8-1-2024

Harnessing NLP and Large Language Models for Pattern Discovery and Information Extraction in Electric Health Reports

Mina Esmail Zadeh Nojoo Kambar

HARNESSING NLP AND LARGE LANGUAGE MODELS FOR PATTERN DISCOVERY AND INFORMATION EXTRACTION IN ELECTRONIC HEALTH REPORTS

By

Mina Esmail Zadeh Nojoo Kambar

Bachelor of Science - Information Technology
Azad University
2014

Master of Science - Information Technology
Urmia University
2017

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy - Computer Science

Department of Computer Science
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
August 2024

UNLV | GRADUATE COLLEGE

Dissertation Approval

The Graduate College
The University of Nevada, Las Vegas

July 9, 2024

This dissertation prepared by

Mina Esmail Zadeh Nojoo Kambar

entitled

Harnessing NLP and Large Language Models for Pattern Discovery and Information Extraction in Electric Health Reports

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy –Computer Science
Department of Computer Science

Kazem Taghva, Ph.D.
Examination Committee Chair

Laxmi Gewali, Ph.D.
Examination Committee Member

Wolfgang Bein, Ph.D.
Examination Committee Member

Mingon Kang, Ph.D.
Examination Committee Member

Emma Regentova, Ph.D.
Graduate College Faculty Representative

Alyssa Crittenden, Ph.D.
Vice Provost for Graduate Education & Dean of the Graduate College

Abstract

In this work, we report on a series of natural language processing tools and models to improve the efficiency and accuracy of information discovery from clinical trials and pharmacological studies. Our main contributions are:

1. **The development of an open-source platform Tri-AL that**
    - Enables dynamic tracking of clinical trials information over time,
    - Excels in data visualization and user interaction with a particular emphasis on enhancing the analysis and representation of race and ethnicity data to foster equity in clinical research, and
    - Includes a predictive model utilizing machine learning to decipher drug mechanisms of action.
2. **Heterogeneous Graph Neural Network for Gene-Chemical Entity Relation Extraction**: We created a supervised deep learning model that adapts a heterogeneous Graph Neural Network to extract gene-chemical components. This model augments word representations using message passing that accurately identifies gene-chemical named entities and their relationships class.
3. **Bipartite Graph Model for Evaluating Summarization Performance**: We proposed a bipartite graph model to evaluate the performance of large language models in summarizing clinical trials. This model provides a robust framework to assess the accuracy and effectiveness of automated summarization tools in the medical domain.

Acknowledgements

I am deeply grateful to my advisor, Dr. Kazem Taghva, for his exceptional impact on my academic and personal development. His unwavering support and guidance throughout my Ph.D. study have been invaluable. Since he first offered me the chance to work with him in 2019, he has continually provided help and encouragement. His expertise, advice, and motivation have significantly shaped my growth. The knowledge and insights he provided will remain with me forever.

I am also thankful to my professors and committee members, Dr. Laxmi Gewali, Dr. Wolfgang Bein, Dr. Mingon Kang, and Dr. Emma Regentova, for equipping me with knowledge and skills in computer science and for their roles on my dissertation committee.

Finally, my heartfelt thanks go to my husband, whose love, patience, and unwavering support have been my emotional anchor throughout this journey. His belief in my abilities and constant encouragement has been a source of motivation that propelled me to persevere even during the most challenging times.

I also want to express my appreciation to my dear friends, especially Pouyan Nahed, my research mate, who has been beside me throughout this academic endeavor, providing mutual academic support.

Mina Esmail Zadeh Nojoo Kambar

University of Nevada, Las Vegas

August 2024

Table of Contents

List of Tables

List of Figures

List of Algorithms

List of Queries

Chapter 1

Introduction

Over the past few decades, the healthcare industry has experienced an exponential increase in the volume of biomedical information, including an extensive array of medical data and scientific literature. This data is syntactically and semantically heterogeneous, varying significantly in structure and meaning depending on the context in which it is used [PPS+23]. Research study [GGS23] classifies the medical data into several key categories:

- **Electronic Health Records (EHRs)**: Comprehensive digital records of a patient's medical history, including diagnoses, treatments, medications, immunization records, allergies, and laboratory test results.
- **Medical Imaging Data**: Visual data from imaging studies used to diagnose and monitor medical conditions, such as X-rays, CT scans, MRI scans, and ultrasound images.
- **Laboratory Test Results**: Data from tests performed on samples of blood, urine, tissue, or other substances to detect or monitor diseases, including blood tests, urine tests, and biopsy results.
- **Genomic Data**: Information about an individual's genetic makeup, including DNA sequences and genetic variations, obtained through methods like whole genome sequencing.
- **Clinical Trials Data**: Data collected during clinical research studies to evaluate the effectiveness and safety of medical interventions, including participant demographics, intervention details, outcomes, and adverse events.
- **Behavioral and Social Determinants Data**: Information about behaviors and social

factors that affect health outcomes, such as smoking status, alcohol consumption, race or ethnicity, and education level.

- **Clinical Notes and Narrative Text**: Unstructured text data from clinician documentation and patient communication, including progress notes, clinical summaries, and referral letters.
- **Sensor and Wearable Data**: Continuous or periodic health data collected from sensors and wearable devices, such as activity levels from fitness trackers, sleep patterns from smartwatches, and continuous glucose monitoring data.

