From Named Entities to Relations: End-to-End Biomedical Information Extraction*

Notebook for the BIU ONLP Lab in the BioASQ Task on Gut-Brain Interplay Information Extraction at CLEF 2025

Ron Keinan[1,*], Amir David Nissan Cohen[1,2] and Reut Tsarfaty[1]

[1]Bar Ilan University, Ramat-Gan, Israel
[2]OriginAI

Abstract

This paper details our contribution to the BioASQ CLEF Lab 2025 [1] GutBrainIE shared task (task 6) [2]. The mission focuses on Named Entity Recognition (NER) and Relation Extraction (RE) from biomedical abstracts concerning the gut-brain axis, Parkinson's disease, and mental health. We developed a system leveraging large language models (LLMs), employing GLiNER for NER and ATLOP for RE, fine-tuned on various backbones including GLiNER Large Bio and *roberta large*. Our approach involved a three-stage pipeline: data processing, model fine-tuning, and prediction generation. We participated in four subtasks: NER (6.1), Binary Tag-Based RE (6.2.1), Ternary Tag-Based RE (6.2.2), and Ternary Mention-Based RE (6.2.3). Our results indicate strong performance in tag-based RE, with our *roberta-large* model achieving a micro-F1 score of 0.6122 in binary RE (ranked 5th out of 11) and 0.5911 in Ternary tag-based RE (ranked 6th out of 12), outperforming the baseline in both cases. However, our system struggled with NER (micro-F1 0.4816, ranking 15th) and particularly with Ternary Mention-Based RE (micro-F1 0.1799, ranking 11th), highlighting challenges with fine-grained entity detection and mention-level relation identification. We conclude that while large transformers are effective for extraction of biomedical relationships, future work must address domain adaptation for NER and explore joint modeling approaches to improve mention-level performance.

Keywords

Biomedical Information Extraction, Named Entity Recognition, Relation Extraction, Deep Learning, Transformers, GLiNER, ATLOP.

1. Introduction

We present a system for extracting biomedical entities and their relations concerning the gut microbiota, Parkinson's disease, and mental health from PubMed abstracts. These associations, while increasingly evidenced, remain buried within unstructured scientific texts. The challenge aims to promote the development of NLP systems for the identification of key biomedical entities and their relations within PubMed abstracts.

The task comprises two primary subtasks: Named Entity Recognition (NER) and Relation Extraction (RE). Subtask 6.1 (NER) involves detecting and classifying spans into 13 biomedical categories, including microorganisms, diseases, and chemicals. Subtask 6.2 (RE) is divided into three phases: binary relation detection (6.2.1), ternary tag-based classification (6.2.2) and mention-level relation identification (6.2.3). The dataset includes multiple annotation tiers (Platinum, Gold, Silver, Bronze), offering varied training and evaluation scenarios.

To address these tasks, we developed a system based on state-of-the-art Transformer models [3], specifically leveraging encoders [4] within a pipeline architecture. The preprocessing steps included annotation conversion and text normalization. For NER, we fine-tuned multiple GLiNER model variants[5], including domain-adapted backbones. For RE, we used the ATLOP architecture[6] , experimenting with *SapBERT from PubMedBERT fulltext* [7], *biobert v1.1 pubmed* [8] and *roberta large* [9]. Training was conducted in a supervised manner using combined datasets from the higher annotation tiers:

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
ronke21@gmail.com (R. Keinan); amirdnc@gmail.com (A. D. N. Cohen); reut.tsarfaty@biu.ac.il (R. Tsarfaty)

Platinum (externally reviewed expert annotations), Gold (in-house expert annotations), Silver (student annotations: A – high accuracy, B – moderate accuracy).

We submitted multiple runs per subtask, each based on a different transformer model and varying in training epochs, batch size, and filtering threshold. Notably, our best results were achieved in the tag-based RE subtasks (6.2.1 and 6.2.2). Our system involved data preprocessing and fine-tuning RuBERTa-Large with an ALTOP-style approach, ranked 5th and 6th and outperforming the baseline. In contrast, our performance in the NER subtask (15th) and the mention-level RE subtask (11th) lagged behind the top systems reveal ongoing difficulties with fine-grained biomedical NER and complex relation extraction. However, our performance in the NER (15th) and mention-level RE (11th) subtasks lagged, indicating that fine-grained biomedical entity and relation extraction remains a persistent challenge.

The reminder of this paper is organized as follows. We begin with a review of related work, followed by a description of our system architecture and data processing pipeline. We then present our experimental setup and results, including comparisons with other participants. We conclude with a discussion of the insights gained, limitations encountered, and directions for future research, particularly in improving domain adaptation, joint modeling, and knowledge integration for biomedical NLP.

2. Related Work

The exponential growth of textual data, particularly in specialized domains such as biomedicine[10][11], requires advanced Information Extraction (IE) techniques to help researchers and practitioners manage the influx of information and uncover new insights [12]. IE encompasses a variety of tasks, with Named Entity Recognition (NER) and Relation Extraction (RE) being fundamental for transforming unstructured text into structured knowledge.

2.1. Named Entity Recognition

Named Entity Recognition (NER), is a foundational task in Natural Language Processing (NLP) that involves identifying and categorizing predefined entities in unstructured text, such as names of persons, organizations, locations, dates, and monetary values [13]. Its primary objective is to locate spans of text that correspond to specific semantic categories, thereby enabling downstream applications such as information retrieval, question answering, and knowledge base construction.

To illustrate the application of NER in the biomedical domain, consider the following sentence extracted from a biomedical abstract [14]:

> "***Probiotics*** are live ***microorganisms*** that confer health benefits on the host when administered in adequate amounts."

