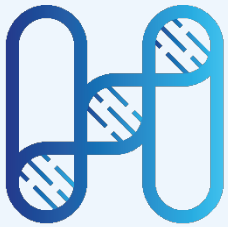

HEREDITARY

HetERogeneous sEmantic Data integratIon for the guT-bRain interplaY

Deliverable D4.9

Evidence-based Knowledge Base Creation: Initial Results

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No GA 101137074. Views and opinions expressed are however those of the author(s) 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.

Funded by the European Union

EXECUTIVE SUMMARY

This deliverable presents the initial results achieved during the first 12 months of task T4.6 "*Evidence-based knowledge graph creation and exploration*", of Work package 4 "*Multimodal Analytics and Learning Platform*" of the HEREDITARY project.

The work conducted during the first period of research includes a review of state-of-the-art approaches and the development of novel unsupervised and supervised models for extracting useful information about gut-brain interplay from biomedical literature, which can provide valuable insights for healthcare. This report presents initial results from the development of methods for information extraction from biomedical literature (named entity recognition, relation extraction) and methods for building Knowledge Bases on top of the extracted entities and relations.

The report presents a description of three main results: PubMiner AI[1] and solutions developed by the consortium for the GutBrainIE Task at CLEF 2025 BioASQ Lab, organized by the HEREDITARY project as part of Task 5.6. The main source of the data used for the development of the proposed approaches is PubMed citations. PubMiner AI provides an end-to-end solution based on large language models (LLMs) that covers the full spectrum of functionalities - information extractions (NER and RE) and transformation into Knowledge Graphs (KGs). The solutions developed for GutBrainIE Task at CLEF 2025 BioASQ Lab, provide approaches for NER and RE from PubMed articles covering 13 categories of entities and 25 relation types with focus on the Gut-Brain axis. The evaluation of the proposed models demonstrates their capabilities to extract gut-brain interplay-related information from PubMed articles. The organised challenge (that will be fully described and reported in Deliverable 5.10 due at Month 24) and the developed tools will bootstrap further research in the area.

[1] https://marketplace.databricks.com/details/6b8ca36d-53c3-4506-bb90-e9a803fb01d6/Ontotext_PubMiner-AI-Automatic-Knowledge-Extraction-from-Scientific-Publications

DOCUMENT INFORMATION

Deliverable ID	D4.9
Deliverable Title	Evidence-based Knowledge Base Creation: Initial Results
Work Package	WP4
Lead Partner	ONTO
Due date	30.06.2025
Date of submission	25.06.2025
Type of deliverable	R
Dissemination level	PU

AUTHORS

Name	Organisation
Ivelina Nikolova-Koleva (Author)	ONTO
Svetla Boytcheva (Author)	ONTO
Aleksis Datseris (Author)	ONTO
Benedikt Kantz (Author)	TUGRAZ
Juan Manuel Rodriguez (Author)	AAU
Chiara Lovati (Author)	OBSERVA
Rute Costa (Author)	UNL
Margarida Ramos (Author)	UNL
Gianmaria Silvello (Internal Reviewer)	UNIPD
Anna Romanovych (Contributor)	UNIPD

REVISION HISTORY

Version	Date	Author	Document history/approvals
0.1.0	24.02.2025	Ivelina Nikolova-Koleva	Initial deliverable structure
0.1.0	27.03.2025	Svetla Boytcheva	2.3 Knowledge graph construction (KBC)
0.1.1	04.04.2025	Benedikt Kantz	Add TUGRAZ CLEANR
0.1.2	23.05.2025	Ivelina Nikolova-Koleva	4.1 PubMiner AI
0.1.3	29.05.2025	Rute Costa, Margarida Ramos	2.2.1 Task definition
0.1.4	01.06.2025	Rute Costa, Margarida Ramos	2.2.2 Approaches

Version	Date	Author	Document history/approvals
0.1.5	01.06.2025	Ivelina Nikolova-Koleva	Input for sections 2, 3 and 4
	01.06.2025	Aleksis Datseris	Input for 4.3
	01.06.2025	Juan M Rodriguez	Input for 4.5
0.1.6	04.06.2025	Svetla Boytcheva	Executive summary, Introduction, Conclusion, and review of the overall deliverable content
0.1.7	05.06.2025	Ivelina Nikolova-Koleva	References, proof-reading, final formatting
0.1.8	06.06.2025	Gianmaria Silvello	Internal review
0.2.1	20.06.2025	Ivelina Nikolova-Koleva	Applying reviewer's comments
0.2.2	20.06.2025	Svetla Boytcheva	Applying reviewer's comments
0.2.3	23.06.2025	Gianmaria Silvello	Second internal review.
0.3.1	24.06.2025	Svetla Boytcheva	Applying reviewer's comments
0.3.2	24.06.2025	Ivelina Nikolova	Applying reviewer's comments
1.0	25.06.2025	Anna Romanovych	Final formatting of the deliverable

Contents

List of Tables

List of Figures

Table of Abbreviations

Acronym	Full name
AI	Artificial Intelligence
ALS	Amyotrophic Lateral Sclerosis
bi-LSTM	Bidirectional Long Short-Term Memory Network
CNN	Convolutional Neural Network
CRF	Conditional Random Fields
DDF	Disease-Disorder-Finding
DDI	Drug-Drug Interaction
DL	Deep Learning
EHR	Electronic Health Record
EL	Entity Linking
FAIR	Findable, Accessible, Interoperable, and Reusable
FN	False Negatives
FP	False Positives
GBNF	GGML Backus-Naur Form
GPT	Generative Pre-trained Transformer
HMM	Hidden Markov Model
IE	Information Extraction
IIIA	Intelligent Interactive Information Access
KB	Knowledge Base
KBC	Knowledge Base Creation
KG	Knowledge Graph
LLM	Large Language Model
LM	Language Model
LSTM	Long Short-Term Memory network
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
NSM	Negative Sample Multiplier
PLM	Pre-trained Language Models
RAG	Retrieval Augmented Generation
RAG4RE	Retrieval Augmented Generation for Relation Extraction
RDF	Resource Description Framework (format)
RE	Relation Extraction
SOTA	State of the Art

Acronym	Full name
SVM	Support Vector Machines
TN	True Negatives
TP	True Positives
UMLS	Unified Medical Language System
WP	Work Package

1 Introduction

We introduce approaches for Information Extraction (IE) from biomedical literature and the transformation of this information into Knowledge Graphs (KGs). We survey the key approaches for IE and Knowledge Base Creation (KBC) and present the novel developed components in the scope of HEREDITARY project. These are initial results developed under Task 4.6 "*Evidence-based knowledge graph creation and exploration*".

1.1 Definition of terms

In this deliverable, we consider the following definitions of the terms:

Natural language processing (NLP) is a field of artificial intelligence (AI) that focuses on enabling computers to "understand", interpret, and generate human (natural) language. NLP combines computational linguistics with machine learning (ML), allowing machines to process and analyse large amounts of natural language data (such as texts, speech, etc.).

Information Extraction (IE) in the biomedical domain refers to the automated process of identifying, classifying, and structuring relevant information from unstructured biomedical texts, such as research articles, clinical reports, and patents discharge summaries. This process typically focuses on specific types of information, such as genes, proteins, diseases, drugs, biological pathways, and their relationships. Biomedical IE aims to transform large volumes of textual data into structured representations, such as knowledge databases or knowledge graphs (KG), that support research, clinical decision-making, and knowledge discovery.

Named Entity Recognition (NER) is a sub-task of information extraction in NLP that involves identifying and classifying key elements (entities) in text into predefined categories such as drugs, diseases, genes, chemicals, etc.

Relation Extraction (RE) is a process in NLP that involves identifying and classifying semantic relationships between entities in text, such as mentions of gene–disease associations from a set of scientific papers or identifying drug–drug interactions reported in patient discharge summaries or clinical trials.

Entity linking (EL) in the biomedical domain is the process of automated matching of biomedical concepts mentions — such as named entities from various categories: drugs, diseases, genes, chemicals, etc. — found in unstructured text (like journal articles or clinical notes) to standardized, unique identifiers in biomedical databases or ontologies (such as MeSH[2], UMLS[3], or UniProt[4]). This step ensures that each mention of a biomedical entity in the text is unambiguously associated with its correct concept, enabling consistent interpretation and integration of information across different sources.

Knowledge Base (KB) is a structured collection of information designed to provide answers to specific questions or support decision-making. It typically includes documents, articles, FAQs, manuals, or databases that are organized for easy retrieval.

[2] https://www.ncbi.nlm.nih.gov/mesh/
[3] https://www.nlm.nih.gov/research/umls/index.html
[4] https://www.uniprot.org/

Knowledge bases are often used in customer support, technical documentation, and internal organizational knowledge sharing.

Knowledge Graph (KG) is a network of real-world entities (such as people, places, or concepts) and the relationships between them. It represents knowledge in a graph format, where nodes represent entities and edges represent relationships. Knowledge graphs are used to enable machines to understand context, infer new information, and support intelligent applications like search engines and recommendation systems.

Knowledge Base Creation (KBC) is the process of systematically gathering, organizing, and structuring knowledge (such as facts, relationships, and rules) about a domain, typically in a machine-readable format. Information is often extracted from unstructured sources (like text) and arranged in structured formats, such as databases or knowledge graphs, for easy access and querying by computers. This task also includes entity linking of all extracted entities into standard classifications and ontologies.

Resource Description Framework (RDF) is a widely adopted standard model for expressing information on the semantic web, especially for exchanging metadata and data. This framework uses a knowledge graph-based structure to describe resources and their relationships. As a fundamental part of the Semantic Web, RDF allows data to be presented in a structured format as <*subject*><*predicate*><*object*> triples, promoting integration and interoperability across different systems.

HEREDITARY Task 4.6 definition objectives include:

- Collecting and filtering from PubMed[5] articles relevant to the project use-cases.
- Development of a text processing pipeline: named entity recognition and relation extraction; entity linking to the HEREDITARY ontology HERO[6] (Task 3.1, Deliverable D3.1 - Ontology and federated execution methods (Cazzaro et al. 2024)).
- Building a large probabilistic KG about gut-brain interplay, gene variants, and risk factors.

Task 4.6 role and position in HEREDITARY project. There are close relations, interconnection and alignment of T4.6 with other tasks in WPs in the HEREDITARY project like T2.2, T2.3, T2.4, T2.5, T3.1, T3.3, T3.6, T5.3 and T5.6.

Indeed, all concepts and relations that we aim to extract in Named Entity Recognition (NER) and Relation Extraction (RE) subtasks are well aligned with the HEREDITARY project use-cases and their specification in T2.2 "*Neurodegenerative use cases and data coordination*" (details are available in D2.16 (Chiò,et al. 2024)), T2.3 "*Linkage and feature extraction from gut-brain*" (see D2.3 (Kohn, et al. 2024)), T2.4 "*Clinical value extraction of gut-brain integration*", and T2.5 "*Parkinson's detection on multimodal and multicenter data*" (D2.8 due at Month 18).

Task 4.6 solutions for Knowledge Base Creation (KBC) and Entity Linking (EL) are closely related to HERO ontology (T3.1 "Multimodal semantic ontology" (Cazzaro et al. 2024)). In addition, the NER task is also related to the extraction of novelties and term extraction for gur-brain interplay (T3.3 "*Multilingual semantic network*" (details are given

[5] https://pubmed.ncbi.nlm.nih.gov/
[6] https://hereditary.dei.unipd.it/ontology/

in Deliverable 3.4 (Costa et al. 2024)) that can help in the enrichment of existing terminologies and taxonomies. There is also a strong relationship between the developed solutions in T4.6 and T3.6 "*FAIRification workflows and 1+MG guidelines compliance for unstructured data*". The created knowledge graphs in T4.6 provide a good basis for further exploration of the identified gut-brain relationships. There is already established collaboration with Task 5.3 "*Visual querying and analytical interaction*", initiated on M12, which will be beneficial for the developments in both tasks.

The alignment with T5.6 "*Shared evaluation campaigns*" allows for developing and evaluating solutions in T4.6 over a custom-selected and annotated dataset for PubMed articles on the topic of gut-brain interplay. Such a shared evaluation effort is the GutBrainIE Task at CLEF 2025 BioASQ Lab[7] described below, along with the partner solutions submitted to the challenge (Martinelli et al. 2025; Nentidis et al. 2025).

1.2 Solutions described in the deliverable

In this deliverable, we present two solutions:

- Pubminer AI - provides an end-to-end solution based on GraphRAG technology and large language models (LLMs) that covers the full spectrum of functionalities - information extraction (NER and RE) and transformation into probabilistic knowledge graphs. The description of PubMiner AI is presented in detail in Section 4.1.
- Solutions for the GutBrainIE Task at CLEF 2025 BioASQ Lab - systems developed by the consortium partners ONTO, TGRAZ, AAU as well as collaborative solutions for the GutBrainIE Task at CLEF 2025 BioASQ Lab, organized by the HEREDITARY project[8]. All solutions are presented in Sections 4.2, 4.3, 4.4, and 4.5.

1.3 Outline

The presentation in this deliverable is structured as follows: Section 2 overviews the state-of-the art approaches for named entity recognition, relation extraction, and knowledge base creation; Section 3 briefly describes some initial data about the available annotated datasets developed in T5.6 for what is related to KBC; Section 4 presents in detail the developed solutions, their evaluation and brief discussion on the results; and Section 5 summarizes in conclusion the presented research results and sketches directions for further work.

[7] GutBrainIE @ CLEF 2025 Guidelines - Gut Brain Information Extraction
[8] The challenge will be described in D5.10 "First evaluation challenge: report on the data, results, and integration with EOSC" due at M24.