Despite the richness of this data, extracting concise and actionable information from such resources remains one of the most significant challenges in the healthcare community. The variability in data formats, terminologies, and contexts can complicate data integration and analysis [SASK23].

1.1 The Impact of NLP on Medical Data

Natural Language Processing (NLP) has become an essential technology in medical data analysis. It offers benefits by enabling the extraction and interpretation of complex unstructured text data from various sources such as medical records, clinical trials, and other healthcare-related documents. By automating the data extraction process, NLP significantly enhances the efficiency and accuracy of data handling, which in turn, delivers real-world results by boosting research capabilities. This improvement includes patient care and support for evidence-based medical practices. NLP enhances clinical decision support by extracting and presenting critical information in a format that is easily interpretable for healthcare providers. Real-time access to this information enables clinicians to make informed decisions quickly to improve patient care outcomes. For example, NLP has been used to flag potential drug interactions or to highlight important trends in a patient's medical history that might require immediate attention. Figure 1.1 illustrates the various NLP areas in the biomedical domain.

1.2 Key Challenges in Clinical Trials and Electronic Health Records Analysis

In this work, we focus on several critical challenges associated with clinical trials and electronic health records (EHRs). These challenges are fundamental obstacles to improving data analysis and utilization in clinical research.

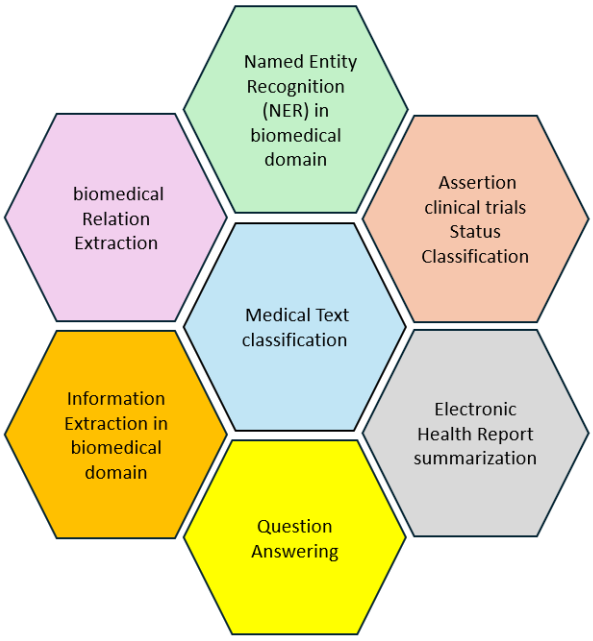


Figure 1.1: NLP Areas on Medical Data

Tracking Clinical Trials Over Time: Monitoring the progress and outcomes of clinical trials over extended periods is challenging due to the vast amount of data generated at different stages. Ensuring that this data remains accurate, organized, and accessible is crucial for longitudinal studies and decision making about ongoing and future research [MBR+23] [CLN+22].

Automating Information Extraction from Clinical Trials: Manual extraction of relevant information from clinical trials is time-consuming and error prone. The diversity in report formats and the complexity of the language used further complicate this task. Automating the extraction process is essential to enhance efficiency and accuracy for analysis and decision-making [ABCED23].

Supporting Diversity and Inclusion in Clinical Trials: Ensuring that clinical trials include diverse and representative populations is vital to generalize the findings. It requires careful analysis of demographic data and targeted efforts to recruit underrepresented groups, which are often overlooked in clinical research [BSM+23] [MMMG23].

Measuring the Reliability of Applying Large Language Models for Clinical Trial Summarization: The use of large language models (LLMs) to summarize clinical trial data presents a significant challenge in accuracy. Summarization must capture essential details while maintaining the information of the original data. Evaluating the performance of LLMs in this context is critical to ensure that these tools can reliably assist in synthesizing complex medical

information [TTE+23] [VVVUB+23].

1.3 Motivation

NLP is a critical technology in clinical research, extraction of biomarkers, analysis of test results, diversity representation, and improvement of decision-making. These concepts are critical to advance medical research and enhance patient care. These often need more comprehensive tools to manage and analyze the vast amounts of unstructured data found in clinical trials and electronic health records (EHRs). Although there are models and systems available to address these challenges, there is currently a lack of comprehensive open-source systems that integrate all the necessary NLP tools to tackle these issues effectively. In this work, we aim to bridge this gap by combining a set of NLP tools designed to implement solutions that improve upon previous methods. By leveraging these tools, we seek to automate the extraction of crucial data points from clinical trials, ensuring that information such as biomarkers and test results are accurately captured and easily accessible. Additionally, our approach addresses the need for more robust demographic data analysis to ensure diversity and inclusion in clinical research. Through these efforts, we aim to enhance the precision and efficiency of data-driven decision-making in healthcare, ultimately contributing to more equitable and effective medical research and practice.

1.4 Contributions

We have made several significant contributions to overcome the challenges outlined in the previous section.

- **Tri-AL, An Open Source System for Tracking Clinical Trials**: We have developed an open-source system called Tri-AL that tracks clinical trials on ClinicalTrials.gov over time. Tri-AL, with its module for analyzing the race and ethnicity of participants in clinical trials, provides an interactive interface and data visualization tools that can be directly applied in exploring data features.