In this example, an effective NER system should correctly identify and classify "*Probiotics*" as a `TherapeuticAgent` and "*microorganisms*" as a `BiologicalEntity`. Recognizing such fine-grained entities is essential for downstream biomedical applications, such as knowledge base construction, therapeutic analysis, and personalized medicine.

Historically, early NER systems were built using hand-crafted rules and domain-specific dictionaries. While effective in constrained environments, these rule-based methods lacked generalization across domains. The advent of statistical machine learning approaches—such as Hidden Markov Models (HMMs) [15], Support Vector Machines (SVMs) [16], and Conditional Random Fields (CRFs) [17]—marked a significant paradigm shift. These models leveraged annotated corpora and engineered features such as word shape, orthographic patterns, part-of-speech tags, and lexical resources like gazetteers to improve generalization and adaptability.

The advent of deep learning has significantly advanced NER capabilities[18]. Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks, often combined with CRFs (e.g., BiLSTM-CRF architectures), became the standard for their

ability to capture contextual information effectively [19, 20]. More recently, transformer-based models, pre-trained on vast text corpora, have set new benchmarks by learning rich contextual representations of words, with advancements seen in models like BERT [4], RoBERTa [21], and ELECTRA [22]. Large Language Models (LLMs) are also increasingly being explored for their capabilities in NER, often through few-shot or zero-shot learning approaches [23].

In the biomedical domain (BioNER), the focus shifts to identifying biologically significant entities such as genes, proteins, diseases, chemicals, cell lines, and microbial taxa [24]. The complexity and ambiguity of biomedical terminology[25][**?**], along with the constant emergence of new terms, make BioNER a particularly challenging task [26]. Initial BioNER systems also relied on dictionary-based methods and rule-based systems, followed by traditional machine learning models like CRFs [27]. However, deep learning methodologies, especially transformer-based pre trained language models like BERT [4] and its domain-specific variants such as *biobert v1.1 pubmed* [8], PubMedBERT [28], and ClinicalBERT [29], have demonstrated state-of-the-art performance. These models leverage vast amounts of unlabeled biomedical text to learn contextual representations, and are then fine-tuned on task-specific annotated datasets to recognize specific entity types relevant to a particular study.

2.2. Relation Extraction

Relation Extraction (RE) is a key task in Information Extraction that seeks to identify and classify semantic relationships between entities. In a general domain context, RE might identify that a person "works for" an organization, or that a company is "headquartered in" a particular city. Traditional RE systems often relied on hand-crafted linguistic patterns, dependency parsing, or kernel-based methods to capture syntactic and semantic interactions between entity pairs [30, 31]. These approaches typically involved extensive feature engineering to encode lexical, syntactic, and positional information surrounding the target entities.

To illustrate, consider the following sentence from a biomedical abstract:

> "Reduced levels of ***Lactobacillus*** have been associated with increased severity of ***depression*** symptoms in individuals diagnosed with ***Parkinson's disease***."

A successful RE system should detect a ternary relation of type `Microbe-Disease-MentalHealthLink`,
involving the entities: "*Lactobacillus*" (a `Bacterium`), "*Parkinson's disease*" (a `Disease`), and "*depression*" (a `MentalHealthCondition`). This example highlights the challenge of n-ary relation extraction, wherein more than two entities must be jointly considered within a unified relational frame. Such n-ary structures are prevalent in biomedical texts and are central to the knowledge discovery objectives of the GutBrainIE task [32, 33, 2].

In the biomedical domain, Relation Extraction (BioRE) is tasked with identifying clinically and biologically significant associations, such as gene-disease links, protein-protein interactions, or therapeutic drug-disease relationships. The GutBrainIE task extends this paradigm to encompass a range of entities, including microbial species, neurodegenerative diseases, mental health conditions, and physiological factors. The biomedical literature's syntactic complexity, domain-specific vocabulary, and frequent presence of long-range dependencies demand robust, domain-adapted modeling strategies [34, 35].

The advent of deep learning has significantly transformed RE by enabling models to learn discriminative features automatically from raw text. Convolutional Neural Networks (CNNs) have been employed to extract local contextual features [36], while Recurrent Neural Networks (RNNs), have been utilized to capture sequential dependencies [37]. More recently, attention mechanisms and transformer-based encoder architectures have achieved state-of-the-art performance by enabling models to dynamically attend to the most informative parts of the sentence [38].

Transformer-based encoder models, including those fine-tuned on biomedical corpora such as PubMedBERT or BioLinkBERT, have shown considerable efficacy in handling the intricacies of BioRE. These models can be adapted to treat relation extraction as a classification task over entity tuples, or through more sophisticated formulations using structured prediction or graph-based reasoning [35, 33].

2.3. Encoder-based Approaches for NER and RE

The landscape of Natural Language Processing has seen rapid advancements. These models, pre-trained on exceptionally large and diverse datasets, exhibit remarkable few-shot or even zero-shot learning capabilities for various tasks, including NER and RE. For instance, language models can be prompted for direct entity identification or fine-tuned with relative efficiency on smaller annotated datasets [39], and they are being explored in specialized areas like the biomedical domain for rapid adaptation and broader text applicability [40, 41]. Similarly, for RE, language models offer potential for open-ended relation extraction and inferential capabilities [42, 43].

However, alongside these developments, Transformer-based encoder architectures, such as BERT [44] and its numerous variants, remain highly effective and widely utilized for NER and RE tasks. These models excel at learning rich contextual representations from text and can be fine-tuned to achieve state-of-the-art performance on specific datasets and domains. For NER, architectures focusing on robust entity span detection and classification, exemplified by models like GLINER, leverage powerful encoder backbones. In the realm of RE, techniques often involve identifying entity pairs and classifying their relationships, with models such as ATLOP demonstrating sophisticated approaches to this task. Furthermore, specialized pre-training objectives, as seen in models like SuMeBERTs , can enhance performance on downstream tasks by tailoring the encoder's understanding to particular nuances of language or information structure.

