Graphwise @ CLEF-2025 GutBrainIE: Towards Automated Discovery of Gut-Brain Interactions - Deep Learning for NER and Relation Extraction from PubMed Abstracts

Aleksis Datseris[1,2,*,†], Mario Kuzmanov[3,1,*,†], Ivelina Nikolova-Koleva[1,4,†], Dimitar Taskov[5,6,†] and Svetla Boytcheva[1,2,†]

[1]*Ontotext, 111R Tsarigradsko Shosse blvd., 1784 Sofia, Bulgaria*
[2]*FMI, Sofia University, 5 "James Bourchier" Blvd., 1164 Sofia, Bulgaria*
[3]*Tübingen University, Geschwister-Scholl-Platz, 72074 Tübingen, Germany*
[4]*IICT, Bulgarian Academy of Sciences, Acad. G. Bonchev Str, bl.2, 1113 Sofia, Bulgaria*
[5]*Multiprofile Hospital for Active Treatment in Neurology and Psychiatry "St. Naum", 1 Dr. Lyuben Rusev St., 1113 Sofia, , Bulgaria*
[6]*Medical University of Sofia, 15 Akademik I. E. Geshov Blvd., 1431 Sofia, Bulgaria*

Abstract

This paper presents a set of approaches to tackle the named entity recognition and relation extraction from scientific literature, specifically targeting the gut-brain axis related terms and relationships between them. The proposed methods participated in the GutBrainIE Task at CLEF 2025 BioASQ Lab. The solutions rely on fine-tuned BERT-based models (BioBERT , BiomedNLP ELECTRA , BioBERT PubMed) and GLiNER for the named entity recognition task, ATLOP and REBEL fine-tuning for the relation extraction task. Hybrid models and ensemble of models are also demonstrated for end-to-end tasks. Notably, one of our proposed solutions ranked 2nd on the most difficult task of the challenge - Ternary Mention-based Relation Extraction, achieving micro-F1 37.29%. Our best system for Named Entity Recognition over the test set achieved micro-F1 80.1%. On the Binary Tag-based Relation Extraction subtask, our best solution achieved micro-F1 65.38% on the test set and on the Ternary Tag-based Relation Extraction subtask, our best result was micro-F1 63.72%. All of the proposed approaches demonstrated good performance, consistently outperforming baseline results across all subtasks of the GutBrainIE Task at CLEF 2025 BioASQ Lab.

Keywords

named entity recognition, relation extraction, gut-brain axis, biomedical NLP

1. Introduction

An increasing amount of research indicates that there exists a complex interaction between the gut and the brain [1, 2]. However, many of the biological mechanisms underlying this relationship remain unclear. In-depth study of the various aspects of these relationships, which are published in scientific publications, could be essential for further progress of the biomedical research.

One step in this direction is to use the wealth of scientific publications in the biomedical field available in PubMed. PubMed contains an extensive collection of peer-reviewed articles from leading scientific journals and conferences in the field of biomedicine and offers a rich resource for systematically identifying, extracting and synthesizing new scientific insights. However, the rapid pace of publication of new articles and the richness of the resource make the manual exploration of current results challenging, therefore it is important to provide tools for automated tracking of new developments to speed up the research in the area.

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

aleksis.datseris@graphwise.ai (A. Datseris); mario.kuzmanov@graphwise.ai (M. Kuzmanov); ivelina.nikolova@graphwise.ai (I. Nikolova-Koleva); dimtaskov@gmail.com (D. Taskov); svetla.boytcheva@graphwise.ai (S. Boytcheva)

0009-0001-7997-3942 (A. Datseris); 0009-0003-6320-6663 (M. Kuzmanov); 0009-0006-6322-0256 (I. Nikolova-Koleva); 0000-0002-0939-5382 (D. Taskov); 0000-0002-5542-9168 (S. Boytcheva)

By applying modern natural language processing (NLP) techniques to scientific literature, researchers can efficiently collect and analyze evidences, reveal hidden patterns, explore comorbidities and risk factors, and accelerate the discovery of new connections within the gut-brain axis, which may ultimately help to better understanding, prevention and treatment of related diseases.

In this paper, we present our research in developing deep learning and hybrid models for NLP, for automated information extraction from PubMed articles. The results we present are part of the GutBrainIE Task at CLEF 2025 BioASQ Lab [3, 4]. This task is divided into two main subtasks. The first one is Named Entity Recognition (NER) - with a focus on important terms such as genes, diseases, microbiomes, chemicals, etc. The second subtask addresses relation extraction (RE), covering the full palette of levels of detail from binary relationships to ternary relationships with an explicit indication of the type of relationship and the entities that are associated with the given relationship. The dataset includes titles and abstracts from PubMed, organized into four categories based on the quality of annotations - from expert validated to automatically generated labels.

The proposed solutions below are based on deep learning models including GLiNER [5], BioBERT [6], BiomedNLP ELECTRA [7], BioBERT PubMed [8] for NER and REBEL [9] and ATLOP [10] for RE.

The paper is organized as follows: Section 2 outlines the related work, state-of-the art models and achievements on the tasks; Section 3 defines the tasks and data; in Section 4 are listed the approaches applied for NER; in Section 5 are presented the results of NER; Section 6 discusses the approaches to RE, Section 7 - the results of RE and in Section 8 are provided discussion and conclusions of the study.

2. Related Work

State-of-the-art NLP approaches for information extraction from Biomedical literature use range of techniques from deep learning (BERT, BiLSTM-CRF, XML-R and LLMs), classical ML methods (CRF, SVM) and even traditional solutions like dictionary- and rule-based ones. The best F1 scores achieved for the NER task on PubMed articles are up to 89.3% and for RE the performance vary in the range 47.7-88.5%, depending on the types of relations and entities [11, 12, 13, 14]. In this overview we focus primarily on NER and RE approaches applied on PubMed articles only, due to the specific nature of scientific literature in contrast with terminology and vocabulary used in clinical trials and clinical texts. The main categories of extracted entities include genes, proteins, diseases, drugs etc. The types of the identified relations include gene-disease, drug-treatment etc. Most of the approaches address both NER and RE tasks [12], but there are also specific techniques for RE only [15], [14] and NER only. Luo et al. [12] applies BiLSTM-CRF, BioBERT-CRF, PubMedBERT-CRF for NER over 600 PubMed abstracts achieving F1-score: 89.3% (strict) and 93.5% (relaxed) and BERT-GT, PubMedBERT for RE achieving F1 score: 47.7% (novelty), 72.9% (entity pair) and 58.9% (relation type). Hassan et al. [15] focuses only on RE task and propose solutions on PubMed abstracts using an unsupervised approach based on BERT, part-of-speech tagging and verb embeddings, achieving F1 score of 88.5% for Drug-Drug Interaction (DDI dataset), while for ChemProt dataset achieves F1 85.8%. Sänger and Leser [14] also focus on RE only, proposing an approach based on Neural Networks, corpus-level entity embeddings and pair embeddings achieving overall improvement of F1 score in the range 4-29% over traditional methods.