2 Related work

2.1 Named entity recognition

2.1.1 Task definition

In the biomedical domain the Named Entity Recognition (NER) task is usually related to the identification of biomedical terms and related entities such as gene and protein identifiers, chemicals, anatomical structures, diseases, clinical findings, and more. The process involves two steps – recognising the scope of the entity in the text and then classifying the given text span into a predefined semantic category. The accurate recognition of drug names, symptoms, diseases, and treatments enhances various applications, such as aggregating vital clinical data, crucial for patient care and medical research; improving automatic structuring of electronic health record (EHR) data; collecting observations related to medications, indications, and adverse drug reactions; automatically mining and summarizing information from scientific literature; and pseudonymizing clinical documents to protect patient privacy while enabling research use. In biomedical research, NER plays an essential role by enabling automated tracking of new developments, which is otherwise challenging due to the rapid pace of publication of new scientific studies.

There are some specific challenges in the biomedical domain that NER faces. Annotated data is costly because it involves expert labour therefore it is difficult to train complex deep learning models which require bigger volumes of data. The use of technical jargon, abbreviations, ambiguous language, and frequent updates in scientific discoveries reflects the variability of the text. The entity boundaries are not clear and often there are nested entities (Park et al. 2024). Scientific literature in this domain is characterised with detailed in lengthy sentences where entities may contain combinations of letters, symbols, and punctuation.

2.1.2 Approaches

The approaches to NER can be summarized as shown on the Figure 1 (Keraghel et al. 2024).

The successful systems for NER are usually hybrid systems using more than one technology. Here are outlined the three basic approaches:

1) Knowledge-based methods,
2) Feature engineering-based methods, and
3) Deep learning-based methods.

Here, more attention is paid to the latter one, because most of the systems developed under T4.6 and reported hereafter rely on deep learning-based methods.

Knowledge-based methods

Knowledge-based methods have been a foundational part of NER, particularly in the early stages of the field. These methods do not require annotated data and are primarily based on rules and gazetteers. Gazetteers serve as essential resources, providing a collection of domain-specific entities that improve recognition performance. Architectures generally involve three key components: (1) a set of rules for identifying named entities,

(2) optional gazetteers for additional context, and (3) an extraction engine that applies the rules and gazetteers to the input text. GATE (Cunningham et al. 2013) and cTAKES (Savova et al. 2010) systems are among the most recognised platforms for developing biomedical knowledge-based NLP applications.

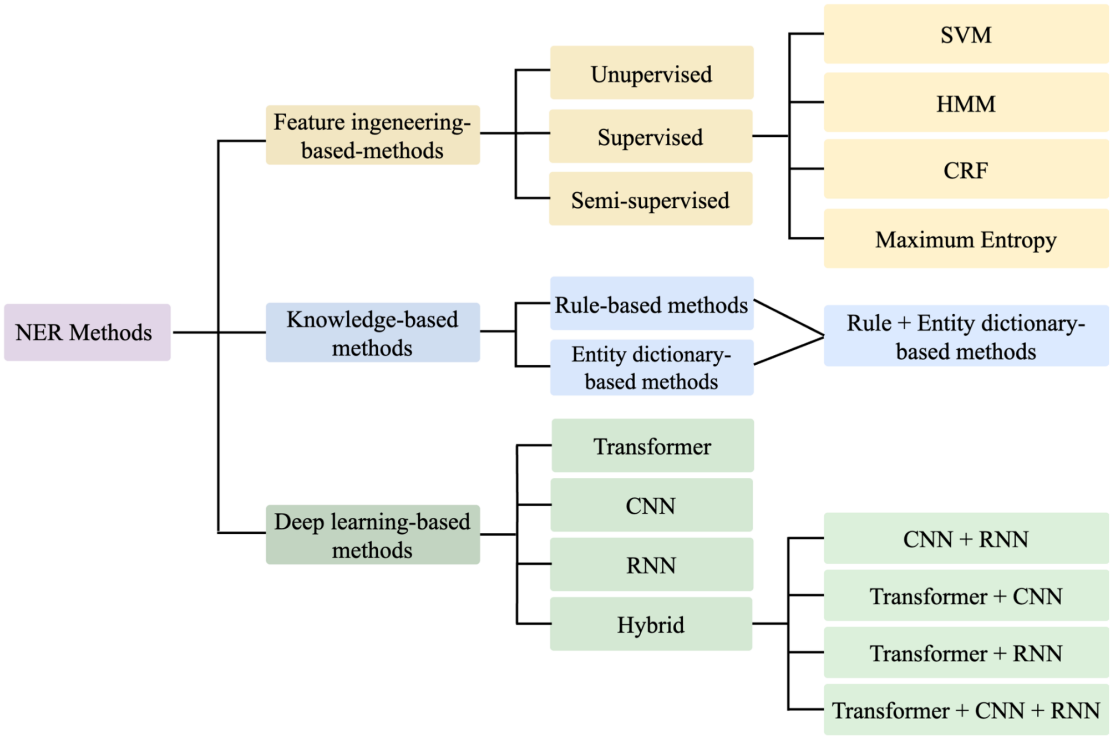

Figure 1. NER approaches (Keraghel et al. 2024)

One of the key applications of knowledge-based methods for biomedical NLP is the MetaMap system (Aronson 2001). MetaMap is developed at the National Library of Medicine to map biomedical text to the UMLS Metathesaurus concepts (Bodenreider 2004). MetaMap used a knowledge-intensive approach based on symbolic, natural language processing, and computational linguistics techniques. It keeps upgrading and has a much more sophisticated architecture nowadays. Another tool worth mentioning is ConceptMapper (Tanenblatt et al. 2010), a tool that was not specifically developed for biomedical term recognition but rather for flexible look-up of terms from a controlled vocabulary. Also, a well-known system is the NCBO's Open Biomedical Annotator9 (Shah et al. 2009), based on a tool from the University of Michigan called mgrep.

ML-based methods

With the evolution of NER and the need to automate the extraction of named entities, the development of methods based on feature engineering arises. These methods switch the focus from manual rule elicitation to identifying key features that could be used for NER tasks within the text. Feature-engineering-based methods can be broadly classified into unsupervised, supervised, and semi-supervised learning methods, each with its advantages and limitations. The most prominent models for supervised learning which have been developed for NER tasks include the Hidden Markov Model (HMM) (Baum et

[9] https://bioportal.bioontology.org/annotatorplus

al. 1970), the Maximum Entropy model (MaxEnt) (Berger et al. 1996), Support Vector Machines (SVM) (Cortes and Vapnik 1995), Conditional Random Fields (CRF) (Lafferty et al. 2001; McCallum and Li 2003).

Deep-learning-based methods

The rise of deep learning has significantly improved the accuracy and flexibility of named entity recognition systems. Methods like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) learn effective representations of entities from large datasets. The introduction of neural probabilistic language models by (Bengio et al. 2023) marked the beginning of deep learning in NLP, demonstrating the power of distributed word representations. Later, in (Collobert 2011) deep learning was applied to NER, showing the potential of CNNs for various NLP tasks.

The **transformer model**, introduced by (Vaswani et al. 2017), revolutionized NLP tasks, particularly named entity recognition, by capturing complex contextual relationships. Transformer-based models like BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), and DeBERTa (He et al. 2020) have achieved state-of-the-art results in NER due to their ability to produce richly contextualized embeddings. BERT derivatives, such as DistilBERT (Sanh et al. 2019) and RoBERTa, have also shown strong performance, with DistilBERT being a smaller and faster variant, and RoBERTa generating deeper contextual embeddings.

Other notable approaches include PLTR (Wang et al. 2023), which uses pre-trained models like RoBERTa to generate contextual embeddings, and TDMS-IE (Hou et al. 2019), which automates information extraction from scientific papers using BERT. The Generalist Model for NER, GLiNER (Zaratiana et al. 2024), integrates global contextual information into NER and has shown promising results. It does not require predefined categories and leverages models like BERT and DeBERTa for efficient processing. It excels in zero-shot scenarios, outperforming general-purpose models and offering flexibility without demanding extensive resources. However, for domain-specific tasks like biomedical text mining, models such as BioBERT, initialized with BERT weights and further pretrained on PubMed abstracts and full-text articles, set new benchmarks for biomedical NER and relation extraction.

While these models have achieved excellent performance, they are computationally intensive. To address this, more efficient models like DistilBERT have been developed. Recent advances also include the development of multilingual models like mBERT (Devlin et al. 2018), which can handle multiple languages and be fine-tuned for specific languages, enhancing performance in low-resource settings.

The top F1 scores achieved for Named Entity Recognition on PubMed articles reach up to 89.3%. In (Luo et al. 2022) BiLSTM-CRF, BioBERT-CRF, PubMedBERT-CRF are applied for NER on over 600 PubMed abstracts and achieve F1-score: 89.3% (strict) and 93.5% (relaxed). More recently, another team (Fabregat et al. 2023) proposed a range of architectures for detecting named entities using negation-based transfer learning techniques based on BiLSTM and CRF. Negation detection is utilized as a preliminary training phase, transferring weights associated with negation triggers and scopes, leading to enhancements in both named entity recognition and relation extraction tasks.

In the context of these approaches follow the partners solutions described in Section 4.

2.1.4 Evaluation metrics

The performance of NER systems is most often evaluated in terms of precision (P), recall (R), and their harmonic mean F1-score (F1), which are a direct result of the number of true positive (TP), false positive (FP), true negative (TN) and false negative (FN) examples. The precision shows how many of the extracted entities are correctly recognised, or $\frac{TP}{TP+FP}$, the recall measures how many of all available entities are among the extracted ones, $\frac{TP}{TP+FN}$. When the task concerns the extraction of more than one category of entities (multi-class), then these three measures are calculated per each class. When assessing the overall performance of a multi-class model, the average is calculated by dividing the overall performance by the number of entities in the gold/test corpus - this is the so-called macro-average. And separately, micro-P, micro-R and micro-F1 are calculated per entity class and averaged over the number of classes. The difference in the two measures is most indicative when the classes are imbalanced.

2.2 Relation extraction

2.2.1 Task definition

Relation extraction in the biomedical domain is a core task within biomedical text mining and large language model applications, focused on detecting and describing semantic relations between entities such as chemicals, diseases, drugs, and proteins. These entities represent biomedical concepts that are conveyed through language using specific terms, names, and other linguistic expressions.

Automated RE plays a pivotal role in transforming unstructured biomedical texts into structured, machine-readable data, thereby facilitating large-scale data integration, efficient knowledge retrieval, and downstream analytical processes. The RE process typically involves a series of computational steps designed to identify and formalize relationships within text.

The process initiates with named entity recognition, where the system identifies and annotates pertinent biomedical entities within the text. For example, in the sentence “The gut-brain axis reduces stress responses”, the phrases “gut-brain axis” and “stress responses” are detected as key biomedical entities. Following entity recognition, the system conducts relation detection, which entails determining whether a relationship exists between the identified entities within a *knowledge-rich context* (Meyer, 2001). This process uses syntactic structures and lexico-semantic cues from the text to accurately infer and characterize the nature of the relationship. In the example provided, the system evaluates whether *“gut-brain axis”* is semantically linked to *“stress responses”*.

Once the relationship is identified, the final step is **relation classification**, where the nature of the connection is categorized based on its semantic type. In this case, the verb *“reduces”* indicates a **causal negative** relation, signifying that the gut-brain axis exerts a downregulating influence on stress responses. This verb acts in a relation:

```
{
  "subject": "gut-brain axis",
  "predicate": "reduces",
  "object": "stress responses",
  "relation_type": "causal_negative"
}
```

Through this multi-step process, automated relation extraction enables the conversion of rich, unstructured biomedical texts into structured relational data. This structured output supports critical biomedical informatics applications, including ontology construction and knowledge graph development, large-scale literature mining, and clinical decision support systems.

2.2.2 Approaches

Relation Extraction has progressed considerably, moving from rule-based systems and manually crafted features to end-to-end models that learn directly from text. This shift has been made possible by recent improvements in deep learning and the progress of powerful pre-trained language models (PLMs). Today's RE methods vary widely, each designed to address different challenges such as handling large volumes of text:

(a) **Pattern-based approaches.** Early RE systems relied on rule-based extraction (e.g., Ramakrishnan et al., 2006; Sateli and Witte, 2015) and manually constructed lexico-syntactic patterns (e.g., Hearst, 1992) or automatically induced patterns using bootstrapping techniques (e.g., Brin, 1999). Manually constructed patterns are predefined templates or rules created by experts to help identify specific relationships between words or entities in text. These patterns are precise, look for certain words, phrases, or grammatical structures that typically signal a relationship, and require significant domain knowledge and manual effort. Instead of writing patterns by hand, bootstrapping methods start with a small set of seed relation instances and search a large set of texts to find common ways these relations appear. They then use those patterns to find more examples and improve themselves iteratively. This self-expanding process allows for broader coverage with less upfront manual intervention, though it can introduce noise and requires strategies to manage it.

(b) **Hybrid approaches integrate the strengths of rule-based extraction and machine learning techniques**. Rule-based components provide high precision by encoding expert knowledge through patterns and linguistic rules, while machine learning models contribute scalability and robustness by learning from data. This combination often improves overall performance, balancing interpretability with adaptability, especially in complex domains where purely statistical or rule-based systems fall short.

A hybrid system might use manually crafted rules to identify explicit phrases like "gut microbiota influences stress response" or "brain-gut axis modulates inflammation," ensuring high precision in detecting known relationships. At the same time, a machine learning model could analyse less structured sentences or infer implicit relations from varied contexts, such as "alterations in gut flora correlate with anxiety behaviours," which may not fit rigid patterns. Combining both allows the system to accurately capture both

well-defined and subtle gut-brain interactions across diverse biomedical literature. (e.g. Miwa and Bansal 2016, Marchesin and Silvello, 2022)

(c) Semantic parsing incorporating syntactic parsing and coreference resolution. Semantic parsing involves converting natural language into structured representations that machines can understand. By combining syntactic parsing, which identifies grammatical roles and dependencies (e.g., subject, object, modifiers), with coreference resolution, which links different expressions that refer to the same entity (e.g., "the microbiome" and "it"), semantic parsing can extract more nuanced and accurate relationships from text (Li et al. 2017).

Consider the sentence:

> "*The gut microbiome influences the immune system, which in turn affects stress-related brain responses*."