- **Predicting Drug Mechanisms of Action**: We have developed a supervised predictive model that can determine the mechanisms of action for various drugs by leveraging machine learning, deep learning, and language models. This model enhances our understanding of drug functions and their interactions.

- **Heterogeneous Graph Neural Network for Gene-Chemical Entity Relation Extraction**: We created a supervised Deep Learning (DP) model that adapts a heterogeneous Graph Neural Network (GNN) to extract gene-chemical components. This model augments word representations using message passing in the GNN and accurately identifies gene-chemical named entities and their relationships class.

- **Bipartite Graph Model for Evaluating Summarization Performance**: We propose a bipartite graph model to evaluate the performance of large language models in summarizing clinical trials. This model provides a robust framework for assessing the accuracy and effectiveness of automated summarization tools in the medical domain.

In this dissertation, we present our work across several detailed chapters. In Chapter 2, we provide a comprehensive review of related work in the field, focusing on the use of NLP to address the challenges mentioned above. This chapter covers various approaches and methodologies previously employed, highlighting their strengths and limitations. Chapter 3 provides details of the medical databases and information sources used in this research including specifics on data selection, preprocessing, and data integration. In Chapter 4, we describe the Tri-AL open-source system, elaborating on its architecture, functionalities, and the different modules it comprises, such as those for tracking clinical trials and analyzing participant demographics. Chapter 5 delves into the DP model we adapted for medical data, specifically for the extraction of gene-chemical named entities and their relationships, using a supervised deep learning approach with a heterogeneous graph neural network. Chapter 6 introduces our innovative bipartite graph solution designed to evaluate the performance of large language models in summarizing clinical trials, explaining the model's structure and evaluation metrics. Finally, in the last section, we conclude our findings, discuss the implications of our work, and suggest future research directions to further advance the field.

Chapter 2

Background

Many NLP systems have been developed for electronic health records and medical datasets. This section provides a comprehensive summary of related work and identifies the gaps that our research aims to address. Initially, we review existing work on ClinicalTrials.gov, focusing on frameworks designed to extract information, track clinical trial progression, and analyze the participants' race/ethnicity. We then delve into the tasks of Named Entity Recognition (NER) and Relation Extraction (RE), highlighting the importance of performing these tasks jointly. We formulate NER and RE tasks and overview their previous research on biomedical textual data, showcasing the efforts and methodologies applied to these tasks. Additionally, we examine prior work involving the use of large language models to summarize clinical and biomedical textual data, underscoring the necessity of having robust and thorough validation methods to ensure the accuracy and reliability of the summaries. Finally, we outline the metrics employed in this dissertation to evaluate the models, ensuring a rigorous assessment of their performance. Through this review, we aim to set the stage for our contributions and demonstrate how this dissertation addresses the existing gaps in the field.

2.1 Information Extraction on Clinical Trial Data

Numerous NLP systems have been developed to automate the extraction of information from clinical trial data, particularly addressing the long, unstructured textual data features in clinical trials, such as eligibility criteria. These criteria, which are divided into inclusion and exclusion categories, define the specific characteristics that determine whether individuals can participate in a study. Inclusion criteria specify participants' traits, such as age range, gender, disease type and stage, health status, previous treatments, and other demographic factors. In contrast, exclusion

criteria outline characteristics that disqualify potential participants, including comorbid conditions, contraindications, specific previous treatments, known allergies or adverse reactions, pregnancy, active substance abuse, and other factors that might compromise the study's integrity or participant safety. [SCH23] highlights that inappropriate criteria can lead to insufficient recruitment, which is a common reason for the failure of many clinical trials. By automating the extraction of these criteria, NLP systems enhance the efficiency and accuracy of managing clinical trial data [TSM+20]. Previous research in this area is categorized into rule-based and ML-based systems

Rule-based Systems: Rule-based systems are a type of artificial intelligence (AI) system that uses predefined logical rules to process data and make decisions. These systems rely on a set of "if-then" rules, which are created by domain experts, to interpret and analyze input data to produce an output. Studies [HLW16], [WWL+11], [TPC+11], and [BTC+12] are rule-based proposed systems to automate the information extraction from clinical trials eligibility criteria. For instance, Valx [HLW16] is a system with an automated method designed to extract and normalize numeric lab test comparison statements. The system leverages semantic knowledge from the Unified Medical Language System (UMLS) [Bod04] and domain knowledge from the Internet. Valx operates through a seven-step process, including text preprocessing, numeric and unit extraction, variable identification, association filtering, measurement unit normalization, and heuristic rule-based verification. An example of this rule-based system is this sentence: "Participants must have a body mass index (BMI) between 20 and 40 kg/m^2 and a fasting blood glucose level of less than 100 mg/dL." Valx processes this by first normalizing special symbols, correcting typos, and identifying sentences with numeric values. It then extracts numeric expressions (20, 40, 100), units (kg/m^2, mg/dL), and comparison operators ("between," "less than"). Using contextual and domain knowledge, Valx identifies the variables "body mass index (BMI)" and "fasting blood glucose level." Next, it associates these numeric values with their respective variables and units, confirming "BMI" with "20 kg/m^2" and "40 kg/m^2" and "fasting blood glucose level" with "¡ 100 mg/dL". Valx then verifies these associations through contextual information, ensuring unit consistency and correcting any missing units. Finally, heuristic rules are applied to verify the structured numeric comparison statements. The system was evaluated using clinical trial data for Type 1 and Type 2 diabetes. Another rule-based system, EliXR system pipeline [WWL+11], developed to automate the semistructured information extraction from clinical research eligibility criteria, integrates syntactic parsing and tree pattern mining to discover common semantic patterns in texts from ClinicalTrials.gov. EliXR combines the Unified Medical Language System's (UMLS) semantic knowledge. This approach

results in 175 semantic patterns forming 12 semantic role labels in a semantic network. Evaluated by three independent raters on 396 sentence segments from 79 eligibility criteria, EliXR achieved a high Fleiss' kappa score of 0.88; Fleiss' kappa score is a statistical measure used to assess the reliability of agreement between multiple raters or judges when they categorize items into mutually exclusive categories.