One of the famous approaches to the problem is to use the innovative method GLiNER [5]. In contrast to the more traditional NER models, GLiNER does not need a pre-defined set of categories which saves the efforts for labeling data and/or retraining. It uses efficient Bidirectional Language Models (BiLMs: BERT, DeBERTa) to process the input: entity type prompt concatenated with the sentence or text. Instead of an autoregressive generation, the core concept of GLiNER is to find the best match between entity type embedding and textual span representation. While GLiNER outperforms large general-purpose models such as ChatGPT and Vicuna in a zero-shot context, it is efficient and does not need extensive resources to run. Due to this great flexibility and easy pipeline for inference, we employ GLiNER as a zero shot classifier to give us a strong starting point by setting the very first baseline. The family of BERT models has the obvious drawback that it has been pretrained on too many general domain texts, which is a strong reason to perform worse in domain-specific tasks. That's why for biomedical text mining, we also consider BioBERT [6] to improve the baseline of GLiNER. It takes advantage of the same "short encoders" mentioned above by being initialized with the weights of BERT - pretrained on English Wikipedia and BooksCorpus. It takes advantage in the subsequent training phase, where a large number of PubMed abstracts and PubMed Central (PMC) full-text articles are used for further pretraining. As a result, the fine-tuned BioBERT sets the new state-of-the-art for biomedical NER and biomedical relation classification/extraction (RE).

3. Tasks and Data

PubMed is one of the largest databases for biomedical literature, comprising more than 38 million citations. The documents provided by the challenge organizers consist of a title, abstract, author and journal metadata of a publication, retrieved from PubMed with explicit focus on the gut-brain interplay and its implications in neurological and mental health. To foster the development of effective Information Extraction (IE) systems within the context of the the GutBrainIE Task at CLEF 2025 BioASQ Lab[1] [4], the annotated training data is organized into 4 collections as demonstrated in Table 1. Each collection has a different name, suggesting the quality of its annotations. The Platinum-Standard annotations are of highest quality - expert-curated and reviewed by external biomedical specialists. The Platinum corpus consists of only 111 documents, making it the smallest among the train collections. Next, the Gold-Standard annotations are only expert-curated with 208 documents followed by the Silver-Standard, which are created by trained students under expert supervision. The Silver corpus is comprised of 499 examples. Finally, the largest but with lowest quality is the Bronze-Standard data. This collection is annotated by distant supervision using fine-tuned GLiNER [5] for NER and fine-tuned ATLOP [16] for RE. The organizers also provide a separate set of documents for validation - Dev set. The final evaluation is performed on an external Test set made up of only Platinum and Gold Standard articles which was released about two weeks before the official deadline.

Table 1
Number of documents per collection.

Collection	Number of documents
Platinum	111
Gold	208
Silver	499
Bronze	750
Dev	40
Test	40

3.1. Task 6.1 - NER

The systems for biomedical NER, described in depth in the next sections, are trained/fine-tuned on different subsets of the aforementioned data. For this task, only the corresponding "*entities*" annotations are taken into account. An entity refers to a tuple and is expressed as shown on the example:

```
{
    "start_idx": 26,
    "end_idx": 35,
    "location": "title",
    "text_span": "BrainBiota",
    "label": "microbiome"
}
```

The property *label* is the category of the entity which is located in *location* (either "title" or "abstract") and starts at position *start_idx* and ends at *end_idx* inclusive.

The goal of this task is to classify a text span (entity mention) into one of 13 pre-defined categories/labels (see the X-axis of Figure 1). We will use only the term "category" in the next sections as the term

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

"label" used by the organizers can be ambiguous in terms of data description (meaning (1) alternative name of the entity and also (2) entity category in the context of the given task). However, some of the categories have only very few annotations. This poses a potential challenge for larger pretrained models which need a substantial amount of data and resources to generalize well. For example, Figure 1 clearly shows the distribution between labels in the Platinum-standard collection. This plot remains almost unchanged for any other annotated dataset. The dominating labels are *Disorder-Disease-Finding (DDF)*, *chemical*, *bacteria*, *human*, and *microbiome*. Categories like *food*, *gene*, *drug*, and *statistical technique* are rarely found in the texts.

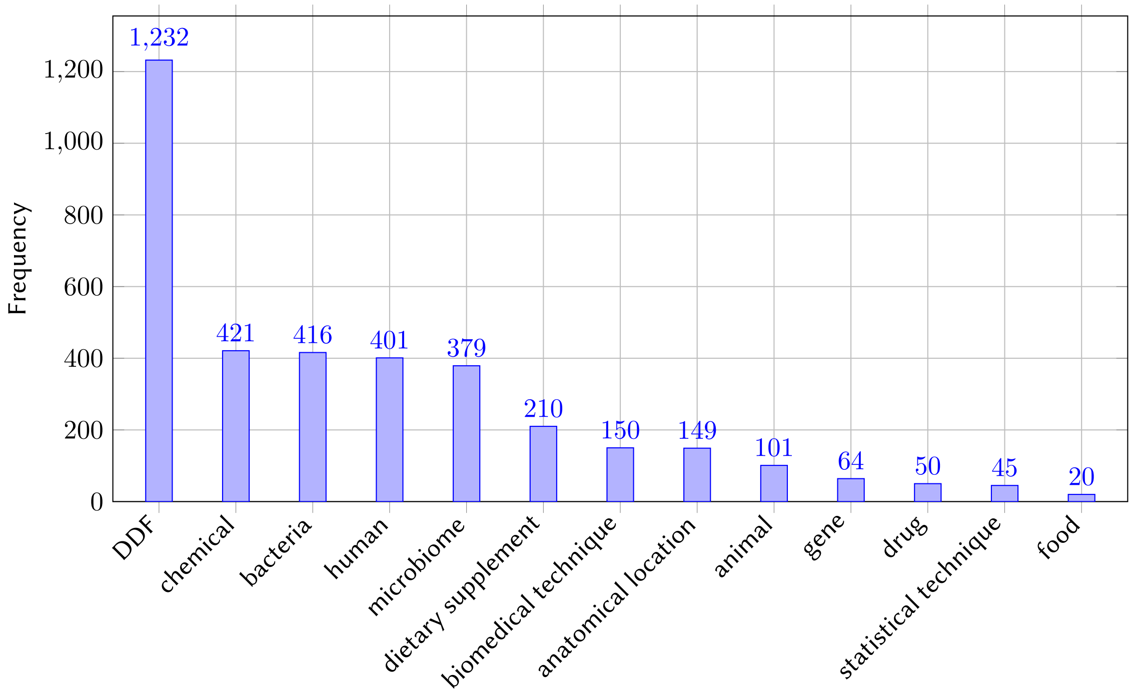

Figure 1: Frequency of named entity categories in the Platinum collection.