While LLMs present exciting avenues, particularly for generative tasks and broad-domain applications, dedicated encoder-based models offer advantages in terms of computational efficiency for fine-tuning and inference, focused performance on specific predictive tasks, and often more direct interpretability of task-specific layers. Challenges in LLMs related to factual accuracy, hallucination mitigation (especially for specific biomedical entities), and structured prediction for complex relations [42] also underscore the continued relevance of developing and employing well-established and robust encoder-centric approaches. This work focuses on leveraging the strengths of such encoder-based models for NER and RE.

2.4. Biomedical Shared Tasks and Benchmark Challenges

Shared tasks have been instrumental in catalyzing progress within biomedical natural language processing by delivering high-quality benchmark datasets, well-defined evaluation metrics, and platforms for community engagement. Notable initiatives—including the BioCreative series for assessment of Information Extraction Systems in Biology [45], the BioNLP Shared Task for structural information extraction from biomedical literature [46], and the BioASQ challenge (focusing on biomedical semantic indexing and question answering) [47]. These shared tasks have each contributed to advances in entity recognition, relation extraction, and broader information-extraction tasks through the provision of domain-specific corpora and rigorous, competitive evaluation frameworks. These challenges foster methodological innovation by encouraging participants to devise models that can handle complex nomenclature, ambiguous terminology, and varied syntactic constructions inherent to biomedical text.

Building on this lineage, the Gut-BrainIE Task[1] concentrates specifically on the gut–brain axis, a domain characterized by intricate, multi-scale interactions between microbial, molecular, and physiological entities. By curating a corpus annotated with specialized entity types (e.g., microbial taxa, neuroactive compounds) and their interrelations (e.g., modulatory effects, transport mechanisms), Gut-BrainIE extends the shared-task paradigm to a highly interdisciplinary context. Moreover, its evaluation schema emphasizes not only exact-match accuracy but also the correct identification of nested and overlapping spans, as well as fine-grained relation categories that reflect causal or correlational links.

Our work leverages state-of-the-art NER architectures—such as GLiNER [5], which employs span classification—and sophisticated document-level RE models—like ATLOP [6], which integrates adaptive thresholding with localized context pooling—to address the specific challenges posed by the gut–brain literature. In doing so, we aim to demonstrate how modern transformer-based encoders [4], when

[1] https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/#

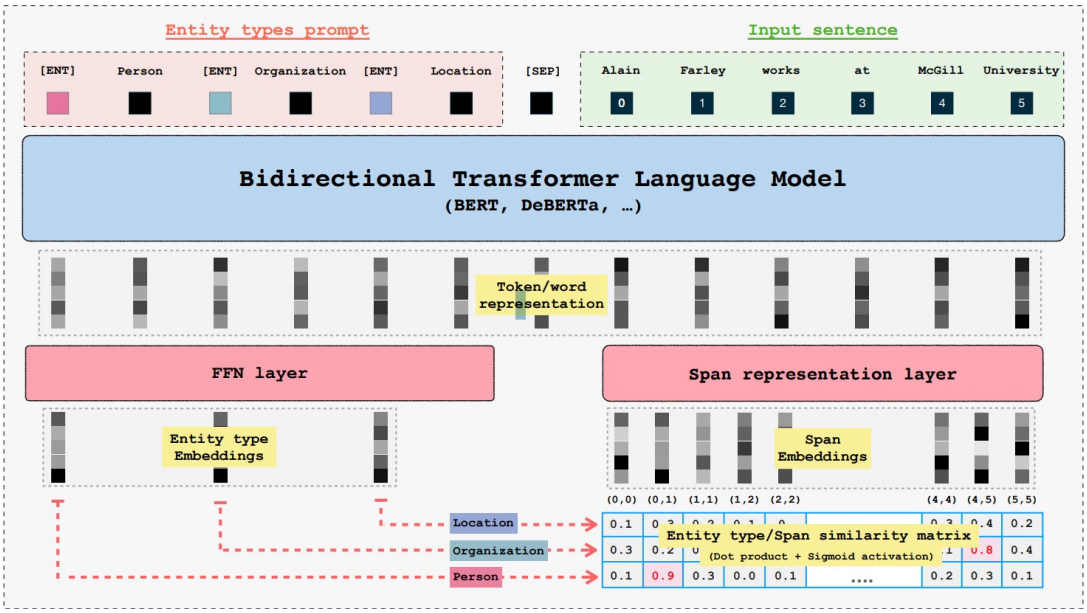

Figure 1: GLiNER Architecture (Figure taken from [5])

coupled with task-tailored pre and post-processing strategies, can achieve robust performance on a domain that demands both linguistic precision and biological insight.

3. Background and Model Foundations

3.1. GLiNER for Named Entity Recognition

GLiNER (Generalist Language model for NER) [5] is a span classification framework built upon a bidirectional language model (BiLM), such as BERT or DeBERTa. At inference time, GLiNER constructs a unified input sequence by prepending special marker tokens for each target entity type to the input sentence, forming a structure such as:

> `[ENT]` PERSON `[ENT]` LOCATION `[ENT]` ORGANIZATION ... `[SEP]` The input sentence.

This augmented sequence is processed by the BiLM, which generates contextualized embeddings for all tokens, including both the entity-type and the sentence words. GLiNER consists of three main components. First, the pretrained BiLM encodes the full sequence, capturing the contextual relationships across tokens. Second, a span representation module computes embeddings for each candidate span—defined as any contiguous sequence of words—by concatenating the embeddings of the span's start and end tokens and passing them through a two-layer feedforward network. Third, an entity representation module constructs embeddings for each entity type by refining the BiLM output for the corresponding `[ENT]` token using another feedforward network.