Hybrid Systems exist to extract critical information from electronic health records (EHRs) and clinical trials. The rule-based component can identify specific medical terms and predefined patterns (e.g., medication dosages, diagnostic codes), while the ML component analyzes the context to recognize more complex relationships (e.g., identifying symptoms related to a condition based on context). The hybrid systems integrated approach ensures comprehensive and accurate extraction of critical patient data. As an example, the Criteria2Query [YRT+19] system utilizes a hybrid information extraction pipeline combining machine learning and rule-based methods to parse eligibility criteria text, transforming it into structured, computable representations that can be executed as SQL queries within clinical databases. The system's modular architecture includes a systematic information extraction pipeline, a query formulation pipeline, and an interface for interactive query review and execution in the ATLAS web application. Criteria2Query was evaluated using 125 criteria from ClinicalTrials.gov and 52 user-entered criteria, achieving F1 scores of 0.795 for entity recognition and 0.805 for relation extraction, with high accuracy in negation and logic detection. The system demonstrated its effectiveness in translating free-text criteria into executable queries, significantly reducing the manual effort required for cohort definition and improving the reproducibility and accuracy of clinical research.

As another hybrid tool, [TSM+20] proposes system frames eligibility criteria extraction as a knowledge base population task and combines machine learning with context-free grammar (CFG) techniques. The methodology involves a seven-step pipeline, including UMLS-based lexicon discovery, semantic term annotation, sentence categorization, syntactic parsing, semantic pattern mining, aggregation, and semantic role labeling. The system implements attention-based conditional random field (CRF) architecture for named entity recognition (NER) and word2vec embedding clustering for named entity linking (NEL). The contributions of this work include achieving a 0.753 end-to-end accuracy, implementing the first attention-based NER for criteria extraction with high precision and recall, and creating a dataset of 121,221 clinical entities, attributes, and limits. The system competes with existing tools like Criteria2Query and provides an open-source library for further research and development.

Machine Learning-based Systems: Machine learning (ML), language models, and large language model systems for information extraction have become popular in handling the complex and unstructured data found in clinical trial eligibility criteria or even the entire trials. For example, Research conducted by [GGS+21] proposes and evaluates the effectiveness of the ExaCT tool in semi-automating data extraction for systematic reviews of randomized trials. The study highlights the challenge of conducting timely systematic reviews due to the rapid increase in clinical trial publications and the labor-intensive nature of manual data extraction. ExaCT integrates machine learning and text mining to identify and extract relevant data elements from full-text trial publications. The methodology involved a prospective evaluation on a sample of 75 randomized trials, with manual extraction and verification of 21 data elements by three reviewers to establish a reference standard. ExaCT then processed these trials, identifying the presence of data elements and extracting relevant sentences and fragments. The study measured the tool's extraction accuracy, the relevance of provided sentences, and time savings compared to manual extraction. The results showed that ExaCT correctly identified the reporting status of data elements with a median accuracy of 91%, and at least one of the top five sentences provided by the tool was relevant in 88% of cases. Pertinent fragments were highlighted with a median relevance of 90%, and entirely correct solutions were provided for 48% of data elements. Using ExaCT resulted in modest time savings, with a total extraction time of 17.9 hours for 75 trials compared to 21.6 hours for manual extraction.

Certain research studies [DBR19], [SMD+19], citehahn2020medical categorize papers in the field of NLP techniques applied to clinical trials. Review study [HO20] focuses on the paradigm shift from traditional Machine Learning (ML) techniques to Deep Neural Networks (DNNs). The study highlights how DNNs have become the dominant approach for tasks such as named entity recognition (NER) and relation extraction (REX) in the medical field, particularly for extracting information about diseases and medications. The authors describe the advantages of DNNs over previous ML methods, such as eliminating the need for manual feature selection and achieving significantly higher performance metrics. Despite these advancements, the paper also acknowledges challenges such as the need for large annotated datasets, the complexity of medical sublanguages, and the computational demands of training DNNs. The survey emphasizes that while DNNs have revolutionized medical NLP, they also present new challenges that require innovative solutions, including transfer learning and domain adaptation to handle medical-specific language nuances better.