3.2. Task 6.2 - RE

The second main task is on RE and consists of three subtasks with increasing difficulty. The first of them is Binary Tag-based Relation Extraction (BT-RE) - requiring to identify whether two entity categories are in relation. Given the input text and/or the found entities, a binary relation is defined as an ordered tuple in the following example format:

```
{
    "subject_label": "DDF",
    "object_label": "human"
}
```

Where *DDF* and *human* are both named entity categories. It is important to note that the least represented categories are usually those involved in relations with only a small set of other categories - objects. For instance, *gene* only participates in relations with three other categories, in contrast to *human* or *microbiome* which appear as subject in relations to more than twice other labels.

In the next subtask - Ternary Tag-based Relation Extraction (TT-RE) the relation type (predicate) between the entity categories is also considered. An example of a ternary tag-based relation is:

```
{
    "subject_label": "microbiome",
    "predicate": "is linked to",
    "object_label": "DDF"
}
```

Table 2 lists the types of all allowed relations and further introduces the possible entities that could take the subject and objects positions for the respective predicate. Some of the possible predicates include *target*, *impact*, *influence*, *change effect*, *located in*, *is linked to*, and others. All relation types are asymmetric, meaning that (subject, predicate, object) **!=** (object, predicate, subject). On average, *influence*, *target*, *located in* are some of the most common predicates. It is worth noting that the Silver-standard collection has the largest number of annotated relations - 10 616, which is 5 times more than the relations annotated in the Gold collection - 1 994. Even though the number of documents is different, the average number of found relations per document is 21.27 for Silver, versus 9.59 for Gold. The highest quality collection - Platinum, has total of 1 455 relations and an average of 13.11 relations per document. For more detailed description on the statistics for the different relation types in the data, refer to the extended overview paper [4].

Finally, the last subtask on Ternary Mention-based Relation Extraction (TM-RE) aims to extract every text mention of the entities with their corresponding relation. The annotated training data has the following format:

```
{
    "subject_text_span": "IgA-Biome",
    "subject_label": "microbiome",
    "predicate": "located in",
    "object_text_span": "AR and TD patients",
    "object_label": "human"
}
```

Across all subtasks, a true positive is only considered when there is a full match of the found instance. Therefore, the possibility of partial matches and potentially accounting for some of the systems errors is eliminated.

Table 2
Types of relations in the provided corpus.

Subject category	Object category	Predicate
Anatomical Location	Human / Animal	Located in
DDF	Anatomical Location	Strike
Bacteria	Bacteria / Chemical / Drug	Interact
DDF	Bacteria / Microbiome	Change abundance
Bacteria	DDF	Influence
DDF	Chemical	Interact
Bacteria	Gene	Change expression
DDF	DDF	Affect / Is a
Bacteria	Human / Animal	Located in
DDF	Human / Animal	Target
Bacteria	Microbiome	Part of
Drug	Chemical / Drug	Interact
Chemical	Anatomical Location / Human / Animal	Located in
Drug	DDF	Change effect
Chemical	Chemical	Interact / Part of
Human / Animal / Microbiome	Biomedical Technique	Used by
Chemical	Microbiome	Impact / Produced by
Microbiome	Anatomical Location / Human / Animal	Located in
Chemical / Dietary Supplement / Drug / Food	Bacteria / Microbiome	Impact
Microbiome	Gene	Change expression
Chemical / Dietary Supplement / Food	DDF	Influence
Microbiome	DDF	Is linked to
Chemical / Dietary Supplement / Drug / Food	Gene	Change expression
Microbiome	Microbiome	Compared to
Chemical / Dietary Supplement / Drug / Food	Human / Animal	Administered

3.3. Augmented Data

As a potential solution to the problem of scarcity in a number of important categories, as observed, PubMed API is used to fetch more articles and augment the training data. As a matching criterion, only articles in the period (2015/01/01 - 2025/01/01) are considered, and a specially designed search query is

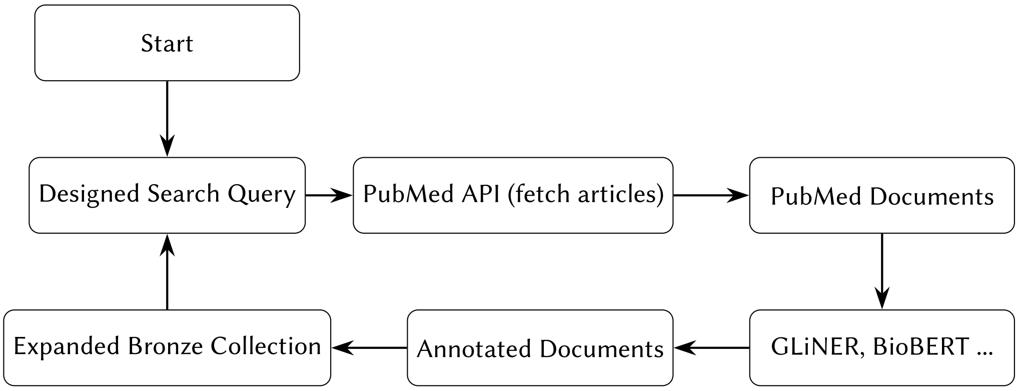

Figure 2: Process of augmentation of the Bronze-standard collection.

created to extract all articles, where the gut-brain axis or the gastrointestinal microbiome are defined as major topics, following the MeSH ontology.

To fully ensure that the newly extracted articles are related to the task requirements, the search query is extended to also account for specific labels relevant to the needed categories. This way, for the rare category *gene* the query extracts all relevant articles within the gut-brain axis that have their main topics defined to include genes, DNA, or genetics. The designed search query looks like this:

```
  ((Brain-Gut Axis[MeSH Major Topic]) OR (Gastrointestinal Microbiome[MeSH Major Topic])) AND
(genes[MeSH Major Topic] OR DNA[MeSH Major Topic] OR genetics[MeSH Subheading])
```

Following the strategy shown on Figure 2 for all categories by changing the query and annotating the entities by distant supervision, usually with different combinations of systems, resulted in the creation of our own Bronze-standard annotated corpus with a total of 6728 articles.

