. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*, 2023.

[5] Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023.

[6] Yucan Guo, Zixuan Li, Xiaolong Jin, Yantao Liu, Yutao Zeng, Wenxuan Liu, Xiang Li, Pan Yang, Long Bai, Jiafeng Guo, and Xueqi Cheng. Retrieval-augmented code generation for universal information extraction, 2023.