2.2 Tracking Clinical Trials Progression

Tracking the history of clinical trials is crucial for ensuring data integrity, regulatory compliance, transparency, and accountability. It allows for accurate verification of data, adherence to ethical and legal standards, and builds trust in research findings. Proper tracking facilitates the replication and validation of studies, continuous monitoring of patient safety and efficacy, and aggregation of data for meta-analyses, providing robust evidence on treatments. Additionally, it informs future research designs, enhances their efficiency, and protects participants' rights and well-being. Overall, it maintains the quality, reliability, and ethical standards of clinical research. Research papers [Car22] and [SSM22] have provided frameworks to track the history of clinical trials while also ensuring data security. [SSM22] presents a method to automate health technology assessment (HTA) processes using R programming, focusing on maintaining data security and enhancing efficiency. HTA involves evaluating medical, social, economic, and ethical issues related to health technology use, which often require sensitive data, posing challenges for data sharing and model development. The pipeline consists of three parts: an economic model constructed with pseudo data, an API hosted on a server containing sensitive data, and an automated workflow that calls the API, retrieves results, and generates a report. The pipeline utilizes R packages like Plumber for creating APIs and RMarkdown for report generation. The automated workflow enhances data security by ensuring that sensitive data remains with the data owner, reducing the risk of data breaches. It allows for scheduled or event-triggered updates, improving efficiency and reducing manual intervention. The separation of data and model enhances transparency, allowing for greater scrutiny and validation of the model. Additionally, handling the computational burden on a remote server speeds up the analysis. The authors provide an example user interface built with the shiny package, enabling non-technical stakeholders to interact with the model. The method has been validated using example data and is available as open-source code on GitHub, inviting further collaboration for validation and improvement. Research papers [Car22] and [SSM22] have provided frameworks for tracking the history of clinical trials while also ensuring data security. [Car22] introduces a novel R package, cthist, designed to automate the extraction of historical data from clinical trial registries, specifically ClinicalTrials.gov and DRKS.de. The motivation behind developing cthist stems from the challenges in accessing historical registry data, which traditionally required manual, labor-intensive efforts. This limitation hindered the feasibility, accuracy, and reproducibility of certain types of research, such as assessing changes in clinical trial enrollment goals over time

or evaluating modifications in trial protocols. The cthist package provides six main functions to facilitate this data extraction, offering users a streamlined process to retrieve and analyze historical versions of clinical trial entries. The methodology involves web scraping techniques to collect data on various aspects of clinical trials, including recruitment status, start and completion dates, enrollment figures, and outcome measures. Three case studies demonstrate the utility of cthist: assessing changes in recruitment period lengths, identifying outcome measure modifications, and correcting for variable follow-up times in meta-research.

Previous methods for tracking clinical trials lack visualization capabilities, customization for specific diseases, and the ability to be extended with machine learning modules for other tasks. However, Tri-AL, as explained in Chapter 4, addresses these needs by offering a comprehensive tool that provides robust visualization, disease-specific customization, and integration with machine learning for enhanced functionality.

2.3 Analyzing Race and Ethnicity

Having a system that provides analysis on the race and ethnicity of clinical trial participants on ClinicalTrials.gov is crucial for ensuring diversity and inclusivity in medical research. Such a system would address the current underrepresentation of minority groups, enhancing the generalizability and reliability of clinical trial outcomes. By systematically tracking and reporting demographic data, researchers and policymakers can identify and mitigate disparities, ensuring that all population segments benefit from advancements in healthcare.

There is much research [FNTW21], [KSMB21], [TSW$^+$22], and [XVL$^+$23] working on race/ethnicity analysis in clinical trial data. For example, the research conducted by [TSW$^+$22] examines the historical and current state of racial and ethnic diversity in US clinical trials. The study analyzed detailed records from all US clinical trials registered in ClinicalTrials.gov between March 2000 and March 2020. The reporting of race and ethnicity for clinical trial enrollees is not strictly required for all trials on ClinicalTrials.gov. However, implementing Section 801 of the Food and Drug Administration Amendments Act (FDAAA 801) in September 2007 and the Final Rule in January 2017 increased the number of trials required to report this information [ZTWC16]. Consequently, the proportion of trials reporting race and ethnicity data on ClinicalTrials.gov rose significantly after the 2007 establishment of the results database. Between 2008 and 2018, the reporting of any race/ethnicity enrollment data increased from 26% (599 out of 2,334) to 91% (194 out of 213), with an annual growth rate of 13.5%. The key findings indicate that only 43% of the 20,692 US-based

trials with reported results included any race/ethnicity data. Among the trials that were reported, the majority of participants were White (median 79.7%), with significantly lower representation of Black (10%), Hispanic/Latino (6%), Asian (1%), and American Indian (0%) participants compared to their respective proportions in the US population. The study found that industry and academic funding were negatively associated with race/ethnicity reporting, while US government-funded trials showed higher reporting rates and greater diversity among enrollees. Over the two decades, there was a modest annual increase of 1.7% in the enrollment of minority groups. The lack of diversity in clinical trials contributes to a data gap, which skews medical evidence and innovation, potentially leading to biased therapeutic outcomes for minority populations. The study highlights the need for improved and standardized reporting of race/ethnicity data to ensure the equitable representation of all demographic groups in clinical research.

Study [XVL+23] provides a systematic review and meta-analysis of the demographic representation in U.S.-based COVID-19 clinical trials. The study analyzed data from 122 trials with 176,654 participants, focusing on including female, racial, and ethnic minority individuals. The findings reveal that female participants and Black and Asian individuals were underrepresented in specific trials, while Hispanic or Latino participants were often overrepresented, particularly in treatment trials. The study underscores the ongoing challenges in achieving equitable representation in clinical trials.