4. NER Approaches

For the NER task several different techniques are presented. Two different pretrained language models (LMs) - BERT [17] and ELECTRA [18] are considered. BERT-based models are pretrained on objectives such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), in contrast to what ELECTRA-based models are trying to learn. This family of models uses a small generator network to corrupt the input, similar to BERT, by replacing some tokens with plausible alternatives and then turns to train a discriminator, learning whether each token was in the original input or was artificially generated. The results of changing the learning objective lead to better contextual representations and outperform BERT on the GLUE natural language understanding benchmark. Furthermore, the learned embeddings are especially more representative for small models, which are actually more suitable for the context of the gut-brain task. To leverage the capabilities of more compute efficient and smaller bidirectional LMs, fine-tuned GLiNER are also included in our experiments. GLiNER can extract any entity type by maximizing the probability of a span `(i,j)` to be of the correct entity type `t` and minimizing it, for the same span `(i,j)` to be of any other type `t`. Finally, our highest-precision and highest-recall system is a combination of several models - ensemble. When using multiple models, the final decision is based on pre-defined rules or majority voting. For example - if model *X* performs best for category *statistical technique*, then on test time we will take the predictions for this category, if any, only from model *X*.

4.1. BERT/ELECTRA-based

To fine-tune models from the BERT/ELECTRA family, standard token classification with Hugging Face is employed. The first step is to adjust the training data or in other words to "tag" the input in the widely adopted BIO-format, first introduced by Ramshaw and Marcus [19]. However, the training data is not always the same. This means, we experiment with different parts of the whole dataset. In some configurations, the models are only given the Platinum and Gold Standard collections. In others, they are fine-tuned on all collections including the expanded Bronze corpus created by us. To convert the provided format to BIO, ScispaCy [20] is used, a python package that provides spaCy [2] models for scientific text preprocessing. Due to its size and complexity, *en_core_sci_sm* model was chosen. The inference happens by creating a pipeline with the saved model and making the predictions on entity level, where the input is no longer a token, but a whole chunk of text. To choose the most successful tokenizer setup was selected through experiments with different maximum sequence lengths starting from 128 and going up to 512. For simplicity, no additional tricks are considred to expand this window. Most of the documents do not exceed 400 tokens, including the subwords, as shown in Figure 3. Experiments with larger maximum lengths like the one in *kiddothe2b/biomedical-longformer-large* do not show better performance for the data.

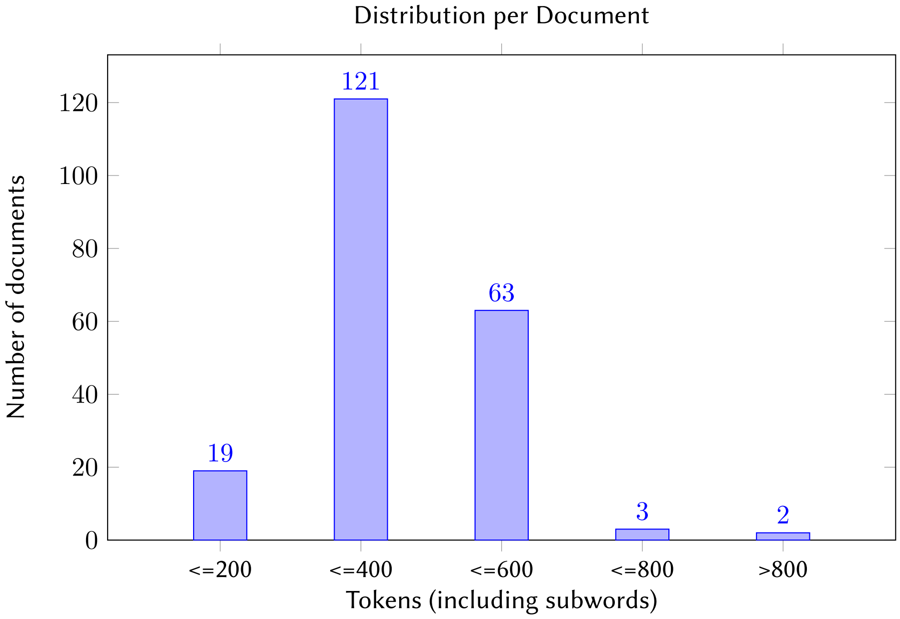

Figure 3: Token distribution per document.

To arrive at this conclusion, the data is tokenized with *biobert-base-cased-v1.1*. Although during fine-tuning different tokenizers are used with the different models, almost all of them use the WordPiece tokenization algorithm. The models in these experiments are domain-specific, pretrained on large biomedical corpora, usually consisting of PubMed abstracts and/or PubMed Central (PMC) full texts.

Here is the list of our most successful models on the development set. Their results on the official test dataset are shown in the next section 5. The exact choice on the hyperparemeters of the submitted models is illustrated in Table 7.

- *dmis-lab/biobert-base-cased-v1.1* [6] - The base version of the model performs better than the large for all categories. A strong indicator for it might be the size of our training corpus. This model is pretrained on PubMed abstracts and PMC full texts. Surprisingly, our most successful submission turned out to be with only fine-tuning on the bronze corpus. It is notable that the model exhibits one of the best precisions on the development set and is used further in a pipeline to form a stronger ensemble.

- *microsoft/BiomedNLP-BiomedELECTRA-base-uncased-abstract* [7] - When using this model, pre-trained on PMC full texts, the performance decreases. For the remaining of this article, this

[2] https://spacy.io/

model is referred as **pubmed-electra-base**. The initial model configuration has 110M parameters before fine-tuning. The best results on the development set are achieved by fine-tuning on all annotated data and the expanded Bronze collection. The specific annotator for this collection is another model - *dmis-lab/biobert-base-cased-v1.1*, which is fine-tuned only on the standard Bronze collection beforehand. The model achieves the best precision after the ensemble approach, which we will further describe. The recall is also slightly worse than this of the ensemble and GLiNER, but overall this configuration shows clear improvements. However, on the official test set the performance decreases. The main reason behind this is that the recall drops by approximately 7% (82.45% - dev, vs. 74.86% test) which also results in a drop in the F1-micro score.

- *monologg/biobert_v1.1_pubmed* [8] - Lighter version of *biobert-base*, only pretrained on PubMed abstracts. Similarly to *pubmed-electra-base*, the model is fine-tuned on all data and the additionally augmented, the annotations are made by the same version of BioBERT. Biobert-PubMed is the most consistent model on the development and test sets, suggesting better generalization.

- *numind/NuNER_Zero* [5] - Generalist Model for NER using Bidirectional Transformer (GLiNER). This is a lightweight zero-shot NER model based on the GLiNER architecture. It is fine-tuned on all data excluding the Bronze collection. We refer to it as **GLiNER**. The model is not achieving the best performance on the dev set, but has the highest micro precision and recall, therefore F1-micro score on the test set.