Here, syntactic parsing helps recognize that “the gut microbiome” is the subject influencing “the immune system,” and that “the immune system” affects “brain responses.” Coreference resolution links “which” back to “the immune system,” allowing the system to correctly infer a chain of relationships:

> (gut microbiome → influences → immune system)
>
> (immune system → affects → stress-related brain responses)

Coreference resolution is the task of identifying when grammatical, lexical, or syntactical expressions in a text refer to the same entity. In simpler terms, it's how a system figures out that pronouns or noun phrases like "it," "they," "this," "the bacteria," or "the system" all refer back to something mentioned earlier in the text. In the example, "The gut microbiota plays a crucial role in brain function. It influences stress responses and emotional behaviour." Here, "It" refers to "The gut microbiota." A coreference resolution system would produce the link "It" → "The gut microbiota".

(d) **Techniques leveraging PLMs, including Large Language Models (LLMs).** Recent developments in PLMs, such as, BERT, GPT and others have significantly improved how we extract relationships from medical and scientific texts. These models are trained on large text corpora and learn deep contextual and semantic representations of language. They can perform relation extraction even without being explicitly trained for a specific task or domain.

These models are capable of detecting and interpreting complex relationships between various scientific entities within a single sentence. For instance, given the sentence: “The gut microbiota influences emotional behaviour via the vagus nerve,” a well-trained model will extract the following relations:

- gut microbiota → influences → emotional behavior
- gut microbiota → via → vagus nerve

In other words, the model recognizes that changes in gut microbiota affect emotional behavior and that this effect occurs through the vagus nerve pathway.

A key advantage of recent PLMs lies in their ability to adapt to new tasks with minimal additional training. Unlike earlier approaches that required extensive retraining for each specific domain, modern PLMs already encode a wealth of biomedical and linguistic knowledge derived from large-scale corpora. As a result, they can be adapted to specialised relation extraction tasks using only a limited amount of annotated data, or, in some cases, through carefully designed prompts without further fine-tuning, e.g. (Ribeiro et al. 2023). This drastically reduces the time and effort typically associated with training a specialized model from scratch. With LLMs the RE task is performed through prompt-based techniques. This shows how PLMs can extract relations flexibly without needing specialized training, making them a powerful tool for exploratory biomedical information extraction.

Depending on the types of relations and entities, the top known systems for relation extraction perform with F1 score of 47.7% to 88.5% (Goyal and Singh 2025; Luo et al. 2022; Névéol et al. 2011; Sänger and Leser 2021). While many approaches tackle relation extraction tasks as one end-to-end task (Huguet Cabot and Navigli 2021), there are also studies dedicated solely to one of the tasks (Hassan et al., 2023; Sänger and Leser 2021, Luo et al. 2022). In Luo et al. 2022 BiLSTM-CRF, BioBERT-CRF, and PubMedBERT-CRF are applied for RE and achieve 47.7% (novelty), 72.9% (entity pair) and 58.9% (relation type) F1-score. The authors of (Hassan et al., 2023) propose an unsupervised BERT-based approach focused only on the RE task from PubMed abstracts and reach up 88.5% F1-score on the Drug-Drug Interaction (DDI dataset). On the ChemProt dataset the same approach achieves F1 85.8%. At the same time (Sänger and Leser 2021) propose a neural network RE approach employing both corpus-level entity embeddings and pair embeddings. It achieves an overall improvement of F1 score about 4-29% over traditional methods. (Kanjirangat and Rinaldi 2021) incorporate BioBERT and grammatical features such as Shortest Dependency Path. Their goal is to select the most representative samples for model learning by pruning out noisy information. These approaches outline the frame in which the partner proposed approaches described in Section 4 are developed.

2.2.3 Evaluation metrics

Relation extraction is usually evaluated with the same measures as NER - Precision, Recall and F1 and their micro- and macro- variations (see Section 2.1.4).

2.3 Knowledge graph construction (KBC)

We individuated ten main studies addressing RDF knowledge graph construction from text using a range of methods. Four studies focus on knowledge graph generation, four - on RDF triple extraction, one - on relationship discovery, and one - on knowledge base construction. Approaches fall into six key categories: 1) Hybrid designs combining rule-based methods with machine learning; 2) Pure rule-based extraction; 3) Semantic parsing; 4) Incorporating syntactic parsing and coreference resolution; 5) Bootstrapping methods that iteratively improve pattern extraction; and 6) Techniques leveraging pre-trained language models, including LLMs.

State of the art (SOTA) research on the topic is primarily focused on NER, syntactic and semantic parsing, and lexical pattern matching. Performance metrics varied from an F1 score of around 50% to 95% precision in a rule-based system, with several studies reporting precision, recall, and F1-score. Language support was rarely detailed—except for one study which explicitly handled English text, another - Spanish, and one - multilingual input. The type of processing ranged from real-time extraction from data streams to batch processing. Integration with existing knowledge bases through entity linking and ontology mapping also emerged as a notable feature.

2.3.1 Text Processing and Entity Extraction

The reviewed studies employ a range of techniques to address text processing and entity extraction:

- Named Entity Recognition: In (Gerber et al. 2013) NER is incorporated in the pattern extraction process. (Kertkeidkachorn and Ichise 2018) use the Stanford Natural Language Processing tool for NER.
- Coreference Resolution: (Exner and Nugues 2012) use a coreference solver to group entity actions and properties across different sentences.
- Syntactic Parsing: In (Ramakrishnan et al. 2006) the authors employ syntactic parsing as an initial step in their text processing pipeline, using a part-of-speech tagger and chunk parser to produce parse trees.
- Semantic Parsing: In (Exner and Nugues 2012) semantic parsing is used to understand the meaning of text and extract structured data.
- Language Model Usage: In (Melnyk et al. 2022) the authors leverage pre-trained language models like T5 for entity extraction, indicating a shift towards more advanced NLP techniques. More recent studies that use LLMs and other pretrained transformer LMs and DL models include (Saponara 2022), (Melnyk et al. 2021) (Kireev and Vasiliev 2024) (Xiao et al. 2022) (Gupte et al. 2021).

2.3.2 Pattern-Based Approaches

Pattern-based approaches are prevalent across several studies, particularly those employing rule-based or hybrid methods:

- Lexical Patterns: In (Rios-Alvarado et al. 2022) lexical patterns are used to extract named entities and semantic relations.
- Bootstrapping: In (Gerber and Ngomo 2012) BOA is introduced - a bootstrapping strategy that extracts natural-language patterns representing predicates found on the Data Web.
- Rule-Based Extraction: The approaches of (Ramakrishnan et al. 2006) and (Sateli and Witte 2015) rely heavily on rule-based extraction methods.

2.3.3 Knowledge Graph Construction Techniques

The reviewed studies employ a variety of techniques for constructing knowledge graphs from processed text. In Table 1 and below, we discuss five different information extraction techniques categorized as follows: 1) Hybrid; 2) Semantic parsing; 3) Bootstrapping; 4) Pre-trained Language Models; 5) and Rule-based Extraction.

Hereditary pays particular attention to probabilistic knowledge graphs. Methods for generating such graphs fall into several distinct categories. Bayesian network models and probabilistic ontologies (as in Freedman, et al. 2023, Freedman, et al. 2024) and

soft logic approaches (Pujara et al. 2013) represent uncertainty with explicit probability distributions, though some applications rely on synthetic data and omit scalability details.

Table 1. Techniques for constructing knowledge graphs from processed text.

Technique Category	Key Features	Limitations	Performance
Hybrid (Rule-based + Machine Learning)	Combines interpretability of rules with adaptability of ML	Complexity in balancing rule-based and ML components	Varies; e.g., T2KG achieves F1 score of ~50%
Semantic Parsing	Deep understanding of text meaning	Computationally intensive	No mention found in reviewed studies
Pre-trained Language Models	Leverages large-scale pre-training	Requires significant computational resources	Matches or outperforms state-of-the-art systems
Bootstrapping	Iterative improvement of extraction patterns	May propagate errors if not carefully controlled	Reported as “very high accuracy” without specific metrics
Rule-based Extraction	High interpretability and precision	Limited adaptability to new domains or languages	95% precision reported in one study

Biomedical texts, primarily electronic health records, support the construction of probabilistic knowledge graphs via several distinct methods. In (Rotmensch et al. 2017), the authors use logistic regression, Naive Bayes, and Bayesian networks with noisy OR gates to extract disease–symptom links, reporting precision of 0.85 at recall of 0.6. Belief-based approaches, as described by (Goodwin et al. 2013), assign qualitative assertion states to captured medical concepts. In (Jiang et al. 2017), the authors combine Boltzmann machines and Markov networks in weighted graph representations that yield average discounted cumulative gains (DGC) about 1, in many cases. Please, note that this is not normalized DCG and the reported values for average DCG in the study are in range between 0 and 2.5. There is also one additional detail - most of the clinical texts have only one actual disease and the model ranks it at the top of the results for diagnosis. This makes the results interpretation more challenging. Nerveless the proposed method provides good basis for further research on probabilistic knowledge graph building.

Implementation Considerations

Language Dependencies and Multilingual Support

The reviewed studies demonstrate varying levels of language dependency and multilingual support:

- **Language-Specific Approaches**: In (Rios-Alvarado et al. 2022) the target is Spanish text.

- **Multilingual Approaches**: The authors of (Gerber and Ngomo 2012) claim their approach is applicable to multiple languages, although specific details on language handling are not provided.
- **Language-Agnostic Methods**: Studies using advanced machine learning techniques, like the one in (Melnyk et al. 2022), may offer more language-agnostic approaches, especially when leveraging multilingual pre-trained models.

Processing Architecture and Scalability

Scalability is a crucial consideration in knowledge graph construction, particularly for large-scale or real-time applications:

- **Real-time Processing**: In (Gerber et al. 2013) the authors focus on real-time RDF extraction from unstructured data streams, addressing the challenge of processing high-volume, dynamic data sources.
- **Batch Processing**: Most studies employ batch processing approaches, which may be suitable for static datasets but could face challenges with large-scale or continuously updating data sources.
- **Distributed Processing**: While not explicitly mentioned in most studies, distributed processing architectures could be crucial for scaling knowledge graph construction to large datasets or real-time scenarios.

Integration with Existing Knowledge Bases

Several studies highlight the importance of integrating constructed knowledge graphs with existing knowledge bases:

- **Entity Linking**: The authors of (Kertkeidkachorn and Ichise 2018) and (Muñoz et al. 2014) emphasize the importance of linking extracted entities to existing knowledge bases like DBpedia.
- **Ontology Mapping**: In (Exner and Nugues 2012), is described an ontology-mapping system that aligns extracted triples to a DBpedia namespace.
- **Background Knowledge Utilization**: In (Gerber and Ngomo 2012), background knowledge from the Data Web is used to inform the extraction process.

3 Data and Organised Challenges

Most of the NER and RE annotated data are result of the open challenges such as the CLEF series[10] which aim to focus the researchers' efforts on a given problem and foster the development of tooling in a concrete research area. Some good overview papers which mention extensively the past challenges and resulting datasets are (Bravo et al. 2015; Yang et al. 2023; Huang et al. 2024).

[10] https://www.clef-initiative.eu/

The GutBrainIE challenge[11], created by the IIIA team within UNIPD[12] under the Task 5.6 umbrella, is a challenge targeting NLP-based extraction of entities and relations in the context of the BioASQ Lab within CLEF 2025.

The CLEF GutBrainIE challenge is organised in 4 subtasks corresponding to the types of provided annotations, namely:

- Task 6.1 - Named-entity recognition: Classify specific text spans, referred to as entities, according to 13 predefined labels.
- Task 6.2.1 - Binary RE: Identify which entity labels are related.
- Task 6.2.2 - Ternary tag-based RE: Identify which entity labels are related and classify each relation according to 17 predefined labels.
- Task 6.2.3 - Ternary mention-based RE: Identify which of the entities are involved in relations and classify each relation according to the 17 predefined labels.

Notice that tasks 6.2.1, 6.2.2, 6.2.3 can be seen as subtasks of a larger Relation Extraction task. More details are available on the challenge website[9].

The released dataset consists of 1607 labelled titles and abstracts from PubMed on the topic of the gut-brain axis. These annotations consist of entities (with their corresponding type) in the text, their relations, relation type, and relevant text location. A sample document is shown on Figure 2.

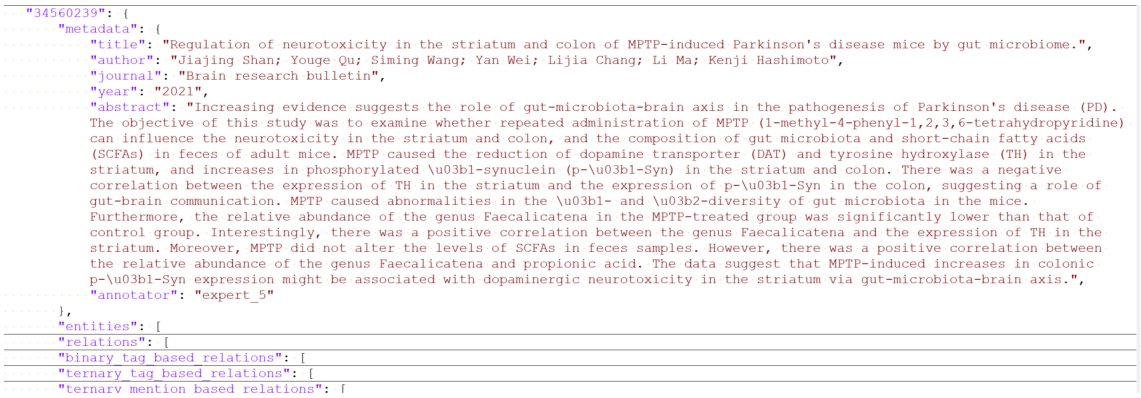

Figure 2. CLEF 2025 GutBrainIE challenge - a sample training data document.