We designed a module within Tri-AL to address these challenges and provide a comprehensive tool for analyzing participants' race and ethnicity on ClinicalTrials.gov. This module enables detailed tracking and reporting of demographic data, ensuring greater transparency and inclusivity in clinical trials. Tri-AL helps identify and mitigate disparities in participant representation, by visualizing and representing statistics of race and ethnicity of trials' participants. We explain this feature of the Tri-AL system in Chapter 4.

2.4 Named Entity and Relation Extraction

Information Extraction (IE) is fundamental to knowledge-based systems. It leverages NLP techniques to uncover hidden information within unstructured texts. These systems are crucial in converting raw text data into structured knowledge, enabling more effective information retrieval, data analysis, and decision-making processes [Kum17].

Over the years, a multitude of knowledge-based systems have emerged, each tailored to assist professionals in diverse fields. For example, a knowledge-based system for troubleshooting PCs

offers automated solutions to identify and resolve computer issues [DAN18]. In the medical domain, a similar system aids in diagnosing diabetes, providing clinicians with valuable insights and decision support [DME+19]. These instances underscore the practicality and impact of knowledge-based systems in various sectors, making the relevance of IE, NER, and RE more tangible.

IE is a complex process that involves two key subtasks: Named Entity Recognition (NER) and Relation Extraction (RE). These tasks, when executed in sequence, play a pivotal role in transforming unstructured data into structured knowledge.

Named Entity Recognition (NER): NER, the first subtask of IE, is the process of identifying and categorizing named entities within a text. These entities can range from common categories like locations, organizations, and persons to more specific ones like drug names, chemical compounds, and biological proteins. NER is a crucial initial step in the development of a knowledge-based system, as it structures the text by highlighting key elements that can be further analyzed [PPB17]. For instance, in a medical document, NER would identify terms like 'insulin,' ' diabetes,' and 'pancreas,'tagging them as relevant entities.

Relation Extraction (RE): RE builds upon the entities identified by NER, focusing on determining the relationships between these entities. RE aims to extract relational triples in the format (Entity 1, Relation type, Entity 2). These triples, often referred to as (subject, relation, object), encapsulate the semantic connections between entities within the text. For instance, in the sentence "Insulin regulates blood sugar levels," an RE system would extract the triple (Insulin, regulates, blood sugar levels). Identifying such relationships is crucial to build comprehensive knowledge graphs that represent complex information in a structured manner [NIR+23]. In another way of classification RE Sequences are categorized into three types [MB16]:

- **Normal Sequences** are those in which there are no triples with common subjects or objects. Each entity pair and their corresponding relationship are unique within the text. For example, consider the medical text: "Aspirin reduces inflammation. Penicillin treats bacterial infections." In this case, there are two distinct triples: (Aspirin, reduces, inflammation) and (Penicillin, treats, bacterial infections).

- **Single Entity Overlapped (SEO) Sequences** contain at least two triples with a common subject or object. This overlap can create complexities in accurately identifying and classifying the relationships due to the shared entities. For instance, in the text: "Aspirin reduces inflammation and is used to prevent heart attacks," the entity "Aspirin" is involved in

two different relationships: (Aspirin, reduces, inflammation) and (Aspirin, is used to prevent, heart attacks).

- **Entity Pair Overlapped (EPO) Sequences** involve at least two triples with the same subject and object but different relation classes. This situation can cause confusion for classifiers, as the same entity pair participates in multiple types of relationships. As an example, consider the text: "Aspirin prevents heart attacks and also reduces the risk of stroke." This text generates the following triples: (Aspirin, prevents, heart attacks) and (Aspirin, reduces the risk of, stroke).

The task of Relation Extraction (RE) can be classified into two main approaches: supervised and unsupervised.

- **Supervised Relation Extraction** relies on labeled training data where experts explicitly annotate the relationships between entities. This approach trains a model on these annotated examples to learn patterns and relationships. Features are extracted from the text, such as lexical, syntactic, and semantic information, to help the model understand the context and nature of the relationships. The trained model can then highly predict relationships in a new, unseen text. Still, this method requires significant annotated data, which can be time-consuming and expensive.

- **Unsupervised Relation Extraction**does not require labeled training data. Instead, it aims to discover relationships directly from the text through clustering, co-occurrence analysis, or pattern mining. This approach leverages the assumption that entities frequently appearing together or in specific patterns are likely related. Techniques such as dependency parsing or Open Information Extraction (OpenIE) [PJC23] are often used to identify these patterns. While unsupervised RE is more accessible when applied to new domains and large datasets, it typically has lower accuracy. It may produce more irrelevant or spurious relationships compared to supervised methods.

In this dissertation, we focus on supervised Relation Extraction (RE). There are two primary types of supervised RE models: Joint Named Entity and Relation Extraction (JNERE) and pipeline methods.

2.4.1 Pipeline Methods

Pipeline methods, such as those described in source [CR11], operate in two distinct phases. The system identifies all possible entity pairs within the text in the first phase. In the second phase, it classifies the relationships between these pairs. However, pipeline methods have several drawbacks:

- **Error Propagation**: Errors in the Named Entity Recognition (NER) subtask can propagate to the Relation Extraction (RE) subtask. For example, if the system misidentifies "Warfarin" (a medication) as a general term rather than a specific drug name, it might incorrectly classify the relationships involving "Warfarin." This could lead to errors in identifying important medical relationships, such as drug interactions or treatment protocols.