4.2. Ensembles

Leveraging the power of a subset of systems, the most successful ensemble approach on the test set turns out to be, this time matching the expectations, with the best models on the development set. In particular, the highest precision and recall system includes *monologg/biobert_v1.1_pubmed*, *pubmed-electra-base*, GLiNER, *monologg/biobert_v1.1_pubmed* and another version of *pubmed-electra-base* fine-tuned on all data and the expanded Bronze corpus, but by changing the annotator model. It is referred in Table 3 as *microsoft/BiomedNLP-BiomedELECTRA-base-uncased (v2).* In this case, the additional data is annotated by *pubmed-electra-base* and GLiNER. In total, 5 different models are used. The strategy to form this ensemble is two-fold. First, error analysis on the development set shows per-label performance for each model. Then, on the basis of this, hand-crafted rules are prepared to take into account only the predictions for the selected categories during testing. For example, GLiNER is used only for *gene and chemical.* The whole model composition is shown on Table 3.

Table 3
Ensemble Model - parts.

Model name	Categories
monologg/biobert_v1.1_pubmed	bacteria
microsoft/BiomedNLP-BiomedELECTRA-base-uncased	statistical technique
numind/NuNER_Zero	gene, chemical
dmis-lab/biobert-base-cased-v1.1	human
microsoft/BiomedNLP-BiomedELECTRA-base-uncased (v2)	anatomical location, animal, bacteria, biomedical technique, DDF, dietary supplement, drug, food, microbiome

5. NER Results

Table 4 lists the results of our models on both the development and test sets for micro precision. The second column shows which subset of the data is used for fine-tuning to achieve the results, including the model split for the ensemble. The ensemble approach discussed in the previous section, Subsection 4.2, improves on precision by nearly 10% compared to GLiNER on the development set. However, on the test set GLiNER is the only model out of all to perform better.

Similarly to the precision table, Table 5 shows once again the better performance of our best models on the development set, this time by micro recall. Although some of the other fine-tuned versions of pubmed-electra-base and biobert-base-v1.1 are not too far behind, the ensemble we form turns out to account for the biggest number of categories according to the authors annotations. In contrast, on the test set, the micro recall of all systems drops. The biggest drop is in pubmed-electra-base, while the most consistent model remains GLiNER.

Finally, Table 6 shows the F1-micro score. This metric is used to rank the systems in the official leaderboard[3]. GLiNER has the lowest score on the development set, because of its weak precision compared to the other systems. On the test set, all of the systems show decreasing performance, except for GLiNER. It is the only model to generalize well and improve on its scores, making it our most successful submission.

Before the final submission, progressive fine-tuning was applied to all of the selected models. Surprisingly, after including the development set and training for a small number of epochs, usually 3-5, the performance decreases on average by around 5% on F1-micro, including the ensemble using these models. In conclusion, lighter models indeed prove to be more beneficial within the context of this task and GLiNER is especially good for NER with limited amount of data.

Table 4
NER Micro-Precision on development and test sets. Here, *all data* means *gold + platinum + silver + bronze* collections; *augmented-biobert* means expanded bronze corpus annotated with BioBERT.

Model	Data Split	Dev P	Test P
ensemble	best-models-on-dev	**84.26%**	78.99%
GLiNER	gold+platinum+silver	75.90%	**80.65%**
biobert-base-v1.1	bronze	81.75%	78.59%
pubmed-electra-base	all data + augmented-biobert	82.60%	78.86%
biobert-pubmed	all data + augmented-biobert	80.99%	78.59%

Table 5
NER Micro-Recall on development and test sets. Here, *all data* means *gold + platinum + silver + bronze* collections; *augmented-biobert* means expanded bronze corpus annotated with BioBERT.

Model	Data Split	Dev R	Test R
ensemble	best-models-on-dev	**82.90%**	76.31%
GLiNER	gold+platinum+silver	82.63%	**79.54%**
biobert-base-v1.1	bronze corpus	81.83%	79.22%
pubmed-electra-base	all data + augmented-biobert	82.45%	74.86%
biobert-pubmed	all data + augmented-biobert	82.01%	79.22%

Table 6
NER Micro-F1 on development and test sets. Here, *all data* means *gold + platinum + silver + bronze* collections; *augmented-biobert* means expanded bronze corpus annotated with BioBERT.

Model	Data Split	Dev F1	Test F1
ensemble	best-models-on-dev	**83.57%**	77.63%
GLiNER	gold+platinum+silver	79.13%	**80.10%**
biobert-base-v1.1	bronze corpus	81.79%	78.91%
pubmed-electra-base	all data + augmented-biobert	81.69%	77.23%
biobert-pubmed	all data + augmented-biobert	81.49%	78.91%

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

Table 7
Hyperparameters of the best-performing NER models with the respective training collection in brackets. Here, *all data* means *gold + platinum + silver + bronze* collections; *augmented-biobert* means expanded bronze corpus annotated with BioBERT.

model (data split)	batch size	max sequence len	epochs	lr	optimizer
GLiNER (gold+platinum+silver)	8	384	20	5e-5	AdamW
biobert-base-v1.1 (bronze corpus)	16	128	10	5e-5	AdamW
pubmed-electra-base (all data + augmented-biobert)	32	512	8	5e-5	AdamW
biobert-pubmed (all data + augmented-biobert)	16	512	10	1e-5	AdamW

The error analysis on the dev set, shows that that only 15% of the false positives are not entities of interest. Obviously the model is good enough in guessing the mentions in the text but not so good in guessing their type. Also the model does not learn very well the annotation guidelines in terms of nested entities - e.g. it extracts gene mentions such as "16 S rDNA gene" which are part of a longer mention "amplification and sequencing of the V4 region of the 16 S rDNA gene" which is a biomedical technique in the gold standard. There are also cases of extracting entities which are true positives in our view but are missing in the gold standard such as "discriminant taxa analyses". Although NER is a classical NLP task with long history and significant progress, this problem remains challenging in its detail.

6. Relation Extraction Approaches

6.1. Babelscape/rebel-large

One of the applied strategies for RE is to use an autoregressive transformer-based approach to extract the entities in relation within a document. For this purpose, Babelscape/rebel-large is employed. REBEL [9] is a seq2seq encoder-decoder model that uses BART-large as a base and solves the RE problem as an end-to-end language generation. This architecture is commonly used in machine translation, where given an input language, the task is to produce an output ("translation") into a target one. In the context of the GutBrainIE challenge, the input is the raw text containing the entities and their implicit relations and the target consists of linearized triplets, explicitly showing the entity mentions and their relations. A triplet follows similar structure:

```
<triplet> head-entity-mention <subj> type-relation <rel> tail-entity-mention
```

At this stage, the goal is to generate the target linearized triplets as accurately as possible from the input text.