Both span and entity-type embeddings are projected into a shared latent space, where GLiNER calculates a similarity score for each span–type pair using a dot product. This score is passed through a sigmoid activation to estimate the probability that the span belongs to the given entity type. During training, the model is optimized using binary cross-entropy loss over all such span–type combinations. Positive pairs are those that appear in the annotated training data, while negative pairs are generated by sampling mismatched entity types from other training examples within the same batch. This negative sampling strategy helps the model learn to differentiate between correct and incorrect associations. It helps recognize that real-world data often lacks certain entity types. The model's performance is evaluated using different negative sampling ratios. Training with only positive entities leads to more

false positives (lower precision), while a high negative sampling ratio makes the model overly cautious, resulting in missed entities (lower recall).

For decoding, GLiNER uses a greedy span selection algorithm that identifies the most probable spans while enforcing task-specific constraints. In flat NER mode, only non-overlapping spans are selected. In nested NER mode, the algorithm allows for nested spans—those that are entirely contained within other spans—while avoiding partial overlaps. This approach allows GLiNER to efficiently extract both flat and hierarchical named entities in a single forward pass.

3.2. ATLOP for Document-Level Relation Extraction

Zhou et al. [6] introduce the ATLOP (Adaptive Thresholding and Localized Context Pooling) model for document-level multi-label relation extraction. ATLOP is built on a pretrained transformer encoder and avoids the use of intermediate graph structures by leveraging two key mechanisms:

- **Localized Context Pooling.** ATLOP repurposes the self-attention heads of the pretrained encoder to derive entity-level attention distributions. For each entity in a candidate pair, the model averages the self-attention scores over all mentions of that entity to obtain an attention distribution. These two distributions are then combined (via elementwise multiplication) to highlight tokens jointly relevant to both entities. The resulting fused weights are used to pool the encoder's token representations into a localized context embedding, which serves as a pair-specific representation for relation classification.
- **Adaptive Thresholding.** To address the multi-label nature of document-level RE, ATLOP replaces the conventional fixed decision threshold with a learnable threshold class. During training, a rank-based loss encourages true relation logits to exceed the threshold logit and non-relation logits to fall below it. At inference, any relation whose score surpasses the threshold class is predicted, thereby removing the need for heuristic threshold tuning and allowing flexibility across different entity pairs.

Extensive experiments on RE datasets such as DocRED, CDR, and GDA [6] demonstrate that ATLOP achieves state-of-the-art performance, validating the effectiveness of localized context pooling and adaptive thresholding in document-level relation extraction.

4. The GutBrainIE Task

4.1. Overview of the GutBrainIE Task

We participated in Task #6 [2] of the BioASQ [1] CLEF Lab 2025 (GutBrainIE[1]), which focuses on extracting entities and relations relevant to the gut–brain axis from PubMed abstracts. This task is divided into two main subtasks. The first subtask, Named Entity Recognition (NER, Subtask 6.1), involves identifying and classifying text spans into one of 13 predefined biomedical categories, such as bacteria, chemical, or microbiota. The expected output format includes the entity label, its location within the title or abstract, and the start and end character offsets. The second subtask, Relation Extraction (RE, Subtask 6.2), consists of three progressive phases. In Subtask 6.2.1, Binary RE, the goal is to detect whether a relation exists between any pair of entities within a PubMed abstract, without the need to specify the relation type. Subtask 6.2.2, Ternary tag-based RE, requires not only identifying related entities but also predicting the type of relation between them. Finally, Subtask 6.2.3, Ternary mention-based RE, focuses on pinpointing the specific entity mentions involved in a relation and classifying the type of relation they share.

4.2. Overview of the GutBrainIE Data

The annotated *GutBrainIE* corpus comprises titles and abstracts of biomedical articles retrieved from PubMed, focusing on the gut–brain interplay and its implications for neurological and mental health.

Table 1
Corpus Statistics by Annotation Tier and Split

Collection	# Docs	Total Entities	Avg Entities/Doc	Total Relations	Avg Relations/Doc
Train Platinum	111	3,638	32.77	1,455	13.11
Train Gold	208	5,192	24.96	1,994	9.59
Train Silver	499	15,275	30.61	10,616	21.27
Train Bronze	749	21,357	28.51	8,165	11.90
Development Set	40	1,117	27.93	623	15.58

Each entry in the dataset corresponds to a PubMed article identified by its PubMed ID (PMID) and is stored in JSON format for ease of integration with NLP pipelines. Entries include rich metadata—such as title, authorship, journal, publication year, and abstract—alongside the identifier of the annotator. Annotators are categorized as expert annotators, student annotators, and automated distant annotations.

The dataset contains two main types of annotations: *entities* and *relations*. Entity annotations consist of individual mentions, each defined by character offsets (start and end), location (title or abstract), the text span itself, and a semantic label (e.g., *bacteria*, *microbiome*). Relation annotations represent links between two entity mentions and include detailed information for both the subject and object: their character offsets, location, text span, and semantic label, as well as the *predicate* describing the type of relationship.

For ease of use in downstream tasks, relations are also provided in three derived formats: (1) *binary tag-based relations*, capturing label pairs of subject and object; (2) *ternary tag-based relations*, representing triplets of subject label, predicate, and object label; and (3) *ternary mention-based relations*, which include mention-level tuples comprising the subject and object text spans, their respective labels, and the predicate.