- **Class Imbalance**: Many entity pairs in medical texts may not belong to any predefined relation class, leading to a highly imbalanced dataset where the "no relation" class dominates. For example, in a medical document, the majority of entity pairs such as "aspirin" and "headache" might not have a direct relationship, making it difficult for the classifier to accurately learn and distinguish between relevant medical relationships like "treats" or "causes."

- **Complexity**: The large number of potential entity pairs in medical documents increases the complexity of the problem. For instance, in a lengthy medical report, the number of possible pairs, such as "doctor-patient," "drug-disease," and "symptom-treatment," can be vast, making the classification task computationally intensive and challenging to manage effectively.

- **Ambiguity in Relations**: The classifier can become confused when the same subject-object pair can belong to multiple relation classes. For example, in the sentence "Metformin treats diabetes and is prescribed for managing blood sugar levels," the subject-object pair "Metformin-diabetes" has both a treatment and a prescription relationship. This ambiguity can confuse the classifier, leading to incorrect or inconsistent relationship extraction.

These SEO and EPO sequences pose additional challenges for pipeline methods because they require the classifier to handle overlapping and potentially conflicting relations.

2.4.2 Joint Named Entity Relation Extraction (JNERE)

To address these challenges, JNERE was proposed. This approach simultaneously extracts entities and their relationships from the text in a single, unified process. JNERE is often referred to by various names in the literature, including joint named entities relation extraction, triple extraction, and end-to-end relation extraction. By extracting entities and relations together, JNERE avoids the error propagation and class imbalance issues inherent in pipeline methods. It also simplifies handling overlapping sequences by considering the context and relations holistically. In summary, while pipeline methods separate the tasks of entity recognition and relation classification, leading to potential errors and complexity, JNERE integrates these tasks, offering a more robust solution for extracting structured information from unstructured texts.

2.4.3 JNERE Task Formulation

In this dissertation, we focus on chemical components as the first named entity and genes as the second named entity, aiming to jointly predict the relationships between them. In this task, joint entity and relation extraction involves identifying all possible (Chemical, Relation-type, Gene) triples within a text. For simplicity, these triples are denoted as $(\mathrm{ch}, \mathrm{r}, \mathrm{g})$ throughout this discussion. The objective function for this task is defined as follows:

$$\prod_{j=1}^{|D|} \left[\prod_{(\mathrm{ch},\mathrm{r},\mathrm{g}) \in T_j} p((\mathrm{ch}, \mathrm{r}, \mathrm{g}) | x_j) \right] \tag{2.1}$$

where:

- x_j is the tokenized sequence.
- $|D|$ is the number of sequences in the training dataset.
- T_j represents the set of all triples in the j^{th} sequence.
- $p((\mathrm{ch}, \mathrm{r}, \mathrm{g}) | x_j)$ is the probability of identifying the triple $(\mathrm{ch}, \mathrm{r}, \mathrm{g})$ in the tokenized sequence x_j.

By applying the chain rule of probability, we can expand the objective function as follows:

$$\prod_{j=1}^{|D|} \left[\prod_{\mathrm{ch} \in T_j} p(\mathrm{ch} | x_j) \prod_{(\mathrm{r},\mathrm{g}) \in (T_j | \mathrm{ch})} p((\mathrm{r}, \mathrm{g}) | \mathrm{ch}, x_j) \right] \tag{2.2}$$

Further expanding the function, we get:

$$\prod_{j=1}^{|D|} \left[\prod_{\mathrm{ch} \in T_j} p(\mathrm{ch} | x_j) \prod_{\mathrm{r} \in (T_j | \mathrm{ch})} p(\mathrm{g} | \mathrm{ch}, \mathrm{r}, x_j) \prod_{\mathrm{r} \in (R \setminus T_j | \mathrm{ch})} p(\mathrm{g}_{\varnothing} | \mathrm{ch}, \mathrm{r}, x_j) \right] \tag{2.3}$$

where:

- ch $\in T_j$ denotes the chemicals within the triples of sequence j.

- r $\in (T_j|\text{ch})$ represents the relations associated with the chemical ch.

- r $\in (R \setminus T_j|\text{ch})$ indicates all relations excluding those that involve the chemical ch.

- $p(\text{g}|\text{ch}, \text{r}, x_j)$ is the probability of the gene g given the chemical ch, relation r, and sequence x_j.

- The term $p(\text{ch}|x_j)$ represents the probability of identifying a chemical entity in the sequence x_j.

- The term $p((\text{r}, \text{g})|\text{ch}, x_j)$ represents the probability of identifying the relation r and gene g given the chemical ch in the sequence x_j.

- $p(\text{g}_\varnothing|\text{ch}, \text{r}, x_j)$ is the probability of having no relation (null gene) given the chemical ch, relation r, and sequence x_j. Our aim is to maximize this objective function to accurately identify chemical entities, their interactions, and the absence of interactions in the given text [WSW+20].

2.4.4 JNER Task Biomedical Domain Related Work

Several studies [ZZ22] [ZWB+17] [SYW+22] [LYC+20] [ZLC+19b] have applied joint Named Entity Recognition and Relation Extraction (JNERE) in the biomedical domain, showcasing its effectiveness in extracting complex information from scientific texts. The following studies serve as the baselines for our model presented in Chapter 1, providing a basis for comparison with our results. Figure 2.1 illustrates a classification of related work for the named entity and relation extraction task, which we only focus on the JNERE methods.