What makes REBEL a good choice are its pretraining methods. The authors of the algorithm solved the problem with the scarce RE datasets by creating a new one. This so-called silver corpus consists of Wikipedia abstracts where Wikipedia hyperlinks are matched with WikiData entities. From these, all the present relations are extracted. In total, REBEL is able to find 220 relation types based on the data it has been trained on.

Here, REBEL is fine-tuned to solve all RE-related subtasks. Because of the nature of the model, first all entity mentions and their corresponding relations are extracted. To make the decoding fit our criteria for the other subtasks, the target is modified to also include the entity types along with the mentions. Therefore, the training linear triplets are in the following format:

```
<triplet> head-entity-mention %head-entity-type% <subj> type-relation <rel>
                   tail-entity-mention %tail-entity-type%
```

As a last postprocessing step, also aiming to improve precision, the pre-defined by the organisers rules for domain and range of the predicates are applied. Therefore, only relations that are valid according to the annotation guidelines [4] are selected as the final REBEL output.

To obtain the most successful version of REBEL, all annotated collections (excluding the development set) are used during fine-tuning. The most efficient combination of hyperparameters is with a batch size of 16 and a learning rate of 5e-5. Furthermore, the maximum length of the generated output during decoding is limited to 512 tokens with the number of beams for beam search set to 5. The model is trained for just 100 iterations on a single RTX 6000 ADA Generation GPU, suggesting strong potential for the encoder-decoder architecture in solving RE-related tasks. The training time was about 2 hours. During it, the model with the best F1-micro score on the development set was selected. Similarly, other experiments with different combinations of input data and hyperparameters showed decreasing performance.

6.2. Adaptive Thresholding and Localized Context Pooling

The most successful approach for relation extraction proved to be the use of Adaptive Thresholding and Localized Context Pooling (ATLOP) [10]. The ATLOP method uses a standard encoder transformer [21], [17] model as its base model. The method requires the extracted entities to be provided to the model, so it relies on the extraction from the NER model to be able to produce a prediction, unlike REBEL, which is an end-to-end entity linking approach. The overall approach of ATLOP is given the input text $d = [x_t]_{t=1}^{l}$ and a set of entities $\{e_i\}_{i=1}^{n}$, which mark the position of entity mentions by inserting a special symbol "*" at the start and end of mentions. After giving the input to a pretrained encoder model, we obtain $\boldsymbol{H} = [\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_l] = \text{BERT}([x_1, x_2, ..., x_l])$. representing the contextualized token embeddings. The embeddings of "*" at the start of the mentions are used to represent the entities, then all the embeddings of the mention are pooled using logsumexp pooling [22]. After the logsumexp pooled embeddings $(\boldsymbol{h}_{e_s}, \boldsymbol{h}_{e_o})$ of an entity pair e_s, e_o are obtained, the entities are mapped to hidden states $\boldsymbol{z}$ with a linear layers followed by tanh activations, then the probability of relation r by bilinear function and sigmoid activation is calculated. This process is formulated as:

$$\boldsymbol{z}_s = \tanh\left(\boldsymbol{W}_s \boldsymbol{h}_{e_s}\right), \tag{1}$$

$$\boldsymbol{z}_o = \tanh\left(\boldsymbol{W}_o \boldsymbol{h}_{e_o}\right), \tag{2}$$

$$\mathrm{P}\left(r \mid e_s, e_o\right) = \sigma\left(\boldsymbol{z}_s^{\top} \boldsymbol{W}_r \boldsymbol{z}_o + b_r\right),$$

ATLOP introduces the "Adaptive Thresholding" (AT) mechanism to address the limitations of global thresholding in the multi-label scenario. This technique replaces the static global threshold with a learnable, entity-dependent threshold, aiming to reduce decision errors during inference. The core of Adaptive Thresholding is the introduction of a special "threshold class," denoted as TH. This TH class is treated and learned similarly to other actual relation classes within the model's architecture. Its purpose is to act as a learned decision boundary. The main benefit of AT is that, instead of using the usual approach of optimizing a value in the range $(0, 1)$ and picking the one that maximizes the evaluation metrics, AT provides a way to learn the optimal threshold during training. Those techniques, plus the use of "Localized Context Pooling" and some other innovations, make ATLOP an efficient and easy-to-use.

To try to improve the model's performance, different pretrained models are used, such as: BERT [17], RoBERTa [23], XLM-R [24], PubMedBERT [8], BiomedELECTRA [7], as base encoder models. Another experiment was done with domain adaptation by continuing the models' pretraining using the masked language modeling objective [17], and task pretraining. The approach for task pretraining is to first fine-tune the base model on the NER objective before plugging the base encoder model into the ATLOP method for further fine-tuning. The training hyperparameters used for fine-tuning are listed in Table 8. While evaluating on the dev set, most models reached their best performance at around the 100th epoch. The models sent for submission are trained on both the train and dev sets. For the test, some submissions are made with intermediate versions of the models before they finish the 200th epoch and with the 200th epoch.

Table 8
ATLOP training hyperparameters.

Hyperparameter	Value
batch size	4
# epochs	200
lr for encoder	5e-5
lr for classifier	1e-4
warmup ratio	6%
max grad norm	1
optimizer	AdamW

7. Relation Extraction Results

7.1. Subtask 6.2.1 - BT-RE

For the BT-RE subtask, results on the dev set are shown in Table 9, demonstrating that ATLOP is a very effective method for relation extraction. While REBEL is a good end-to-end alternative, performance is slightly worse than ATLOP models, which suggests that there is a strong potential for encoder-decoder or decoder-only models for relation extraction. Another observation is that task pretraining gives improvement for BiomedELECTRA, but it isn't very significant. Here, all ATLOP models use NER predictions from the baseline GLiNER model.

Table 9
BT-RE - model performance on dev set. Here, "+ DA" suffix means that the model went through an additional pretraining phase with masked language modeling before being fine-tuned on the specific task; "+ TP" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

Model	Dev Micro-P	Dev Micro-R	Dev Micro-F1
rebel-large	62.79%	55.56%	61.50%
ATLOP + BERT base	67.50%	61.36%	61.03%
ATLOP + RoBERTa base	67.54%	58.64%	62.77%
ATLOP + XLM-R base	67.50%	61.36%	64.29%
ATLOP + XLM-R base + DA	64.25%	**64.55%**	64.44%
ATLOP + XLM-R base + TP	70.11%	58.64%	63.86%
ATLOP + BiomedELECTRA	62.26%	60.00%	64.71%
ATLOP + BiomedELECTRA + TP	**72.58%**	61.36%	**66.50%**

The results on the test set are shown in Table 10. While XLM-R and REBEL do not fall too much behind BiomedELECTRA, BiomedELECTRA, as a base model with task pretraining, is the best-performing model. Here again, all ATLOP models use NER predictions from the baseline GLiNER model.