Annotations are organized into four hierarchical tiers, reflecting varying levels of quality and provenance. The overall distribution of annotations across these tiers is summarized in Table 1. Specifically, the tiers are defined as follows:

- **Platinum-Standard Annotations:** Highest-quality annotations, curated and externally reviewed by biomedical specialists to ensure maximal precision.
- **Gold-Standard Annotations:** High-quality annotations produced in-house by domain experts.
- **Silver-Standard Annotations:** Mid-quality annotations created by trained students under expert supervision, subdivided into two clusters:
  - *Student A:* Annotators demonstrating consistently high accuracy.
  - *Student B:* Annotators with less consistent performance.
- **Bronze-Standard Annotations:** Automatically generated annotations using fine-tuned GLiNER for NER and ATLOP for relation extraction.

These hierarchical tiers enable a nuanced understanding of annotation quality and provenance, which is critical for downstream evaluation and model training. By distinguishing between levels of human expertise and automation, the dataset supports flexible experimentation across a spectrum of reliability, allowing researchers to benchmark their systems under varying degrees of annotation fidelity.

5. Experimental Setup

5.1. System Architecture and Workflow

Our pipeline comprises three main stages: data processing, model fine-tuning, and prediction. During data processing, we standardize and aggregate annotated data to prepare it for training, structuring named entity recognition (NER) data in the GLiNER format, which includes entity spans and types

Table 2
Hyperparameters for GLiNER (NER task)

Hyperparameter	Value
Epochs	30
Batch size	8
Max sequence length	384
Warmup	10%
Optimizer	AdamW
Learning rate	5×10^{-5}
Inference threshold	0.9

in a structured JSON representation, and transforming relation extraction (RE) data into the ATLOP format, which captures entity pairs and their associated relations within the context of a document. In the model fine-tuning stage, we adapt the GLiNER model for NER using five variants—GLiNER multipii-v1, Medium, Large, Large Bio v0.1, and Large Bio v0.2—each offering different balances between general language capabilities, biomedical specialization, and model size to evaluate performance across diverse settings. For RE, we fine-tune the ATLOP model on four pretrained language models: *SapBERT from PubMedBERT fulltext*, *bert-base-cased*, *biobert v1.1 pubmed* v1.1, and *roberta large*, leveraging their distinct strengths in general and biomedical language representation to enhance relation extraction. Finally, in the prediction stage, we generate entity and relation outputs, merge the results from NER and RE, and convert them into the appropriate evaluation format.

5.2. Key Pre-processing Steps

Consistent pre-processing was applied to optimize model performance:

1. **Lowercase**: All input text was converted to **lowercase** to reduce vocabulary size and mitigate data sparsity. This standardizes input, treating "Organization" and "organization" as identical, which aids generalization despite potential loss of casing signals.
2. **Space Normalization**: This involved standardizing whitespace by collapsing multiple consecutive spaces into one and removing leading/trailing spaces. This ensures input consistency, preventing models from learning spurious patterns from irregular spacing.
3. **Tokenization**: Text was broken down into tokens using the specific tokenizer trained with the corresponding fine tuned model. This **sub-word tokenization** handles out-of-vocabulary words and captures morphological similarities. For NER, entity labels were carefully aligned with tokens, often assigning the primary label to the first sub-token.

5.3. Data and Training

For both Named Entity Recognition (NER) and Relation Extraction (RE) tasks, we utilized a combination of Platinum, Gold, and Silver datasets to maximize training coverage and robustness.

The NER model was trained using the GLiNER architecture. Training was conducted for 30 epochs with a batch size of 8 and a maximum sequence length of 384 tokens. The detailed hyperparameters are listed in Table 2.

For the RE task, we adopted the ATLOP model. Training was performed over 500 epochs with a batch size of 4 and a maximum input length of 1024 tokens. The complete configuration is summarized in Table 3.

Table 3
Hyperparameters for ATLOP (RE task)

Hyperparameter	Value
Epochs	500
Batch size	4
Max sequence length	1024
Warmup	6%
Optimizer	AdamW
Learning rate	5×10^{-5}

5.4. Evaluation Metrics

To rigorously evaluate system performance across both subtasks, we used the metrics provided by the task organizers[2]. The metrics Precision, Recall, and F1-score, computed under both macro- and micro-averaging schemes. These metrics allow for the assessment of model effectiveness at both the label level and across the entire label distribution. The evaluation criteria are consistent across all subtasks, and system outputs are benchmarked against manually annotated ground truth labels.

Let L denote the set of target labels:

- For Subtask 6.1, L includes all entity labels.
- For Subtask 6.2.1, L refers to pairs of (subject label, object label).
- For Subtasks 6.2.2 and 6.2.3, L comprises triples of (subject label, predicate, object label).

The macro-average scores compute metric values independently for each label and then average them, treating all labels equally regardless of their frequency. In contrast, micro-average scores aggregate the contributions of all classes to compute overall metrics, giving more weight to frequent labels.

The evaluation metrics are formally defined as follows:

$$P_{\text{macro}} = \frac{1}{|L|} \sum_{l \in L} \frac{TP_l}{TP_l + FP_l}, \quad R_{\text{macro}} = \frac{1}{|L|} \sum_{l \in L} \frac{TP_l}{TP_l + FN_l},$$

$$F1_{\text{macro}} = 2 \cdot \frac{P_{\text{macro}} \cdot R_{\text{macro}}}{P_{\text{macro}} + R_{\text{macro}}}, \quad P_{\text{micro}} = \frac{\sum_{l \in L} TP_l}{\sum_{l \in L} (TP_l + FP_l)},$$

$$R_{\text{micro}} = \frac{\sum_{l \in L} TP_l}{\sum_{l \in L} (TP_l + FN_l)}, \quad F1_{\text{micro}} = 2 \cdot \frac{P_{\text{micro}} \cdot R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}}.$$

The primary leaderboard metric for ranking participating systems is the micro-averaged F1-score, as it more effectively captures model performance in the presence of class imbalance, a common characteristic of real-world relation extraction tasks. Nonetheless, macro-averaged scores and per-relation metrics are also reported to provide a comprehensive performance profile, including sensitivity to rare and long-tail classes.