The annotations are generated using four different skill levels, beginning with *Gold* and *Platinum* annotations (present in the test dataset), annotated by highly skilled experts in the field. The *Silver* collection is annotated by linguistic Students, while the *Bronze* collection is generated using automated systems. There is a predefined annotation scheme including the possible entity types (and their description) and allowed links between them. A detailed description of the relation types and data distribution is available on the challenge website[11] and will be available in project deliverable D5.10 due M24.

[11] https://hereditary.dei.unipd.it/challenges/gutbrainie/2025
[12] https://iiia.dei.unipd.it/team

The challenge dataset[13] is used by the challenge participants to create and evaluate their own solutions. It is also used by the organisers to evaluate the submitted approaches and to improve comparability and reproducibility of the claims on publicly available data.

4 Proposed solutions (per partner, including data, experiments, results)

The solutions, result of partner work on T4.6 presented below, are all dealing with relation extraction despite other topics. Therefore, they tackle the essence of the knowledge base creation. With the development of reliable approaches for relation extraction focused on the gut-brain axis, we get one step closer to the automatic creation of probabilistic KBs in this area. Once developed, these methods could be iteratively applied for updating, refining and reassessing available facts in the KB. Here follow the partner solutions.

4.1 PubMiner AI

4.1.1 Motivation

PubMiner AI is inspired by the attractive results demonstrated by the GraphRAG approach (GraphRAG 2024) approach and knowledge base creation initiatives relevant to HEREDITARY project (Marchesin and Silvello 2022, Giachelle et al. 2023, Marchesin et al. 2023). It is a result of combining available linked resources in the Ontotext's LinkedLifeData Inventory[14] with AI intelligence to identify relationships between biomedical entities, which are of interest to the HEREDITARY. With PubMiner AI, the extracted relationships are intuitively visualized and added to a knowledge graph which could be separately explored or involved in further research tasks. PubMiner AI focuses on *Amyotrophic Lateral Sclerosis (ALS)*, therefore it focuses on HEREDITARY use cases 1 and 2 that are about neurodegenerative diseases. These use cases are detailed in deliverable D2.16 (Chiò et al. 2024).

4.1.2 Methodology

PubMiner AI explores, extracts and integrates facts from vast scientific knowledge such as PubMed. Its source data is a slice of PubMed, including all publications from year 2022. Given a research question like: "***What are the potential susceptibility genes and biomarkers for ALS?***" PubMiner AI goes through the following RAG scenario.

1. Select the question context - PubMiner AI leverages Ontotext's LinkedLifeData[15], including datasets like MeSH, UMLS, NCBI Gene, and SemMedDB, to select PubMed articles related to ALS diagnosis and genetics. This filtering is based on the abstracts MeSH semantic annotation and additionally on SemMedDB annotations

[13] https://hereditary.dei.unipd.it/challenges/gutbrainie/2025/#three
[14] https://web.archive.org/web/20250322033318/https://www.ontotext.com/solutions/healthcare-and-life-sciences/linked-life-data-inventory/
[15] https://web.archive.org/web/20250322033318/https://www.ontotext.com/solutions/healthcare-and-life-sciences/linked-life-data-inventory/

(see a blog post[16] on this topic). The resulting articles define the context of the research question. For the given example these are 53 articles from 2022 alone.

2. Query LLM with augmented context – The question is sent to and LLM along with the selected context from Step 1. For comparison, with the application of the RAG technology this approach produces more than 10 times more results than a generic query which includes only the question without the context.

The overall workflow of PubMiner AI is presented on Figure 3. The Graph RAG technology is used to identify new ALS susceptibility genes and biomarkers along with provenance information for validation and further source exploration. Expert validation confirms its accuracy; more details are available in the Evaluation section 4.1.3. The extracted relationships (<gene/biomarker> <ALS>) from the scientific publications at the RAG steps are used for creating an evidence-based knowledge graph with focus on ALS susceptibility genes and biomarkers with potential for new discoveries. This is a scalable workflow because it allows for modifications of the research goal and used resources at scale. It can be adapted for other research questions and has potential for unlocking insights hidden in scientific publications.

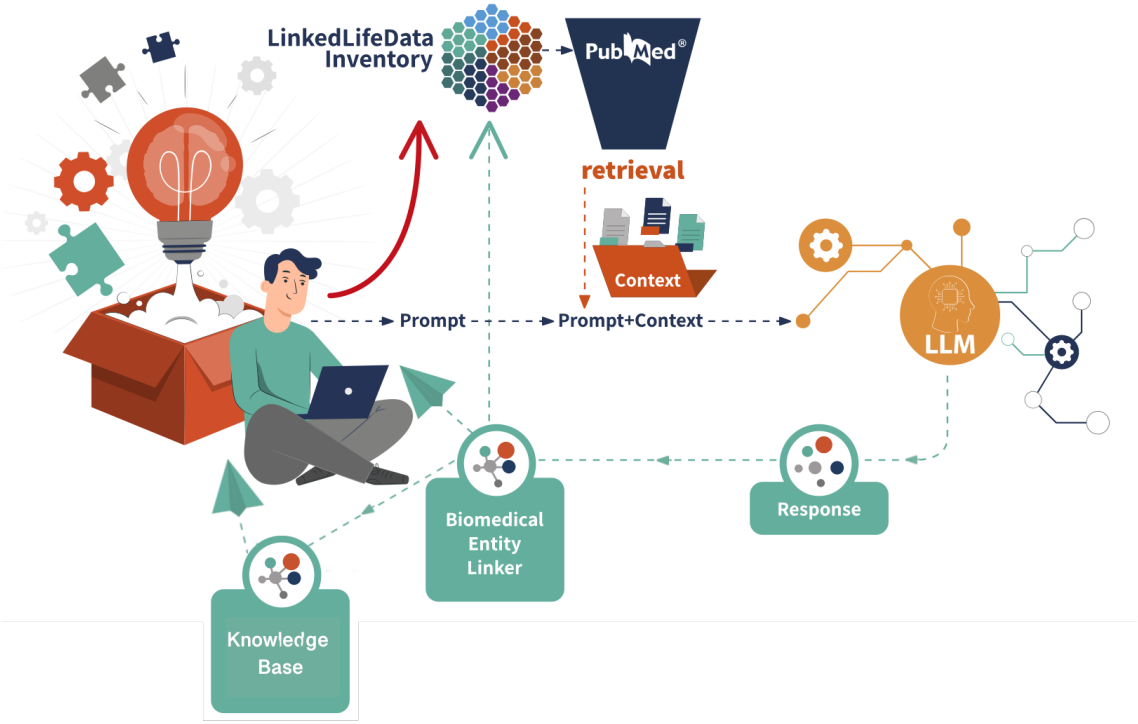

Figure 3. PubMiner AI workflow.

Figure 4 is a Venn diagram of the results of a sample run of PubMiner AI. The circles contain the results from 1) querying a large language model[17] (LLM) without context - this yields fewer genes and biomarkers (the smaller circle on the left) and 2) running the PubMiner AI Graph RAG method (the larger circle on the right containing on average

[16] https://graphwise.ai/blog/unlocking-genetic-insights-how-pubminer-ai-enhances-als-research-in-hereditary/
[17] https://www.ontotext.com/knowledgehub/fundamentals/what-is-a-large-language-model/

more than 10 times more results than the one on the left). This diagram highlights the effectiveness of the PubMiner AI approach in uncovering novel knowledge. Details on expert evaluation of the results are given in Section 4.1.3.

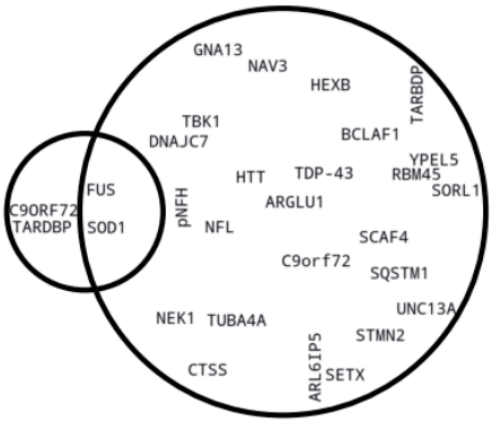

Figure 4. PubMiner AI's GraphRAG approach superiority vs generic query.

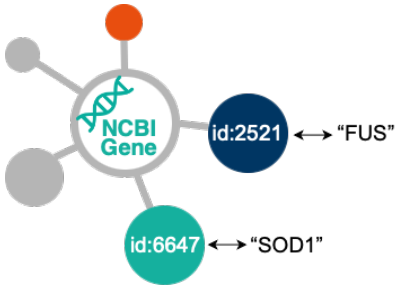

Figure 5. Biomedical Entity Linker - each gene name is getting linked to the corresponding NCBI gene concept.

Knowledge Graph Construction and Exploration

The extracted gene names like "SOD1" or "FUS" in Figure 4, are disambiguated using the Biomedical Entity Linker, a service produced by Ontotext, that takes as input a gene name and returns as output the respective NCBI Gene concept (Figure 5). This way the Biomedical Entity Linker transforms the textual representations of genes into interconnected concepts. These concepts are later used to build a probabilistic knowledge graph illustrating relationships between ALS and its susceptibility genes (Figure 6). The approach in line with the recent developments relevant to HEREDITARY (Marchesin and Silvello 2022, Giachelle et al. 2023, Marchesin et al. 2023). The confidence score associated with each relationship at extraction time, is used as a measure of reliability and relatedness of the extracted facts to the asked question. The confidence score is preserved in the graph and will be leveraged in future iterations to reassess and refine the reliability of facts already present in the knowledge base. Moreover, PubMiner AI conforms to the WP requirement of iterative KB building because with iterative runs of PubMiner AI on different documents (and even on the same ones, considering the LLM non-deterministic behaviour), new facts are added to the KB and previously extracted information is refined or updated. The generated graph, aligned with the HERO ontology (see D3.1), can be integrated into existing systems for further exploration. Additionally, PubMiner AI provides multiple ways for researchers to explore the extracted relationships, including access to LinkedLifeData Inventory[16] resources, PubMed, and real-time analytics, enabling deeper investigation beyond initial findings.

PubMiner AI aims to contribute to achieving one of the main HEREDITARY goals - to speed up disease detection and treatment for neurodegenerative conditions like ALS. PubMiner AI supports this by providing a structured workflow to identify and evaluate genetic links and by building standardized RDF knowledge graphs aligned with the HEREDITARY ontology for easy exploration and verification. This approach follows FAIR principles, making the data findable, accessible, interoperable, and reusable, helping

researchers quickly discover and validate potential therapeutic targets from complex biomedical literature.

A detailed description of PubMiner AI[1] and the HEREDITARY case study can also be found in an online blog post[18].

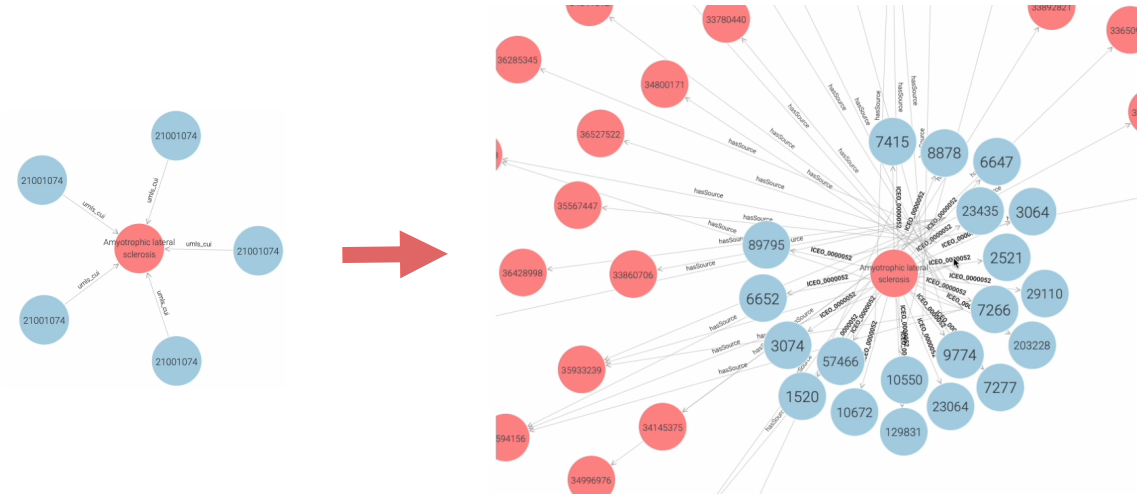

Figure 6. The "Amyotrophic Lateral Sclerosis" node in LinkedLifeData Inventory before (on the left) and the augmentation with susceptibility genes relations (on the right).

4.1.3 Evaluation Results

The system has been manually evaluated by an expert who checked the results of 4 consecutive system runs providing as input the question: "***What are the potential susceptibility genes and biomarkers for ALS?***". The output of the system was evaluated in two directions: (i) gene validation checking whether the results are indeed gene names, (ii) degree of relatedness of the output gene mentions, in the context of the original paper they are extracted from, to the given ALS scenario: strongly related, slightly related, irrelevant.

The results of the evaluation show that 100% of the gene names in the results were present in the NCBI Gene database. In each of the 4 runs, the generic query returned 4 results, and the RAG query returned between 47 and 50 results (more than 10 times more). Out of these, 1 of the generic query results was not directly related to the asked question and the remaining 3 were strongly related. For comparison, in the RAG results, 4 results were slightly related which is less than 10%, 41-44 were strongly related, which is 85-88% of the results and 2-3 of the results were still gene names but irrelevant to the scenario. The expert conclusion is that in these 85-89% of the RAG results there was causal relation of these genes being susceptibility gene for ALS.

[18] https://graphwise.ai/blog/advancing-automatic-knowledge-extraction-with-pubminer/

4.2 TUGRAZ@CLEF 2025 Gut-BrainIE - Constrained Linked Entity Annotation using RAG (CLEANR)

The CLEANR system (illustrated in Figure 7) builds on existing RAG and few-shot systems that are used for Relation Extraction. We extend the existing systems with a constrained LM output to allow only correct responses from the LM.