Table 10
BT-RE - model performance on the test set. Here, "+ TP" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

Model	Test Micro-P	Test Micro-R	Test Micro-F1
rebel-large	68.59%	56.71%	62.10%
ATLOP + XLM-R base (100 epochs)	70.62%	54.11%	61.28%
ATLOP + BiomedELECTRA + TP (100 epochs)	**74.17%**	**65.38%**	**63.86%**
ATLOP + BiomedELECTRA + TP (200 epochs)	72.37%	47.61%	57.44%

7.2. Subtask 6.2.2 - TT-RE

The dev and test sets results for TT-RE are shown in Table 11 and Table 12, respectively. Here, the figures are similar to the ones from the previous task, with the main difference being that on the test set this time "ATLOP + BiomedELECTRA + TP" trained with 200 epochs performs slightly better compared to the one trained with 100 epochs. Here, all ATLOP models use the NER predictions from the baseline GLiNER model.

Table 11
TT-RE - model performance on dev set. Here, "+ DA" suffix means that the model went through an additional pretraining phase with masked language modeling before being fine-tuned on the specific task; "+ TP" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

Model	Dev Micro-P	Dev Micro-R	Dev Micro-F1
rebel-large	63.08%	58.70%	60.81%
ATLOP + BERT base	69.36%	52.17%	59.55%
ATLOP + RoBERTa base	67.36%	56.52%	61.47%
ATLOP + XLM-R base	67.00%	59.13%	62.82%
ATLOP + XLM-R base + DA	63.39%	**61.74%**	62.56%
ATLOP + XLM-R base + TP	70.05%	56.96%	62.83%
ATLOP + BiomedELECTRA	64.90%	58.70%	61.64%
ATLOP + BiomedELECTRA + TP	**72.34%**	59.13%	**65.07%**

Table 12
TT-RE - model performance on the test set. Here, "+ TP" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

Model	Test Micro-P	Test Micro-R	Test Micro-F1
rebel-large	68.88%	55.56%	61.50%
ATLOP + XLM-R base (100 epochs)	68.37%	55.14%	61.05%
ATLOP + BiomedELECTRA + TP (100 epochs)	71.98%	53.90%	61.65%
ATLOP + BiomedELECTRA + TP (200 epochs)	**73.26%**	**56.38%**	**63.72%**

7.3. Subtask 6.2.3 - Ternary Mention-based Relation Extraction

For the final subtask, Ternary Mention-based Relation Extraction (TM-RE), the results are again fairly similar. The results on the dev set can be found in Table 13 and again demonstrate that ATLOP with BiomedELECTRA as a base model is the best performing model, with XLM-R being slightly behind and rebel-large showing strong performance while being an end-to-end model. Here, all ATLOP models use the NER predictions from the baseline GLiNER model.

The results on the test set are shown in Table 14. This time XLM-R slightly outperforms BiomedELECTRA, and *rebel-large* is not much behind. Here, is observed significant degradation in performance compared to the previous tasks. This is partially due to the accumulation of errors following each task, as the final task contains each of the other subtasks in itself. This could also partially be due to the fact that the search space becomes much larger with the final subtask. Let N be the input length (number of tokens). Let S be the number of possible spans (ordered contiguous sub-sequences of tokens) in an input of length N. Then the number of such spans is $S = \frac{N(N+1)}{2}$. Let the number of possible classes for entities be C classes. Then, for the NER subtask, the search space is C^N, which is exponential with respect to the input length. For the relation extraction subtask, the number of possible relations is C^2 assuming that each entity can be in relation to each other since all relations could be in the document or none of them the total search space is 2^{C^2}. Which in our case is much smaller because the number of possible types is much smaller than the number of tokens in a document. But we still see some

Table 13
TM-RE - model performance on dev set. Here, "+ DA" suffix means that the model went through an additional pretraining phase with masked language modeling before being fine-tuned on the specific task. And "+ TP" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

Model	Dev Micro-P	Dev Micro-R	Dev Micro-F1
rebel-large	38.81%	36.25%	37.49%
ATLOP + BERT base	47.55%	32.86%	38.86%
ATLOP + RoBERTa base	46.44%	33.75%	39.09%
ATLOP + XLM-R base	46.04%	**37.32%**	41.22%
ATLOP + XLM-R base + DA	42.65%	**37.32%**	39.81%
ATLOP + XLM-R base + TP	45.52%	34.46%	39.23%
ATLOP + BiomedELECTRA	43.33%	37.14%	40.00%
ATLOP + BiomedELECTRA + TP	**48.56%**	36.25%	**41.51%**

performance degradation for this subtask. Some of the models, like ATLOP, require the predictions of the NER subtask, which explains the degradation of the performance. REBEL is a generative model, so technically it needs to search through an infinite search space to generate the answer, but potentially instead of generating the answer one could take take the averaged token log probabilities by giving each answer to the model, and could achieve better performance, but this was not done in time for the competition. For the TM-RE subtask, the search space increases to 2^{RC^2} where R is the number of possible relation types, assuming that all relation types and all entities are possible. This is not a significant increase in the search space, which is also shown in the results as the performance for this subtask is just a few percent lower than the BT-RE subtask. For the TM-RE, the search space increases to $(RC^2+1)^{S^2} = (RC^2+1)^{(\frac{N(N+1)}{2})^2}$, assuming that each entity can be in relation with each other and with every relation type, and assuming there could be an overlap between entities. Although this is an overestimation of the search space, it still shows that for this subtask, the search space is significantly larger, thus at least partially explaining the significantly lower results for the last subtask compared to the other subtasks. Here, all ATLOP models use the NER predictions from the baseline GLiNER model.

Table 14
TM-RE - model performance on the test set. Here, "+ TP" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

Model	Test Micro-P	Test Micro-R	Test Micro-F1
rebel-large	43.36%	31.10%	36.22%
ATLOP + XLM-R base (100 epochs)	**46.86%**	**30.96%**	**37.28%**
ATLOP + BiomedELECTRA + TP (100 epochs)	44.15%	26.80%	33.36%
ATLOP + BiomedELECTRA + TP (200 epochs)	46.12%	29.49%	35.98%

7.4. Hybrid Systems

For subtasks 6.2.1, 6.2.2 and 6.2.3, a hybrid system with another participant in the challenge - CLEANR[4] [25]) is proposed. The results are combined by taking the union and intersection of both solutions to combine the systems' strengths. The results of the union achieve higher results than any of the systems alone on subtasks 6.2.1 and 6.2.2. On subtask 6.2.3 our system alone demonstrates better capabilities. This means that both systems are complementary. Results are shown on Table 15 and Table 16.