6. Results

6.1. Participation Overview

The shared task comprised four subtasks. Subtask 6.1 (NER) saw 16 teams submit at least one run, ranging from the organizer's baseline to ensemble and transformer-based systems. Subtask 6.2.1 (Binary Tag-Based RE) included 11 teams, with top systems leveraging both rule-based and deep-learning approaches. Subtask 6.2.2 (Ternary Tag-Based RE) attracted 12 teams, many extending their binary-tag RE pipelines to support a neutral relation label. Subtask 6.2.3 (Ternary Mention-Based RE) saw 12 teams tackling mention-level relation extraction with three labels, including several strong neural-model submissions.

[2] https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/#five

Table 4
Results for BIU-ONLP submissions across all subtasks.

Task	Run	System Description	Macro-P	Macro-R	Macro-F1	Micro-P	Micro-R	Micro-F1
T6.1	1	*multi_pii-v1*	**0.4627**	0.3687	0.3846	0.4908	0.4721	**0.4813**
T6.1	2	*gliner_medium-v2.1*	0.4049	0.3717	**0.3864**	0.4866	0.4568	0.4712
T6.1	3	*gliner_large_bio-v0.1*	0.4393	0.3585	0.3711	0.4916	**0.4721**	**0.4816**
T6.1	4	*gliner_large-v2.1*	0.4029	0.3710	0.3842	0.4893	0.4632	0.4759
T6.1	5	*gliner_large_bio-v0.2*	0.4488	**0.3633**	0.3775	**0.4961**	0.4632	0.4791
T6.2.1	1	*SapBERT*	0.3846	**0.3598**	0.3545	0.6554	0.5022	0.5686
T6.2.1	2	*bert-base-cased*	0.4383	0.2912	0.3273	**0.7955**	0.4545	**0.5785**
T6.2.1	3	*biobert v1.1*	0.4309	0.2965	0.3293	0.7519	0.4199	0.5389
T6.2.1	4	*roberta large*	**0.4632**	0.3379	**0.3713**	0.7453	**0.5195**	0.6122
T6.2.2	1	*SapBERT*	0.3734	**0.3454**	0.3430	0.6497	0.4733	0.5476
T6.2.2	2	*bert-base-cased*	0.4467	0.2799	0.3182	**0.7803**	0.4239	**0.5493**
T6.2.2	3	*biobert v1.1*	0.4134	0.2866	0.3187	0.7519	0.3992	0.5215
T6.2.2	4	*roberta large*	**0.4725**	0.3288	**0.3630**	0.7362	**0.4938**	0.5911
T6.2.3	1	*SapBERT*	0.0777	**0.0807**	0.0765	0.2033	0.1327	0.1606
T6.2.3	2	*bert-base-cased*	**0.1274**	0.0777	**0.0899**	**0.2929**	0.1166	0.1668
T6.2.3	3	*biobert v1.1*	0.0935	0.0682	0.0683	0.2459	0.1206	0.1619
T6.2.3	4	*roberta large*	0.1171	0.0854	0.0879	0.2339	**0.1461**	**0.1799**

6.2. System Results

The results for the BIU-ONLP submissions across all subtasks and runs are shown in Table 4. For Subtask T6.1, the multi pii v1 model achieved the highest Macro-F1 and Micro-F1 scores among all submissions, indicating strong overall performance on entity recognition. Among the GLiNER variants, gliner large bio-v0.1 performs best in terms of Micro-F1. For Subtask T6.2.1 and T6.2.2, *roberta large* consistently outperforms other models across both macro and micro metrics, demonstrating its robustness in different settings. Finally, Subtask T6.2.3 shows significantly lower scores overall, reflecting the higher difficulty of this task, although *roberta large* still maintains a marginal lead in performance.

6.3. Comparison to Other Participants

Table 5 presents our rank, micro-F1, the best micro-F1, and the organizer's baseline for each subtask. We see that our system performed competitively on the Binary and Tag-Based Relation Extraction (RE) subtasks, ranking mid-field and surpassing the baseline in both cases. In Binary RE (6.2.1), we achieved a micro-F1 of 0.6122, placing 5th out of 11, and similarly in Tag-Based RE (6.2.2), we ranked 6th out of 12 with a score of 0.5911. However, our performance in the NER (6.1) and Mention-Based RE (6.2.3) subtasks was substantially lower, with our NER system ranking near the bottom and our Mention-Based RE system showing the largest performance gap relative to the best and baseline systems.

- **NER (Subtask 6.1):** Our best micro-F1 (0.4816) places us near the bottom (15th of 16), trailing the baseline by over 0.31. The large gap to the top system (0.8408) suggests specialized biomedical NER models or ensembling are crucial.
- **Binary Tag-Based RE (Subtask 6.2.1):** Ranking 5th of 12 with 0.6122 micro-F1, we outperform the baseline by 0.0175. The margin to the best (0.6864) is 0.0742, indicating competitive performance but room for relation-specific fine-tuning.
- **Ternary Tag-Based RE (Subtask 6.2.2):** Our 6th place with 0.5911 micro-F1 is above the baseline by 0.0160. The 0.0955 gap to the top performer highlights challenges in neutral-relation distinction.
- **Ternary Mention-Based RE (Subtask 6.2.3):** Our performance (0.1799) is well below both baseline and top, reflecting the challenge of mention-level extraction. Low recall suggests the mention detection module needs enhancement via joint modeling.