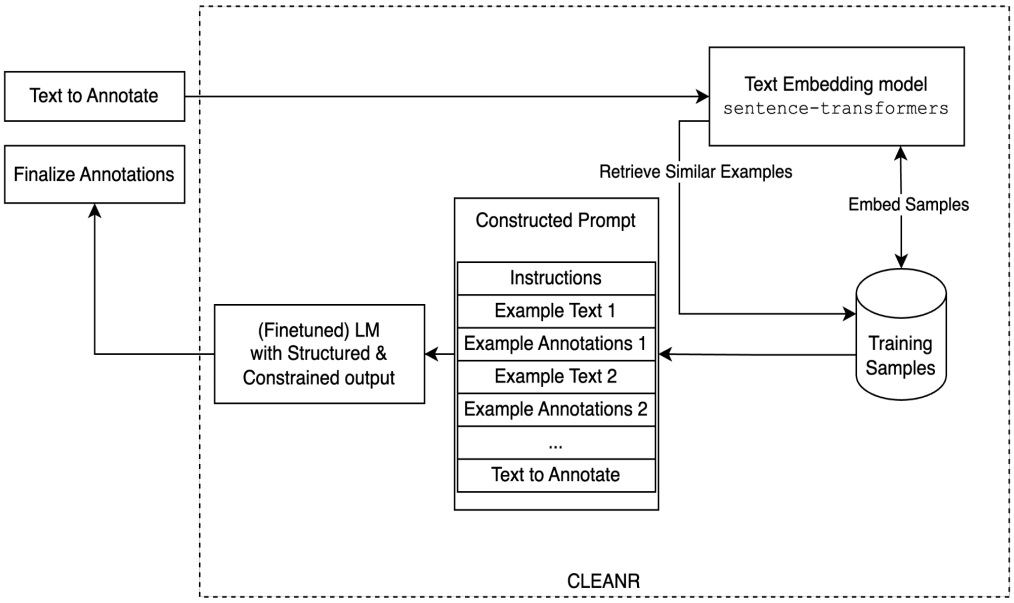

Figure 7. CLEANR workflow

4.2.1 Similar and Related Systems

CLEANR is inspired by existing Relationship Extraction systems, such as RAG4RE (Retrieval-Augmented Generation-based Relation Extraction), which 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 samples.

4.2.2 Methodology & Implementation

CLEANR extends the approach to use RAG 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). We use the sentence-transformer system (Reimers and Gurevych 2019) to embed the given training samples and store them in a Postgres database using the pgvector extension.19

Another key contribution of our methodology is the addition of constrained LM generation. We utilize llama-cpp20 and llama-cpp-agent21 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

[19] https://github.com/pgvector/pgvector
[20] https://github.com/ggml-org/llama.cpp/tree/master/grammars
[21] https://github.com/Maximilian-Winter/llama-cpp-agent

syntax, which is then used to constrain the LM output to the exact schema provided by the challenge.

4.2.3 Evaluation Results

We performed our evaluation on the dev-set provided within the CLEF-GutbrainIE challenge using the provided script. The results are below the baseline posted by the Challenge due to a late update to the evaluation set, removing duplicates, which led to inflated scores in our testing. Nevertheless, our system combines RAG and structured generation to retrieve data without fine-tuning or adaptation to the model. We performed additional fine-tuning on the LM, increasing the performance only for the last task. Our results for the CLEANR approach are shown in Tables 2 through 4.

Table 2. Task 6.2.1 Results for CLEANR on the Dev Set

Model	RAG	Reorder	Precision	Recall	F1	Micro Precision	Micro Recall	Micro F1
Hermes LLama 3.1 8B	TRUE	FALSE	0.25	0.12	0.15	0.55	0.24	0.33
OpenAI 4o Mini	TRUE	FALSE	0.15	0.17	0.15	0.44	0.3	0.36

Table 3. Task 6.2.2 Results for CLEANR on the Dev Set

Model	RAG	Reorder	Precision	Recall	F1	Micro Precision	Micro Recall	Micro F1
Hermes LLama 3.1 8B	TRUE	FALSE	0.26	0.11	0.15	0.52	0.22	0.31
OpenAI 4o Mini	TRUE	FALSE	0.16	0.17	0.15	0.42	0.29	0.35

Table 4. Task 6.2.3 Results for CLEANR on the Dev Set

Model	RAG	Reorder	Precision	Recall	F1	Micro Precision	Micro Recall	Micro F1
Hermes LLama 3.2 3B Finetune	FALSE	FALSE	0.03	0.02	0.02	0.19	0.076	0.11
Hermes LLama 3.2 3B Finetune	FALSE	TRUE	0.03	0.026	0.02	0.19	0.076	0.11

4.3 Ontotext@CLEF 2025 GutBrainIE - challenge solutions

4.3.1 Methodology

4.3.1.1 NER

In the context of the related approaches described in the beginning of the deliverable, Ontotext created several systems solving the named entity recognition problem for the GutBrainIE Task at CLEF 2025 BioASQ Lab. The systems use transformer-based

models among which are BioBERT (Lee et al. 2020), BiomedNLP-BiomedELECTRA (Tinn et al. 2021), BioBERT PubMed (Gu et al. 2022), and the well-recognised GLiNER (Zaratiana et al. 2024). The results of the study are submitted for publication at the Gut-Brain track of BioASQ Lab Workshop22. The tables in the next section summarize the published results. Here is the list of best-performing models on the dev set of the task.

Single-model

- *dmis-lab/biobert-base-cased-v1.1* (Lee et al. 2020) - this was the model with the best precision on the development set. We also tried the large version of the model but it did not show any improvement.
- microsoft/BiomedNLP-BiomedELECTRA-base-uncased-abstract (Tinn et al. 2021) - Here, we will refer to this model as pubmed-electra-base. This configuration showed clear improvements over the baseline on the dev set, however, on the test set, the performance decreased - the recall dropped by approximately 7% (82.45% - dev, vs. 74.86% test), therefore the F1-micro dropped too.
- monologg/biobert_v1.1_pubmed (Gu et al. 2022) - lighter version of biobert-base, pretrained on PubMed abstracts only. This was the most consistent model on the development and test sets, suggesting good capabilities of generalization.
- numind/NuNER_Zero (Zaratiana et al. 2024) - Generalist Model for Named Entity Recognition using Bidirectional Transformer (GLiNER). The model was a baseline for the NER task performing worse than the other models on the dev set however it achieved highest precision, recall and F1 on the test set.

Table 5 shows the hyperparameters for each of the models. In the first column is given the model's name and the data split it was trained on. For more details about the data splits see Section 3.2.2.

Table 5. Ontotext's NER models hyperparameters.

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

Comparison of the results of the NER systems on the dev and test datasets are shown in Table 8 in the next section.

[22] https://www.bioasq.org/workshop2025

Ensembles

By analyzing the performance of the separate systems on the split of entity categories, we selected the best-performing system per each category on the dev set. This way we formed an ensemble of systems (best-on-dev) that includes 5 systems and each one of them contributes to the annotation only of given categories where it performs optimal, e.g. `GLiNER` output is considered only for gene and chemical, all other annotations are disregarded. The whole model composition is shown on the table below.

Table 7. Ontotext's ensemble model - split of categories per base model.

Model name	Categories
monologg/biobert_v1.1_pubmed	bacteria
microsoft/BiomedNLP-BiomedELECTRA-base-uncased	statistical technique
numind/NuNER_Zero (GLiNER)	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

4.3.1.2 Relation Extraction

For the relation RE task, we applied two alternative approaches shortly outlined below.

Babelscape/rebel-large

REBEL (Huguet Cabot and Navigli 2021) 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. In our case, 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, e.g. for the task Ternary Tag-based relation extraction the expected output structure would look like this:

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

Or if we take a concrete example:

```
<triplet> DDF <subj> target <rel> human
```

For the Ternary Mention-based relation extraction task, the expected output structure would look like this:

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

And with sample data it would be as follows:

```
<triplet> Alzheimer disease %DDF% <subj> target <rel> patients %human%
```

The model learns to predict these sequences which represent the extracted relations.

As a postprocessing step, aiming to improve precision, we have created a dictionary of domain and range for each of the predicates. Where domain is every possible subject for a given predicate and range is every possible object of that predicate. This dictionary was used to restrict the output of the model to only possible relations in the context of the task. The most successful version of REBEL was obtained by fine-tuning on all annotated collections, excluding the development set.

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) (Zhou et al. 2020). The ATLOP method uses a standard encoder transformer (Vaswani et al. 2017) (Devlin et al. 2018) model as its base model. Unlike REBEL, which is an end-to-end entity linking approach, this method requires the extracted named entities to be provided to the model. So a NER annotator is needed and expected to make entity predictions and the error from this step accumulates in the RE error.

Different base encoder models were used in order to improve the model's performance like BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), XLM-R (Conneau et al. 2020), PubMedBERT (Gu et al. 2022), BiomedELECTRA (Tinn et al. 2021). Some experiments with domain adaptation by continuing the models' pretraining using the masked language modeling objective (Devlin et al. 2018) were also made. Finally, we first fine-tuned 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 as follows: *batch size* - 4, *# epochs* - 200, *lr for encoder* - 5e-5, *lr for classifier* - 1e-4, *warmup ratio* - 6%, *max grad norm* - 1, *optimizer* - AdamW.

4.3.1 Evaluation Results

4.3.1.1 NER

Table 8 shows a comparison of all systems results on the dev and test datasets. For more details on the data splits refer to Section 3.2.2. Our best architecture on the dev set did not achieve that high results on the test set. Our expectation was that fine-tuning on the dev set would increase the performance, however surprisingly the models which achieved highest scores on the test set are not trained on the dev set at all.

Table 8. Ontotext's NER results on dev and test set.

Model	Data Split	Dev P %	Test P %	Dev R %	Test R %	Dev F1 %	Test F1 %
ensemble	best-models-on-dev	**84.26**	78.99	**82.90**	76.31	**83.57**	77.63
GLiNER	gold+platinum+silver	75.90	**80.65**	82.63	**79.54**	79.13	**80.10**
biobert-base-v1.1	bronze	81.75	78.59	81.83	79.22	81.79	78.91
pubmed-electra-base	all data + augmented-biobert	82.60	78.86	82.45	74.86	81.69	77.23
biobert-pubmed	all data + augmented-biobert	80.99	78.59	82.01	79.22	81.49	78.91

Only the ensemble model which included models re-trained on dev showed higher results on test in comparison with the other ensemble models but performed worse than the single models which were not retrained on dev. This could be either due to the ability of the models to progressively train or in a comparatively high distance in the concept's annotation in the dev and test set.

4.3.1.1 Relation Extraction

The results of all RE subtasks *Binary Relation Extraction*, *Ternary Tag-based Relation Extraction*, *Ternary Mention-based Relation Extraction* are summarized on Table 9, which shows dev set evaluations and Table 10 - test dataset evaluations. Here, all ATLOP models use NER predictions from the baseline GLiNER model. On Table 9, "+ DA" suffix means that the model went through an additional pre-training phase with masked language modeling before being fine-tuned on the specific task. And "+ TPT" suffix means the model went through fine-tuning on the NER task before being fine-tuned on the specific task.

The models perform comparatively consistently throughout the three tasks. Our general observation is that REBEL is a good end-to-end alternative for RE, however its performance is slightly worse than ATLOP models which are very effective. In the Binary RE task, we see that task pretraining gives improvement for the BiomedELECTRA, but it is not very significant.

Table 9. Ontotexts's Binary RE, Ternary Tag-based, Ternary Mention-based RE-model performance on dev set.

DEV Set	**Binary RE**			**Ternary Tag-based RE**			**Ternary Mention-based RE**		
Model	**Micro-P %**	**Micro-R %**	**Micro-F %1**	**Micro-P %**	**Micro-R %**	**Micro-F1 %**	**Micro-P %**	**Micro-R %**	**Micro-F1 %**
rebel-large	62.79	55.56	61.50	63.08	58.70	60.81	38.81	36.25	37.49
ATLOP + BERT base	67.50	61.36	61.03	69.36	52.17	59.55	47.55	32.86	38.86
ATLOP + RoBERTa base	67.54	58.64	62.77	67.36	56.52	61.47	46.44	33.75	39.09
ATLOP + XLM-R base	67.50	61.36	64.29	67.00	59.13	62.82	46.04	37.32	41.22
ATLOP + XLM-R base + DA	64.25	64.55	64.44	63.39	61.74	62.56	42.65	37.32	39.81
ATLOP + XLM-R base + TPT	70.11	58.64	63.86	70.05	56.96	62.83	45.52	34.46	39.23
ATLOP + BiomedELECTRA	62.26	60.00	64.71	64.90	58.70	61.64	43.33	37.14	40.00

ATLOP + BiomedELECTRA + TPT	72.58	61.36	66.50	72.34	59.13	65.07	48.56	36.25	41.51

On the **test set** while XLM-R and REBEL do not fall too much behind BiomedELECTRA, we see that BiomedELECTRA is the best-performing model, as a base model with task pretraining. In the Ternary Tag-based RE task, the main difference is that on the test set this time "ATLOP + BiomedELECTRA + TP" trained with 200 epochs performed slightly better in comparison to the one trained with 100 epochs. In the Ternary Mention-based task XLM-R slightly outperforms Biomed-ELECTRA, and rebel-large is not much behind. There is also significant degradation in performance compared to the previous tasks. The reasons here could be two-fold: (i) error accumulation throughout the subtasks and (ii) the search space becomes much larger in the final more complex subtask.

Table 10. Ontotext's Binary, Ternary Tag-based RE and Ternary Mention-based RE-model performance on the test set.