CLEANR utilizes RAG as its approach to incorporate detailed training data through semantic retrieval processes in the prompt for the language model (LM). This few-shot approach, combined with dynamic retrieval, enables the system to be extended or "retrained" by simply adding or reweighting the training

[4] https://github.com/Dakantz/CLEANR

Table 15
Results on the dev set obtained by the hybrid system.

Task	System description	Micro-Precision	Micro-Recall	Micro-Recall
T6.2.1	intersection	0.89	0.11	0.19
T6.2.1	union	0.69	0.65	0.67
T6.2.2	union	0.69	0.62	0.65
T6.2.3	union	0.37	0.38	0.38

Table 16
Results on the test set obtained by the hybrid system.

Task	System description	Micro-Precision	Micro-Recall	Micro-Recall
T6.2.1	intersection	0.89	0.13	0.23
T6.2.1	union	0.71	0.61	0.66
T6.2.2	union	0.71	0.59	0.64
T6.2.3	union	0.35	0.32	0.34

samples. CLEANR extends the approach by introducing a reweighting of the samples in the retrieval process to prefer samples with a higher degree of confidence (i.e., prefer the Gold annotations over the Bronze annotations in our setting). A sentence-transformer system is used to embed the given training samples and store them in a Postgres database using the pgvector extension. CLEANR utilizes llama-cpp and llama-cpp-agent for both efficient inference of pre-trained models and constrained generation from a provided grammar. The grammar is generated using dynamically created models, which are transformed into the GBNF syntax, which is then used to constrain the LM output to the exact schema provided by the challenge. CLEANR's output was combined with the results from *ATLOP + BiomedELECTRA + TP (100 epochs).*

8. Discussion & Conclusion

This challenge is of major importance for bootstrapping the development of tools for automated analysis of gut-brain related literature and therefore facilitating the research in the area. The provided data is of good quantity as a starting point for fine-tuning deep learning models, however, some categories of named entities are underrepresented and need alternative approaches. E.g. for genes extraction, gazetteers may increase recall but then many efforts to remove noise are also necessary. And, of course, more annotated data would potentially help for better training of these models.

Our takeaways from the NER task are two: (i) GLiNER deserves more attention, it is our top performing system on the NER task, it shows best capabilities to generalize and it is worth exploring more in depth; (ii) progressive training of the models on the dev set is not showing effective results on the test set - all these models degrade their performance on the test set in comparison with the dev set. This could also mean that the dev and train set are not really close in means of annotation agreement.

For the relation extraction task and ATLOP especially, we come to the conclusion that the model is very sensitive to the initial set of named entities it works on. If the model is provided with more entities than the ones entering in relations then the performance heavily drops. E.g. the ATLOP results are worse when the golden named entities are provided as input than when working on the GLiNER entities. Therefore initial pre-processing of the supplied named entities improves the results.

The hybrid system in which our RE model was part of outperformed our own system which means that the results of both systems are complementary and not overlapping, therefore combing a transformer-based language model with an LLM-based approach seems to be promising for further research.

Acknowledgments

This work is partially supported by European Union's Horizon research and innovation programme projects RES-Q PLUS [Grant Agreement No. 101057603] and HEREDITARY [Grant Agreement No. 101137074]. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] T. S. Dong, E. Mayer, Advances in brain–gut–microbiome interactions: a comprehensive update on signaling mechanisms, disorders, and therapeutic implications, Cellular and molecular gastroenterology and hepatology 18 (2024) 1–13.

[2] M. H. Mohajeri, G. La Fata, R. E. Steinert, P. Weber, Relationship between the gut microbiome and brain function, Nutrition reviews 76 (2018) 481–496.

[3] 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. Maria 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, in: J. Carrillo-de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[4] 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.

[5] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist model for named entity recognition using bidirectional transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300/. doi:10.18653/v1/2024.naacl-long.300.

[6] 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.

[7] R. Tinn, H. Cheng, Y. Gu, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Fine-tuning large neural language models for biomedical natural language processing, 2021. URL: https://arxiv.org/abs/2112.07869. doi:10.48550/ARXIV.2112.07869.

[8] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2020. arXiv:arXiv:2007.15779.

[9] P.-L. Huguet Cabot, R. Navigli, REBEL: Relation extraction by end-to-end language generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 2370–2381. URL: https://aclanthology.org/2021.findings-emnlp.204.

[10] W. Zhou, K. Huang, T. Ma, J. Huang, Document-level relation extraction with adaptive thresholding and localized context pooling, 2020. URL: https://arxiv.org/abs/2010.11304. arXiv:2010.11304.

[11] N. Goyal, N. Singh, Named entity recognition and relationship extraction for biomedical text: A

comprehensive survey, recent advancements, and future research directions, Neurocomputing (2024) 129171.

[12] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, Biored: a rich biomedical relation extraction dataset, Briefings in Bioinformatics 23 (2022) bbac282.

[13] A. Névéol, R. I. Doğan, Z. Lu, Semi-automatic semantic annotation of pubmed queries: a study on quality, efficiency, satisfaction, Journal of biomedical informatics 44 (2011) 310–318.

[14] M. Sänger, U. Leser, Large-scale entity representation learning for biomedical relationship extraction, Bioinformatics 37 (2021) 236–242.

[15] N. A. A. Hassan, R. A. A. A. A. Seoud, D. A. Salem, Open information extraction methodology for a new curated biomedical literature dataset, International Journal of Advanced Computer Science and Applications 14 (2023).

[16] W. Zhou, K. Huang, T. Ma, J. Huang, Document-level relation extraction with adaptive thresholding and localized context pooling, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 14612–14620. doi:10.1609/aaai.v35i16.17717.

[17] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL: https://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[18] M.-T. Clark, Kevin andLuong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators (2020).

[19] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107/.

[20] M. Neumann, D. King, I. Beltagy, W. Ammar, Scispacy: Fast and robust models for biomedical natural language processing, in: Proceedings of the 18th BioNLP Workshop and Shared Task (2019) 319-327, 2019. URL: https://arXiv:1902.07669.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[22] R. Jia, C. Wong, H. Poon, Document-level n-ary relation extraction with multiscale representation learning, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3693–3704. URL: https://aclanthology.org/N19-1370/. doi:10.18653/v1/N19-1370.

[23] 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, 2019. URL: https://arxiv.org/abs/1907.11692. arXiv:1907.11692.

[24] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, 2020. URL: https://arxiv.org/abs/1911.02116. arXiv:1911.02116.

[25] B. Kantz, P. Waldert, S. Lengauer, T. Schreck, Constrained linked entity annotation using rag (cleanr), in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS, 2025.