Table 5
BIU-ONLP micro-F1 ranking versus top and baseline systems.

Task	Our Rank	Our micro-F1	Best micro-F1	Baseline micro-F1
NER (6.1)	15/16	0.4816	0.8408	0.7927
Binary RE (6.2.1)	5/11	0.6122	0.6864	0.5947
Tag-Based RE (6.2.2)	6/12	0.5911	0.6866	0.5751
Mention-Based RE (6.2.3)	11/12	0.1799	0.4635	0.3288

7. Discussion and Conclusions

Our participation in the CLEF 2025 biomedical shared task focused on four challenging subtasks, including named entity recognition (NER) and multiple forms of relation extraction (RE). The evaluation of our systems yields several important insights into model performance, dataset complexity, and directions for future research.

The strongest results were obtained using large pretrained transformer architectures, such as *roberta large*. These models demonstrated particularly robust performance in the RE subtasks, achieving a micro-F1 score of 0.6122 in Binary RE and 0.5911 in Ternary tag based RE. These outcomes highlight the value of contextualized representations in modeling biomedical relations, especially under limited supervision.

In contrast, the NER subtask proved significantly more difficult in our settings. Our best-performing NER model achieved a micro-F1 score of 0.4816. This performance gap underscores the challenges posed by the dataset, including the presence of fine-grained and domain-specific entity types that are often rare or absent in general pretraining corpora. The resulting domain shift and data sparsity hindered generalization, particularly in recognizing low-frequency entities. Broadening the training dataset to include a wider range of annotated entities across related domains could enhance the model's ability to generalize to rare or unseen entity types.

In stark contrast to the RE tasks, our performance on the NER subtask (6.1) was unexpectedly poor, with our best system ranking 15th out of 16 participants. This result fell significantly short of our expectations, as our Micro-F1 score (0.4816) trailed the official baseline (0.7927) by a substantial margin. This wide gap suggests that our chosen models were ill-suited for the specific challenges of this dataset. While we initially hypothesized a domain shift, the scale of the underperformance indicates a more fundamental mismatch. The top-performing systems and the baseline likely leveraged models with extensive pre-training on biomedical corpora (e.g., *biobert v1.1 pubmed*, PubMedBERT), giving them a decisive advantage in recognizing the domain-specific entity types that our more general models struggled with.

A particularly noteworthy and surprising finding from this subtask was the relative performance of our GLiNER model variants. Counterintuitively, *gliner medium v2.1* achieved a Macro-F1 score (0.3864) that was not only comparable but slightly superior to its larger counterparts, *gliner large v2.1* (0.3842) and *gliner large bio v0.1* (0.3711). This outcome was somewhat unexpected. While larger models usually outperform smaller ones due to their greater capacity, in this case, *gliner medium v2.1* may have benefited from better regularization or simply a more favorable initialization. Alternatively, the performance convergence may indicate that the available training data was insufficient for the larger models to fully realize their capacity, leading to overfitting or unstable learning. This points to the importance of model-data fit, particularly in low-resource domains such as biomedical NER.

The low recall in Ternary Mention-Based RE (macro-F1 < 0.09) points to limitations in current pipeline architectures for span detection and relation classification. It might be that the low reults in our NER module created a cascade of errors, as relations cannot be correctly identified if the constituent entities are missed. This confirms that for complex, mention-based RE, a pipeline architecture is suboptimal. We suggest that future systems may benefit from joint models that integrate entity and relation extraction into a unified framework. Second, while large transformer models excel in high-resource conditions, they are prone to overfitting when applied to tasks with limited annotated data. Addressing this requires

exploring lightweight alternatives such as domain-specific data augmentation to enhance generalization.

Moreover, error patterns in both NER and RE indicate confusion between semantically similar entity types and fine-grained relation labels. These ambiguities suggest that leveraging external knowledge bases to better disambiguate closely related entities and relations.

In conclusion, while our models achieved competitive performance in several subtasks, the CLEF 2025 dataset exposes persistent challenges in biomedical information extraction. Progress in this field will require not only more sophisticated architectures but also greater emphasis on domain adaptation, structured knowledge integration, and robust learning from sparse data.

Acknowledgments

This work was supported by the Israeli Innovation Authority.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini, and Grammarly in order to: Grammar and spelling check, Paraphrase, and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, G. M. Di Nunzio, N. Ferro, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering, volume TBA of *Lecture Notes in Computer Science*, Springer, 2025, p. TBA.

[2] M. Martinelli, G. Silvello, V. Bonato, G. M. Di Nunzio, N. Ferro, O. Irrera, S. Marchesin, L. Menotti, F. Vezzani, Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

[3] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations, 2020, pp. 38–45.

[4] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/.

[5] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, Gliner: Generalist model for named entity recognition using bidirectional transformer, arXiv preprint arXiv:2311.08526 (2023). URL: https://arxiv.org/abs/2311.08526.

[6] W. Zhou, K. Huang, T. Ma, J. Huang, Document-level relation extraction with adaptive thresholding and localized context pooling, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, AAAI Press, 2021, pp. 7093–7101. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17164.

[7] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, N. Collier, Self-alignment pretraining for biomedical entity representations, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 4220–4230. URL: https://aclanthology.org/2021.naacl-main.334/.