TEST Set	**Binary RE**			**Ternary Tag-based RE**			**Ternary Mention-based RE**		
Model	**Micro-P %**	**Micro-R %**	**Micro-F1 %**	**Micro-P %**	**Micro-R %**	**Micro-F1 %**	**Micro-P %**	**Micro-R %**	**Micro-F1 %**
rebel-large	68.59	56.71	62.10	68.88	55.56	61.50	43.36	31.10	36.22
ATLOP + XLM-R base (100 epochs)	70.62	54.11	61.28	68.37	55.14	61.05	46.86	30.96	37.28
ATLOP + BiomedELECTRA + TPT (100 epochs)	74.17	65.38	63.86	71.98	53.90	61.65	44.15	26.80	33.36
ATLOP + BiomedELECTRA + TPT (200 epochs)	72.37	47.61	57.44	73.26	56.38	63.72	46.12	29.49	35.98

4.4 ONTUG@CLEF 2025 GutBrainIE - challenge solution (TUGRAZ+ONTO)

4.4.1 Methodology

We combine the results of the two proposed methods by taking the union and intersection of the ONTO and TUG solutions to combine the strengths of our systems, resulting in our collaborative entry *ONTUG*.

4.4.2 Evaluation Results

Tables 11, 12 and 13 summarize the joint systems results on the three separate tasks - Binary RE, Ternary Tag-based RE and Ternary Mention-based RE respectively. The intersection enhances precision further, suggesting that the combination of both systems

yields a highly precise output, albeit with a higher number of results. The union, on the other hand, improves the F_1-score significantly.

Table 11. TUG+ONTO results from Binary RE on the Dev Set

Model	Set	RAG	Precision	Recall	F1	Micro Precision	Micro Recall	Micro F_1
Hermes LLama 3.2 3B & ATLOP + BiomedELECTRA + TPT (100 epochs)	∩	TRUE	0.25	0.07	0.10	0.85	0.15	0.25
Hermes LLama 3.2 3B Finetune & ATLOP + BiomedELECTRA + TPT (100 epochs)	∪	FALSE	0.59	0.54	0.54	0.71	0.65	0.67

Table 12. TUG+ONTO results from Ternary Tag-based RE on the Dev Set

Model	Set	RAG	Precision	Recall	F1	Micro Precision	Micro Recall	Micro F1
Hermes LLama 3.2 3B & ATLOP + BiomedELECTRA + TPT (100 epochs)	∩	TRUE	0.24	0.07	0.10	0.85	0.14	0.25
Hermes LLama 3.2 3B Finetune & ATLOP + BiomedELECTRA + TPT (100 epochs)	∪	FALSE	0.59	0.53	0.54	0.70	0.62	0.66

Table 13. TUG+ONTO results from Ternary Mention-based RE on the Dev Set

Model	Set	RAG	Precision	Recall	F1	Micro Precision	Micro Recall	Micro F1
Hermes LLama 3.2 3B & ATLOP + BiomedELECTRA + TPT (100 epochs)	∩	TRUE	0.11	0.05	0.06	0.27	0.13	0.18
Hermes LLama 3.2 3B Finetune & ATLOP + BiomedELECTRA + TPT (100epochs)	∪	TRUE	0.34	0.30	0.30	0.37	0.38	0.38

4.5 AAU@CLEF 2025 GutBrainIE challenge solutions

We propose a methodology for named entity recognition and relation extraction from biomedical documents related to the gut-brain axis. Our model is based on the ensemble of transformer-based models and careful model selection. The model is structured in a two-step phase, firstly for NER and secondly for RE. Figure 8 depicts the general architecture of the proposed model, consisting of a NER ensemble and a RE ensemble. This solution was built as part of a master thesis supervised by Daniele Dell'Aglio and Juan Manuel Rodriguez. This approach was developed in the context of the GutBrainIE challenge. This challenge defined 4 tasks as described in Section 3.2.1.

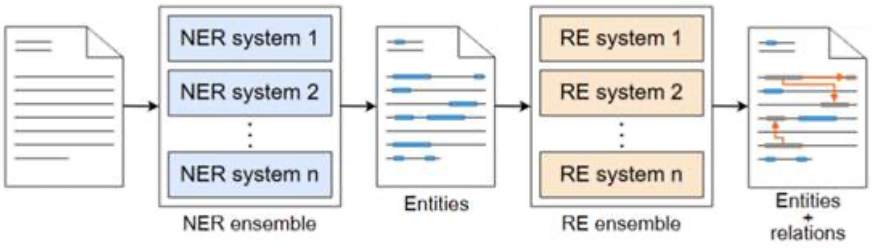

Figure 8. General view of the AAU NER+RE pipeline.

4.5.1 Similar and Related Systems

Currently, there exist many transformed based pretrained models that can be finetune for the NER and RE tasks. Moreover, our system actually builds on several of those models (Yasunaga et al. 2022; Gu et al. 2022; Tinn et al. 2021; Lee et al. 2020) (Alrowili and Shanker 2021), namely:

- BioLinkBERT-uncased-base,
- BioLinkBERT-uncased-large,
- BiomedBERT-base-uncased-abstract-fulltext,
- BiomedBERT-base-uncased-abstract,
- BiomedBERT-large-uncased-abstract,
- BiomedELECTRA-base-uncased-abstract, and
- BiomedELECTRA-large-uncased-abstract.

In addition to these models, other authors have leveraged model ensembles to handle noisy or inconsistent data, as averaging predictions helps suppress errors from individual models (Meng et al. 2021; Nguyen et al. 2019). Additionally, ensembles enhance model stability and generalization, making them more reliable for real-world applications.

4.5.2 Methodology

Both the NER and the RE components are model ensembles. In the text below, we describe how those models work. Notice that the models for each task follow a similar template, where different components can be replaced. For instance, the language models used by a particular NER or RE model can be any of the models mentioned in the previous section.

The system's NER process focuses on accurately identifying and classifying entity mentions within text. The primary method involves obtaining the tokens' embeddings to classify and using a classification head to obtain the final result. We used either a dense

linear layer, a Conditional Random Field (CRF) layer, or a bi-LSTM and a CRF for the classification head (Lample et al. 2016) (Dai et al. 2019). Both the bi-LSTM and CRF layer can consider each word's context, looking at the words before and after, to make a more informed decision about whether it is part of an entity and what type of entity it is (food, genome, chemical, etc.). Although the dense layer does not consider the neighbours, the embeddings computed by the pretrained language model are affected by other tokens. Figure 9 presents an overview of the NER system components. Notice that this template allows combining any of the pretrained models with any of the proposed token classification heads.

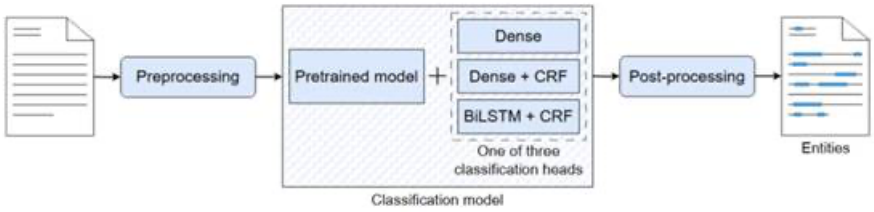

Figure 9. AAU NER system components.

For the RE, we introduce a set of special tokens, namely [E1], [\E1], [E2], and [\E2]. For detecting the relationship between two entities, each entity is demarked by a pair of these special tokens. Then, the new text is passed through a pretrained language model, and the embeddings for [E1] and [E2] are extracted, concatenated, and classified using a dense layer (Baldini Soares et al. 2019). The components to perform the NE task are depicted by Figure 10.

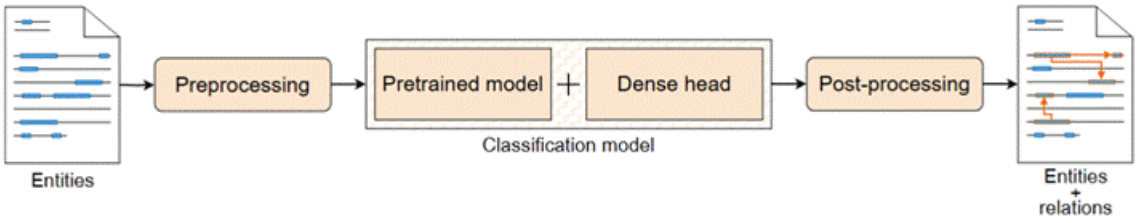

Figure 10. AAU RE system components.

The training strategy focuses on utilizing all available data, even noisier datasets, while giving higher-quality annotations more influence during optimization. A weighted average training loss is computed for each batch of samples, allowing the model to account for dataset quality. All models, except for NER models with a CRF layer, use the cross-entropy loss function. For the CRF, the loss function is negative log-likelihood. In the case of the RE, since there are no explicit negative instances, we use negative sampling to simulate the negative instances.

4.5.3 Evaluation Results

For the evaluation, we focus on F1-micro score and present the Precision-micro and Recall-micro. This is for both NER and RE tasks as both can be described as classification tasks. Notice that from now on, we will refer to F1-micro, Recall-micro, and Precision-micro, as F1, Recall, and Precision for brevity's sake.

Firstly, we perform an evaluation first step to determine the best combination of dataset weights. Table 14 presents the scores for the NER task on the development dataset for 18 dataset weight combinations using BioLinkBERT-base with a dense classification head. We use these distributions of weights to train other models. From this table, we can observe that a weighting schema where platinum and gold have the highest weight, silver is slightly lower than gold, and bronze is slightly lower than silver reports the best results.

Table 14. AAU NER evaluation for different weights using BioLinkBERT-base and Dense layer

W. Platinum	W. Gold	W. Silver	W. Bronze	Precision	Recall	F1
1.5	1.5	1.5	0.75	0.7866	0.838	0.8114
1.5	1.5	1.5	0.5	0.7918	0.8478	0.8188
1.5	1.5	1.5	0.25	0.792	0.8317	0.8114
1.5	**1.5**	**1**	**0.75**	**0.8157**	**0.8478**	**0.8314**
1.5	1.5	1	0.5	0.7888	0.8523	0.8193
1.5	1.5	1	0.25	0.7949	0.8397	0.8167
1.25	1.25	1.25	0.75	0.7911	0.8478	0.8185
1.25	1.25	1.25	0.5	0.8001	0.8424	0.8211
1.25	1.25	1.25	0.25	0.7923	0.8505	0.8204
1.25	**1.25**	**1**	**0.75**	**0.8003**	**0.8469**	**0.823**
1.25	1.25	1	0.5	0.8015	0.8317	0.8163
1.25	1.25	1	0.25	0.7892	0.8344	0.8111
1	1	1	0.75	0.7913	0.8451	0.8173
1	1	1	0.5	0.7852	0.838	0.8107
1	1	1	0.25	0.7938	0.8514	0.8216
1	1	1	1	0.7935	0.846	0.8189
1	1	1	-	0.8064	0.8317	0.8189
1	1	-	-	0.7264	0.8228	0.8093

After selecting the two best weighting schemas, Table 15 presents the F1 results for the different combinations of based pretrained models and classification heads. In this table,

it can be observed that in general the Dense layer performs slightly worse than Dense+CRF ($p < 0.01$) and BiLSTM+CRF ($p < 0.05$), yet the statistical significance of these differences is low. Moreover, there is no statistically significant difference between the CRF strategy. Notice that this analysis is not valid for a particular model, but for the performance in general, i.e., there might be a statistical significance on the results for BioLinkBERT-base Dense+CRF vs BioLinkBERT-base LSTM+CRF.

Table 15. AAU NER results for different pretrained base models and classification modules.

Pretrained Model	Dense	Dense + CRF	LSTM + CRF
Weights = [1.5, 1.5, 1, 0.75]			
BioLinkBERT-base	0.8151	**0.8213**	0.8152
BioLinkBERT-large	0.8094	0.8185	0.8149
BiomedBERT-base-uncased-abstract	0.8103	0.8193	0.8092
BiomedBERT-base-uncased-abstract-fulltext	0.8118	0.8176	**0.8268**
BiomedBERT-large-uncased-abstract	0.8148	0.817	0.8178
BiomedElectra-base-uncased-abstract	0.805	0.8174	0.8123
BiomedElectra-large-uncased-abstract	0.8072	**0.8201**	0.8198
Weights = [1.25, 1.25, 1, 0.75]			
BioLinkBERT-base	**0.8217**	0.8166	0.8158
BioLinkBERT-large	0.8097	0.8085	**0.8202**
BiomedBERT-base-uncased-abstract	0.8019	0.8109	0.8077
BiomedBERT-base-uncased-abstract-fulltext	0.8067	0.8174	0.8152
BiomedBERT-large-uncased-abstract	0.7831	0.8183	0.8176
BiomedElectra-base-uncased-abstract	0.8013	0.8156	0.81
BiomedElectra-large-uncased-abstract	**0.8211**	0.8165	0.8162
Average	**0.8085**	**0.8168**	**0.8156**

The final step for developing our NER ensemble is to select the number of the NER systems to ensemble. The strategy was to select the top-K performant NER systems. Table 16 presents the performance for the different ensemble sizes. As it can be observed, NER ensemble strategy achieves better performance than any single NER systems. We found the best performing configuration to be the top-9 systems, but in all cases the F1 was greater than 0.84, where no single NER system obtained a F1 greater than 0.83. When evaluated in the GutBrainIE Challenge test dataset, the F1 performance of this NER ensemble is 0.8382, which is lower than the obtained performance in the dev dataset.

Table 16. AAU NER ensemble results.

Ensemble size	Precision	Recall	F1
3	0.8351	0.8612	0.848
5	0.8376	0.8585	0.8479
7	0.836	0.8532	0.8445
9	**0.8436**	**0.8594**	**0.8514**
11	0.8398	0.8585	0.849
13	0.8406	0.8594	0.8499
15	0.8405	0.8585	0.8494
17	0.8393	0.8559	0.8475

Regarding RE tasks, our RE system performs the task 6.2.3, but the output can be converted to the other tasks as 6.2.1 and 6.2.2 requires less information. Regarding the dataset, we noticed that most of the documents have less than 100 relations, but there are a few outliers with more annotations in the silver and bronze dataset. Therefore, the first step is to determine the impact of these outliers in training along with the negative sample multiplier (NSM), i.e., the number of negative instances per positive instance, and the weight for the different datasets. To do this, we consider task 6.2.1 and BioLinkBERT-base as a pretrained model. Table 17 presents the results for this experiment. Notice that in most cases removing outliers and increasing the NSM improves the results. Notice that when silver and bronze dataset are not used for training, we consider the dataset to be without outliers, as no outlier is present in gold and platinum datasets. All in all, the best results were obtained when NSM is 10 and the weighting schema is 1.25, 1.25, 1, and 0.75 for the platinum, gold, silver and bronze datasets, respectively.

Using the best configuration according to the previous experiment, we test the performance of the different pretrained RE systems for the task. Table 18 depicts the results for the different RE systems, based on different pretrained base models. In this case, the goal is to rank the pretrained RE systems for creating the ensembles.

Finally, Table 19 presents the results for the three RE subtasks. Unlike the NER task, adding more RE systems to the ensemble seems not to improve the performance significantly. Hence, we decided to build our RE ensemble using only the top 3 performing RE system configurations.

Table 17. AAU F1 scores RE system BioLinkBERT-base under different dataset weights and NSM.

W. Platinum	W. Gold	W. Silver	W. Bronze	NSM	With Outliers	Without Outliers
1	1	-	-	1	-	0.7505
1	1	1	-	1	0.7208	0.7186
1.5	1.5	1	0.75	1	0.7125	0.7391
1.25	1.25	1	0.75	1	0.7171	0.7305
1	1	-	-	3	-	0.7562
1	1	1	-	3	0.7345	0.7606
1.5	1.5	1	0.75	3	0.7612	0.7458
1.25	1.25	1	0.75	3	0.7603	0.7736
1	1	-	-	5	-	0.7645
1	1	1	-	5	0.7543	0.7621
1.5	1.5	1	0.75	5	0.7612	0.7732
1.25	1.25	1	0.75	5	0.768	0.7757
1	1	-	-	10	-	0.7805
1	1	1	-	10	0.7642	0.7826
1.5	1.5	1	0.75	10	0.7708	0.7934
1.25	**1.25**	**1**	**0.75**	**10**	**0.7711**	**0.7965**

For submitting the challenge, we retrain the models using training and dev results. Therefore, we have two versions of the models, trained only with training and trained with all the data. For the NER task, adding the dev data increased the performance of the model. However, for the RE task, the best performance was achieved when trained only with the training data. However, notice that the RE performance depends on the NER performance as for the testing we do not know the named entities. Therefore, the RE submission was based on the NER ensemble trained with all the data.

When evaluated in the GutBrainIE Challenge test dataset, the F1 performance NER ensemble is 0.8382, which is lower than the obtained performance in the dev dataset but it is expected as there might be a slight overfit in the dev dataset. For the RE tasks, the F1 score is 0.6864, 0.6866, and 0.4635 for tasks 6.2.1, 6.2.2, and 6.2.3, respectively. Although these results are significantly lower than the ones reported in Table 19, it is important to note that the original results do not depend on the NER model for inferring the entities, while for the challenge the final results consider the whole pipeline.

Table 18. AAU F1 scores on the development dataset of RE classification for different base models.

Pretrained Model	Precision	Recall	F1
BioLinkBERT-base	**0.7379**	**0.8318**	**0.7821**
BioLinkBERT-large	0.7358	0.8227	0.7768
BiomedBERT-base-uncased-abstract	0.7438	0.8182	0.7792
BiomedBERT-base-uncased-abstract-fulltext	0.7094	0.8545	0.7753
BiomedBERT-large-uncased-abstract	**0.7354**	**0.8591**	**0.7925**
BiomedElectra-base-uncased-abstract	0.7059	0.8727	0.7804
BiomedElectra-large-uncased-abstract	**0.7629**	**0.8045**	**0.7832**

Table 19. AAU results for the ensembles for the three RE subtasks.

Ensemble Size	Precision	Recall	F1
Binary RE (Task 6.2.1)			
3	0.7530	0.8591	0.8025
5	0.7540	0.8500	0.7991
Ternary Tag-based RE (Task 6.2.2)			
3	0.7090	0.8261	0.7631
5	0.7209	0.8087	0.7623
Ternary Mention-based RE (Task 6.2.3)			

3	0.5489	0.6518	0.5959
5	0.5786	0.6375	0.6066

5 Conclusion and next steps

In this deliverable, two types of innovative solutions are presented, all developed in Task 4.6 of HEREDITARY project: PubMiner AI and solutions addressing the GutBrainIE Task at CLEF 2025 BioASQ Lab. The presented systems include components for: 1) information extraction (NER, RE) from biomedical literature, 2) EL for biomedical concepts and 3) KBC, present great potential for extracting valuable information about gut-brain interplay and contribute to the overall development of tools on the topic. The information extracted with such means can be integrated with existing resources and systems and used for iterative creation of knowledge bases.

In the next period, we foresee and extension of PubMiner AI. Its source dataset will include PubMed articles from years other than 2022 and also PubMed slices of longer periods. The current version addresses the ALS use case, and we plan to extend its functionalities to cover also other HEREDITARY use cases about Parkinson's disease and the connections between brain-related diseases and the gut microbiome. The generated KG, result of new use case scenarios, will be aligned with the new extended versions of HERO ontology.

The solutions for the GutBrainIE Task at CLEF 2025 BioASQ Lab will be studied in-depth with further error analysis that will allow for better understanding of their strengths and weaknesses and can lead to further improvement and higher performance respectively. The provided annotated datasets will streamline this process for both tasks, NER and RE being the core of knowledge base creation. In addition, the envisioned challenges organised by the project partners on similar topics in 2026 and 2027 will provide new testbeds and ideas for KBC.

A possible further step is the exploration of various AI agent-based and LLM-based approaches to benefit from few-shot learning for other entities and relations for gut-brain interplay, beyond those investigated in the GutBrainIE Task at CLEF 2025 BioASQ Lab.

REFERENCES

Key	Reference
Alrowili et al. 2021	Alrowili, S., & Shanker, V. (2021). BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. Proceedings of the 20th Workshop on Biomedical Language Processing. Proceedings of the 20th Workshop on Biomedical Language Processing, Online. https://doi.org/10.18653/v1/2021.bionlp-1.24
Apache UIMA, 2025	Apache UIMA - Apache UIMA. (n.d.). Retrieved June 4, 2025, from http://uima.apache.org/
Aronson et al. 2001	Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proceedings of the AMIA Annual Symposium, 17–21.
Baldini Soares et al. 2019	Baldini Soares, L., FitzGerald, N., Ling, J., & Kwiatkowski, T. (2019). Matching the blanks: Distributional similarity for relation learning. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. https://doi.org/10.18653/v1/p19-1279
Baum et al. 1970	Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains.The annals of mathematical statistics, 41(1):164–171, 1970.
Bengio et al. 2003	Bengio, Y., Ducharme, R. , Vincent, P. and Janvin, C. (2003). A neural probabilistic language model. J. Mach. Learn. Res. 3, null (3/1/2003), 1137–1155.
Berger et al. 1996	Berger, A. , Della Pietra, S. A. and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1):39–71, 1996.
Bertolo et al. 2024	Bertolo, R., Ditonno, F., Veccia, A., De Marco, V., Migliorini, F., Porcaro, A. B., Rizzetto, R., Cerruto, M. A., Autorino, R., Antonelli, A., & PubMed-indexed collaborators. (2024). Postoperative outcomes of transperitoneal versus retroperitoneal robotic partial nephrectomy: a propensity-score matched comparison focused on patient mobilization, return to bowel function, and pain. Journal of Robotic Surgery, 18(1), 96.
Bodenreider et al. 2004	Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Research, 32(Database issue), D267-70.
Bravo et al. 2015	Bravo, À., Piñero, J., Queralt-Rosinach, N., Rautschka, M., & Furlong, L. I. (2015). Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research. BMC Bioinformatics, 16(1), 55.
Brin et al. 1999	Brin, S. (1999). Extracting patterns and relations from the World Wide Web. In Proceedings of the WebDB Workshop at EDBT (Vol. 27, pp. 172–183).
Bunescu et al. 2005	Bunescu, R., Ge, R., Kate, R. J., Marcotte, E. M., Mooney, R. J., Ramani, A. K., & Wong, Y. W. (2005). Comparative experiments on

Key	Reference
	learning information extractors for proteins and their interactions. Artificial Intelligence in Medicine, 33(2), 139–155.
Cazzaro et al. 2024	Cazzaro, M., Menotti, L., Primov, T., Rueda, M., & Silvello, G. (2024). HEREDITARY: Deliverable 3.1: Ontology and federated execution methods. Zenodo. https://doi.org/10.5281/zenodo.14627940
Chiò et al. 2024	Chiò, A., Grassano, M., & Atzori, M. (2024). Deliverable 2.16: Design of neurodegenerative use cases. Zenodo. https://doi.org/10.5281/zenodo.14034445
Collobert et al. 2011	Collobert, R. (2011). Deep learning for efficient discriminative parsing.In AISTATS, pages 224–232. JMLR Workshop and Conference Proceedings, 2011.
Conneau et al. 2020	Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual representation learning at scale. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online. https://doi.org/10.18653/v1/2020.acl-main.747
Constantopoulos et al. 2020	Constantopoulos, P., & Pertsas, V. (2020). From publications to knowledge graphs. In Communications in Computer and Information Science (pp. 18–33). Springer International Publishing.
Cortes et al. 1995	Cortes, C. and Vapnik, V. (1995) Support-vector networks.Machine learning, 20:273–297, 1995.
Costa et al. 2024	Costa, R., Vezzani, F., Di Nunzio, G. M., Ramos, M., Silva, R., Carvalho, S., Canelas, M., Bonato, V., Romanovych, A., & Boytcheva, S. (2024). Deliverable 3.4: Medical terminology. Zenodo. https://doi.org/10.5281/zenodo.14628022
Cunningham et al. 2013	Cunningham, H., Tablan, V., Roberts, A., & Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Computational Biology, 9(2), e1002854.
Dai et al. 2019	Dai, Z., Wang, X., Ni, P., Li, Y., Li, G., & Bai, X. (2019, October). Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Suzhou, China. https://doi.org/10.1109/cisp-bmei48845.2019.8965823
Devlin et al. 2018	Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional Transformers for language understanding. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1810.04805
Ding et al. 2001	Ding, J., Berleant, D., Nettleton, D., & Wurtele, E. (2001). Mining MEDLINE: abstracts, sentences, or phrases? In: Biocomputing (pp. 326–337). World Scientific.
Exner et al. 2012	Exner, Peter, and Pierre Nugues. "Entity extraction: From unstructured text to DBpedia RDF triples." In The Web of Linked Entities Workshop (WoLE 2012), pp. 58-69. CEUR-WS, 2012. http://ceur-ws.org/Vol-906/paper7.pdf

Key	Reference
Fabregat et al. 2023	Fabregat, H., Duque, A., Martinez-Romo, J., & Araujo, L. (2023). Negation-based transfer learning for improving biomedical Named Entity Recognition and Relation Extraction. Journal of Biomedical Informatics, 138(104279), 104279.
Fundamentals et al. 2024	Fundamentals, O. (2024, February 23). What Is Graph RAG? Ontotext. https://www.ontotext.com/knowledgehub/fundamentals/what-is-graph-rag/
Fundel et al. 2007	Fundel, K., Küffner, R., & Zimmer, R. (2007). RelEx-relation extraction using dependency parse trees. Bioinformatics (Oxford, England), 23(3), 365–371.
Gerber et al. 2013	Gerber, D., Hellmann, S., Bühmann, L., Soru, T., Usbeck, R., & Ngonga Ngomo, A.-C. (2013). Real-time RDF extraction from unstructured data streams. In Advanced Information Systems Engineering (pp. 135–150). Springer Berlin Heidelberg.
Gerber et al. 2012	Gerber, D., & Ngomo, A.-C. N. (2012). Extracting multilingual natural-language patterns for RDF predicates. In Lecture Notes in Computer Science (pp. 87–96). Springer Berlin Heidelberg.
Giachelle et al. 2023	Giachelle, F., Marchesin, S., Silvello, G., & Alonso, O. (2023). Searching for reliable facts over a medical knowledge base (demo). In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23) (pp. 3205–3209). ACM. https://doi.org/10.1145/3539618.3591822
Goyal et al. 2025	Goyal, N., & Singh, N. (2025). Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. Neurocomputing, 618(129171), 129171.
Gu et al. 2022	Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare, 3(1), 1–23.
Gupte et al. 2021	Gupte, A., Sapre, S., & Sonawane, S. (2021). Knowledge graph generation from text using neural machine translation techniques. 2021 International Conference on Communication Information and Computing Technology (ICCICT), 1–8.
Gurulingappa et al. 2012	Gurulingappa, H., Rajput, A. M., Roberts, A., Fluck, J., Hofmann-Apitius, M., & Toldo, L. (2012). Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. Journal of Biomedical Informatics, 45(5), 885–892.
Hassan et al. 2023	Hassan, N. A. A., Seoud, R. A. A. A. A. and Salem, D. A. (2023) Open information extraction methodology for a new curated biomedical literature dataset, International Journal of Advanced Computer Science and Applications 14 (2023).
He et al. 2020	He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with disentangled attention. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2006.03654
Hearst. M.A. 1992	Hearst. M.A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics.