[8] 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 (2020) 1234–1240.
[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.
[10] R. Keinan, E. Margalit, D. Bouhnik, Analysis of user trends in digital health communities using big data mining, Plos one 19 (2024) e0290803.
[11] R. Keinan, E. A. Margalit, D. Bouhnik, Impacts of a public health crisis on health-centered online social networks, Informing Science: The International Journal of an Emerging Transdiscipline 28 (2025) 022.
[12] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, Journal of biomedical informatics 77 (2018) 34–49.
[13] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes 30 (2007) 3–26.
[14] H. Mirzaei, S. Sedighi, E. Kouchaki, E. Barati, E. Dadgostar, M. Aschner, O. R. Tamtaji, Probiotics and the treatment of parkinson's disease: An update, Cellular and Molecular Neurobiology 42 (2022) 2449–2457. doi:10.1007/s10571-021-01128-w, epub 2021 Jul 20.
[15] L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (1989) 257–286.
[16] C. Cortes, V. Vapnik, Support-vector networks, Machine Learning 20 (1995) 273–297.
[17] J. Lafferty, A. McCallum, F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann, 2001, pp. 282–289.
[18] D. Bareket, R. Tsarfaty, Neural modeling for named entities and morphology (NEMO2), Transactions of the Association for Computational Linguistics 9 (2021) 909–928. URL: https://aclanthology.org/2021.tacl-1.54/. doi:10.1162/tacl_a_00404.
[19] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, in: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016, pp. 260–270.
[20] J. P. Chiu, E. Nichols, Named entity recognition with bidirectional lstm-cnns, Transactions of the Association for Computational Linguistics 4 (2016) 357–370.
[21] Y. Liu, H. Zao, M. Ghosal, A. Goyal, Y. Ding, J. Yang, L. Yang, D. Huang, D. Ma, Z. Yang, X. He, L. Du, H. Hu, J. Wang, J. Yang, L. Yang, D. Huang, D. Ma, Z. Yang, X. He, L. Du, H. Hu, J. Wang, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[22] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, ELECTRA: Pre-training text encoders as discriminators rather than generators, International Conference on Learning Representations (2020).
[23] I. Keraghel, S. Morbieu, M. Nadif, Recent advances in named entity recognition: A comprehensive survey and comparative study, arXiv preprint arXiv:2401.10825 (2024).
[24] S. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, 2019, pp. 2145–2158.
[25] R. Keinan, E. Margalit, D. Bouhnik, Emotional analysis in a morphologically rich language: Enhancing machine learning with psychological feature lexicons, Electronics 14 (2025) 3067.
[26] S. Kim, J. Lee, S.-Y. Shin, J. Kang, Bioner: a comprehensive survey of named entity recognition in the biomedical domain, Briefings in bioinformatics 21 (2020) 2111–2127.
[27] C. E. Kusuma, K. H. Chen, M. H. Hsieh, W. J. Lee, Y. C. Chang, Bioner: A survey of pre-trained language models for biomedical named entity recognition, Briefings in Bioinformatics 24 (2023) bbad346. doi:10.1093/bib/bbad346, published online Oct 2023, covers recent trends. If your Kusuma2024 is different, please replace this.
[28] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Pubmedbert:

Domain-specific language model pretraining for biomedical text mining, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 8537–8547.
[29] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical bert embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019, pp. 72–78.
[30] D. Zelenko, C. Aone, A. Richardella, Kernel methods for relation extraction, Journal of machine learning research 3 (2003) 1083–1106.
[31] R. C. Bunescu, R. J. Mooney, A shortest path dependency kernel for relation extraction, in: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 724–731.
[32] N. Peng, H. Poon, C. Quirk, K. Toutanova, W.-t. Yih, Cross-sentence n-ary relation extraction with graph lstms, Transactions of the Association for Computational Linguistics 5 (2017) 101–115.
[33] G. Westerfield, K. Lee, R. Xu, K. C.-C. Chang, TRIDENT: A TRansformer-based method for open-world n-ary relation extraction, in: The Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Washington, DC, USA, February 7-14, 2023, AAAI Press, 2023, pp. 13889–13897. URL: https://ojs.aaai.org/index.php/AAAI/article/view/26615.
[34] S. Zhao, L. Wang, Z. Liu, F. Li, Gutbrainie: A dataset for information extraction of gut-brain axis from scientific literature, Database 2021 (2021) baab028. doi:10.1093/database/baab028.
[35] C.-H. Lin, D. Ji, F. Li, A survey on biomedical relation extraction: methods and applications, Briefings in Bioinformatics 22 (2021) bbab149.
[36] D. Zeng, K. Liu, S. Lai, G. Zhou, J. Zhao, Relation classification via convolutional deep neural network, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 2335–2344.
[37] D. Zhang, D. Wang, Relation classification via recurrent neural network, arXiv preprint arXiv:1508.01006 (2015).
[38] A. D. Cohen, S. Rosenman, Y. Goldberg, Supervised relation classification as two-way span-prediction., in: AKBC, 2022.
[39] J. Wang, Y. Li, S. Chang, S. Liu, P. S. Yu, Is chatgpt a good ner annotator? a preliminary study, arXiv preprint arXiv:2304.12790 (2023).
[40] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (2023) 172–180.
[41] M. Agrawal, S. Hegselmann, D. Sontag, Large language models are diverse nlu models, arXiv preprint arXiv:2305.12539 (2023).
[42] X. Han, W. Zhao, Z. Liu, Language Model Is a Good ReLation Extractor, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2896–2907. URL: https://aclanthology.org/2023.findings-acl.184.
[43] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, Briefings in bioinformatics 23 (2022) bbac409.
[44] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.
[45] R. Islamaj, P.-T. Lai, C.-H. Wei, L. Luo, T. Almeida, R. A. A. Jonker, S. I. R. Conceição, D. F. Sousa, C.-P. Phan, J.-H. Chiang, et al., The overview of the biored (biomedical relation extraction dataset) track at biocreative viii, Database 2024 (2024) baae069.
[46] D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024. URL: https://aclanthology.org/2024.bionlp-1.0/.
[47] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, Bioasq-qa: A manually curated corpus for biomedical question answering, Scientific Data 10 (2023) 170.