Key	Reference
Herrero-Zazo et al. 2013	Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P., & Declerck, T. (2013). The DDI corpus: an annotated corpus with pharmacological substances and drug-drug interactions. Journal of Biomedical Informatics, 46(5), 914–920.
Hou et al. 2019	Hou, Y., Jochim, C., Gleize, M., Bonin, F., & Ganguly, D. (2019). Identification of tasks, datasets, evaluation metrics, and numeric scores for scientific leaderboards construction. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1906.09317
Huang et al. 2024	Huang, M.-S., Han, J.-C., Lin, P.-Y., You, Y.-T., Tsai, R. T.-H., & Hsu, W.-L. (2024). Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource. Briefings in Bioinformatics, 25(3). https://doi.org/10.1093/bib/bbae132
Huguet Cabot et al. 2021	Huguet Cabot, P.-L., & Navigli, R. (2021). REBEL: Relation Extraction By End-to-end Language generation. Findings of the Association for Computational Linguistics: EMNLP 2021. Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.findings-emnlp.204
Jia et al. 2019	Jia, R., Wong, C., & Poon, H. (2019). Document-level N-ary relation extraction with multiscale representation learning. Proceedings of the 2019 Conference of the North. Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota. https://doi.org/10.18653/v1/n19-1370
Jiang, et al 2017	Jiang, J., Li, X., Zhao, C., Guan, Y., & Yu, Q. (2017). Learning and inference in knowledge-based probabilistic model for medical diagnosis. *Knowledge-Based Systems*, *138*, 58-68.
Keraghel et al. 2024	Keraghel, I., Morbieu, S., & Nadif, M. (2024). Recent advances in named Entity Recognition: A comprehensive survey and comparative study. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2401.10825
Kertkeidkachorn et al. 2018	Kertkeidkachorn, N., & Ichise, R. (2018). An automatic knowledge graph creation framework from natural language text. IEICE Transactions on Information and Systems, E101.D(1), 90–98.
Kireev et al. 2024	Kireev, V., & Vasiliev, F. (2024). Development of a tool for constructing a knowledge graph using Gan. 2024 6th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA), 720–723.
Kohn et al. 2024	Kohn, N., Boleij, A., Ciompi, F., Nebbia, G., & Romanovych, A. (2024). Deliverable 2.3: Linkage and feature extraction from gut-brain, initial evaluation. Zenodo. https://doi.org/10.5281/zenodo.14627764
Kordjamshidi et al. 2015	Kordjamshidi, P., Roth, D., & Moens, M.-F. (2015). Structured learning for spatial information extraction from biomedical text: bacteria biotopes. BMC Bioinformatics, 16(1), 129.
Lafferty et al. 2001	Lafferty, J., McCallum, A. ,Pereira, F. et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data.In Icml, volume 1, page 3. Williamstown, MA, 2001.
Lample et al. 2016	Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Proceedings of the 2016 Conference of the North

Key	Reference
	American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, California. https://doi.org/10.18653/v1/n16-1030
Lee et al. 2020	Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics (Oxford, England), 36(4), 1234–1240.
Lee et al. 2004	Lee, K.-J., Hwang, Y.-S., Kim, S., & Rim, H.-C. (2004). Biomedical named entity recognition using two-phase model based on SVMs. Journal of Biomedical Informatics, 37(6), 436–447.
Li et al. 2017	Li, F., Zhang, M., Fu, G. et al. (2017) A neural joint model for entity and relation extraction from biomedical text. BMC Bioinformatics 18, 198. https://doi.org/10.1186/s12859-017-1609-9
Liu et al. 2019	Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1907.11692
Luo et al. 2022	Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N., & Lu, Z. (2022). BioRED: a rich biomedical relation extraction dataset. Briefings in Bioinformatics, 23(5). https://doi.org/10.1093/bib/bbac282
Luo et al. 2023	Luo, L., Wei, C.-H., Lai, P.-T., Leaman, R., Chen, Q., & Lu, Z. (2023). AIONER: all-in-one scheme-based biomedical named entity recognition using deep learning. Bioinformatics (Oxford, England), 39(5). https://doi.org/10.1093/bioinformatics/btad310
Marchesin and Silvello, 2022	Marchesin, S., & Silvello, G. (2022). TBGA: A large-scale gene-disease association dataset for biomedical relation extraction. *BMC Bioinformatics, 23*, 111. https://doi.org/10.1186/s12859-022-04646-6
Marchesin et al. 2023	Marchesin, S., Menotti, L., Giachelle, F., Silvello, G., & Alonso, O. (2023). Building a large gene expression–cancer knowledge base with limited human annotations. Database: The Journal of Biological Databases and Curation, 2023, baad061. https://doi.org/10.1093/database/baad061
Martinelli et al. 2025	Martinelli, M., Silvello, G., Bonato, V., Di Nunzio, G. M., Ferro, N., Irrera, O., Marchesin, S., Menotti, L., & Vezzani, F. (2025). Overview of GutBrainIE@CLEF 2025: Gut-Brain Interplay Information Extraction. In G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), *CLEF 2025 Working Notes, Ceur-Ws.*
McCallum et al. 2003	McCallum, A., & Li, W. (2003). Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 -. the seventh conference, Edmonton, Canada. https://doi.org/10.3115/1119176.1119206
Melnyk et al. 2021	Melnyk, I., Dognin, P., & Das, P. (2021). Grapher: Multi-Stage Knowledge Graph Construction using Pretrained Language Models. https://openreview.net/forum?id=N2CFXG8-pRd
Melnyk et al. 2022	Melnyk, I., Dognin, P., & Das, P. (2022). Knowledge Graph Generation From Text. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2211.10511

Key	Reference
Meng et al. 2021	Meng, Y., Zhang, Y., Huang, J., Wang, X., Zhang, Y., Ji, H., & Han, J. (2021). Distantly-supervised named entity recognition with noise-robust learning and language model augmented self-training. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic. https://doi.org/10.18653/v1/2021.emnlp-main.810
Meng et al. 2024	Meng, Z., Zhan, S., Xu, R., Mayer, W., Zhu, Y., Zhang, H.-Y., He, C., He, K., Cheng, D., & Feng, Z. (2024). Domain ontology-driven knowledge graph generation from text. ACM Transactions on Probabilistic Machine Learning. https://doi.org/10.1145/3708478
Meyer et al. 2001	Meyer, I. (2001). Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), Recent Advances in Computational Terminology (pp. 279–302). Amsterdam: John Benjamins.
Miwa and Bansai, 2016	Miwa, M and Bansal, M. 2016. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1105–1116, Berlin, Germany. Association for Computational Linguistics.
Muñoz et al. 2014	Muñoz, E., Hogan, A., & Mileo, A. (2014, February 24). Using linked data to mine RDF from wikipedia's tables. Proceedings of the 7th ACM International Conference on Web Search and Data Mining. WSDM 2014: Seventh ACM International Conference on Web Search and Data Mining, New York, New York USA. https://doi.org/10.1145/2556195.2556266
Nédellec et al. 2005	Nédellec, C. (2005). Learning language in logic - genic interaction extraction challenge.
Nentidis et al. 2025	Nentidis, A., Katsimpras, G., Krithara, A., Krallinger, M., Rodríguez-Ortega, M., Rodriguez-López, E., Loukachevitch, N., Sakhovskiy, A., Tutubalina, E., Dimitriadis, D., Tsoumakas, G., Giannakoulas, G., Bekiaridou, A., Samaras, A., Di Nunzio, G. M., Ferro, N., Marchesin, S., Martinelli, M., Silvello, G., & Paliouras, G. (2025). Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In G. Faggioli, N. Ferro, P. Rosso, & D. Spina (Eds.), *Lecture Notes in Computer Science*. Springer.
Névéol et al. 2011	Névéol, A., Islamaj Doğan, R., & Lu, Z. (2011). Semi-automatic semantic annotation of PubMed queries: a study on quality, efficiency, satisfaction. Journal of Biomedical Informatics, 44(2), 310–318.
Névéol et al. 2007	Névéol, A., Shooshan, S. E., & Humphrey, S. M. (2007). Multiple approaches to finegrained indexing of the biomedical literature. Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 292-e303.
Nguyen et al. 2019	Nguyen, D. T., Mummadi, C. K., Ngo, T. P. N., Nguyen, T. H. P., Beggel, L., & Brox, T. (2019). SELF: Learning to filter noisy labels with self-ensembling. In arXiv [cs.CV]. arXiv. http://arxiv.org/abs/1910.01842

Key	Reference
Oralbekova et al. 2024	Oralbekova, D., Mamyrbayev, O., Zhumagulova, S., & Zhumazhan, N. (2024). A comparative analysis of LSTM and BERT models for named entity recognition in Kazakh language: A multi-classification approach. In Communications in Computer and Information Science (pp. 116–128). Springer Nature Switzerland.
Park et al. 2024	Park Y, Son G, Rho M. Biomedical Flat and Nested Named Entity Recognition: Methods, Challenges, and Advances. Applied Sciences. 2024; 14(20):9302. https://doi.org/10.3390/app14209302
Pujara, et al. 2013	Pujara, J., Miao, H., Getoor, L., & Cohen, W. (2013). Knowledge graph identification. In The Semantic Web–ISWC 2013: 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part I 12 (pp. 542-557). Springer Berlin Heidelberg.
Pyysalo et al. 2007	Pyysalo, S., Ginter, F., Heimonen, J., Björne, J., Boberg, J., Järvinen, J., & Salakoski, T. (2007). BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics, 8(1), 50.
Ramakrishnan et al. 2006	Ramakrishnan, C., Kochut, K. J., & Sheth, A. P. (2006). A framework for schema-driven relationship discovery from unstructured text. In Lecture Notes in Computer Science (pp. 583–596). Springer Berlin Heidelberg.
Reimers et al. 2019	Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In arXiv [cs.CL]. https://doi.org/10.48550/ARXIV.1908.10084
Ribeiro et al. 2023	Ribeiro, M. A. O., Gutierrez, M. A., & Nunes, F. L. S. (2023). Improving deep learning shape consistency with a new loss function for left ventricle segmentation in cardiac MRI. 435–440.
Riedl et al. 2018	Riedl, M., & Padó, S. (2018). A named entity recognition shootout for German. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Melbourne, Australia. https://doi.org/10.18653/v1/P18-2020
Rios-Alvarado et al. 2022	Rios-Alvarado, A. B., Martinez-Rodriguez, J. L., Garcia-Perez, A. G., Guerrero-Melendez, T. Y., Lopez-Arevalo, I., & Gonzalez-Compean, J. L. (2022). Exploiting lexical patterns for knowledge graph construction from unstructured text in Spanish. Complex & Intelligent Systems. https://doi.org/10.1007/s40747-022-00805-7
Roller and Erk 2016	Roller, S. and Erk, K. 2016. PIC a Different Word: A Simple Model for Lexical Substitution in Context. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1121–1126, San Diego, California. Association for Computational Linguistics.
Rotmensch et al. 2017	Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports*, *7*(1), 5994.
Sänger et al. 2021	Sänger, M., & Leser, U. (2021). Large-scale entity representation learning for biomedical relationship extraction. Bioinformatics (Oxford, England), 37(2), 236–242.

Key	Reference
Sanh et al. 2019	Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1910.01108
Saponara et al. 2022	Saponara, M. (2022). Hybrid Neural Knowledge Graph-to-Text and Text-to-Text Generation [Laurea, Politecnico di Torino]. https://webthesis.biblio.polito.it/21933/
Sateli et al. 2015	Sateli, B., & Witte, R. (2015). Automatic construction of a semantic knowledge base from CEUR workshop proceedings. In Semantic Web Evaluation Challenges (pp. 129–141). Springer International Publishing.
Savova et al. 2010	Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association: JAMIA, 17(5), 507–513.
Schweter et al. 2019	Schweter, S., & Baiter, J. (2019). Towards robust named entity recognition for Historic German. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1906.07592
Segura et al. 2013	Segura, B. I., Martínez, P., Zazo, H., & Semeval, M. (2013). SemEval-2013 Task 9: Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013) (Vol. 9, pp. 341–350).
Segura-Bedmar et al. 2011	Segura-Bedmar, I., Inez, P. M. ´., & Sánchez-Cisneros, D. (2011). The 1st DDIExtraction-2011 challenge task: Extraction of Drug-Drug Interactions from biomedical texts. Proceedings of the 1st Challenge Task on Drug-Drug Interaction Extraction, Huelva Spain, 761, 1–9.
Shah et al. 2009	Shah, N., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., & Musen, M. (2009). Comparison of concept recognizers for building the Open Biomedical Annotator. BMC Bioinformatics, 10, S14–S14.
Tanenblatt et al. 2010	Tanenblatt, M., Coden, A., & Sominsky, I. (2010). The ConceptMapper approach to named entity recognition. International Conference on Language Resources and Evaluation.
Tinn et al. 2021	Tinn, R., Cheng, H., Gu, Y., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2021). Fine-tuning large neural language models for biomedical natural language processing. In arXiv [cs.CL]. arXiv. https://doi.org/10.48550/ARXIV.2112.07869
Vaswani et al. 2017	Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/1706.03762
Wang et al. 2023	Wang, Z., Zhao, Z., Chen, Z., Ren, P., de Rijke, M., & Ren, Z. (2023). Generalizing few-shot named entity recognizers to unseen domains with type-related features. In arXiv [cs.IR]. arXiv. http://arxiv.org/abs/2310.09846
Xiao et al. 2022	Xiao, G., Pfaff, E., Prud'hommeaux, E., Booth, D., Sharma, D. K., Huo, N., Yu, Y., Zong, N., Ruddy, K. J., Chute, C. G., & Jiang, G. (2022). FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP Common data Model. Journal of Biomedical Informatics, 134(104201), 104201.

Key	Reference
Yang et al. 2023	Yang, X., Saha, S., Venkatesan, A., Tirunagari, S., Vartak, V., & McEntyre, J. (2023). Europe PMC annotated full-text corpus for gene/proteins, diseases and organisms. Scientific Data, 10(1), 722.
Yasunaga et al. 2022	Yasunaga, M., Leskovec, J., & Liang, P. (2022). LinkBERT: Pretraining language models with document links. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland. https://doi.org/10.18653/v1/2022.acl-long.551
Zaratiana et al. 2024	Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2024). GLiNER: Generalist model for named entity recognition using bidirectional transformer. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico. https://doi.org/10.18653/v1/2024.naacl-long.300
Zhou et al. 2020	Zhou, W., Huang, K., Ma, T., & Huang, J. (2020). Document-level relation extraction with Adaptive Thresholding and Localized cOntext Pooling. In arXiv [cs.CL]. arXiv. http://arxiv.org/abs/2010.11304