Study [ZZ22] proposes a novel span-based approach to jointly extract entities and their relations from biomedical text involving bacteria biotopes (BBs). The method addresses the challenge of recognizing nested and discontinuous entities common in the BB corpus. The authors employ a BERT model pre-trained on domain-specific corpora to encode sentences, capturing rich contextual information [DCLT18a]. The span-based model considers all possible spans within a sentence as potential entity mentions, computing relation scores between spans using their representations and the context between them. The model involves several steps: preprocessing the text with ScispaCy [NKBA19] for sentence segmentation and tokenization, using BERT to generate contextual embeddings for each token, and constructing span representations through max pooling over the embeddings within each span. These representations are then used in feed-forward neural networks to predict entity and relation types. A pruning strategy reduces the number of spans considered for

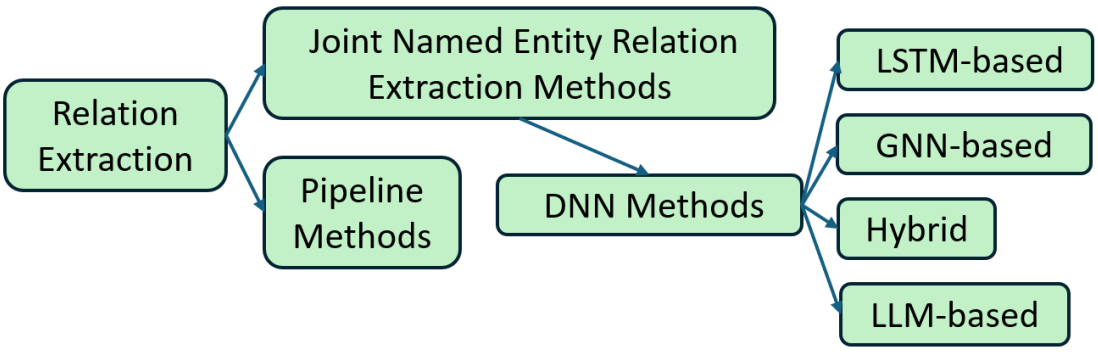


Figure 2.1: Classification of Related Work on NERE Task

relation extraction, focusing on the most likely entity spans. The model utilizes multi-task learning, simultaneously optimizing for entity recognition and relation extraction, enhancing both tasks' performance. Experimental results on the BB-rel+ner 2019 corpus [DBC+16] demonstrate the model's superior performance, significantly reducing the slot error rate (SER) compared to state-of-the-art methods. The model's effectiveness in recognizing nested entities and its generalizability to other datasets, such as the CHEMPROT corpus [KKB+16], are highlighted, showcasing its potential for broader applications in biomedical information extraction.

Research [SYW+22] presents a novel method for extracting biomedical entities and their relations from unstructured literature. The proposed method utilizes a machine reading comprehension (MRC) framework to address the challenge of overlapping triplets, which are common in biomedical datasets. They employ BERT for encoding sentences and introduced a tagging strategy for overlapping triplets. The MRC4BioER model converts the joint extraction task into a sequence of (Query, Context, Answer) tuples, allowing the model to focus on relation-specific semantic features and effectively handle overlapping entities. The model was evaluated on the CHEMPROT and DDIExtraction2013 datasets, demonstrating superior performance to existing joint extraction methods. Specifically, MRC4BioER outperformed baseline methods such as NovelTagging and CasRel, achieving higher F1 scores in both CPI [KKB+16] and DDI extraction tasks [HZSBMD13]. Ablation experiments further confirmed the effectiveness of the proposed tagging scheme and the overall MRC framework, showing significant improvements in handling sentences with multiple

triplets.

The study [LYC+20] proposes a joint learning approach to extract entities and relations from biomedical texts simultaneously. The method employs a novel tagging scheme designed to handle overlapping ties, which are common in biomedical literature. The model integrates an Att-BiLSTM-CRF architecture, which leverages attention mechanisms to enhance the long-distance dependencies between related entities and focuses on critical words for accurate predictions. Additionally, the model uses contextualized ELMo embeddings pre-trained on biomedical texts to improve performance further. The model was evaluated on the DDI and CPI datasets, where it significantly outperformed existing methods. Specifically, it achieved an F1-score of 0.751 on the DDI dataset and 0.551 on the CPI dataset, marking substantial improvements in overlapping relation extraction.

The study [ZWB+17] proposes a novel tagging scheme to convert the joint extraction of entities and their relations into a tagging problem, facilitating the use of end-to-end models without the need for complex feature engineering. The method employs a Bi-LSTM-CRF architecture, leveraging the capabilities of Bi-directional Long Short-Term Memory (Bi-LSTM) networks to encode input sentences and Conditional Random Fields (CRF) to decode the tag sequence. This approach effectively captures long-term dependencies in the data, enhancing the accuracy of entity and relation extraction. The model was evaluated on a public dataset created using a distant supervision method, achieving superior performance compared to traditional pipelined and joint learning methods. Specifically, the proposed model, LSTM-LSTM-Bias, outperformed other models with a precision of 0.615, recall of 0.414, and an F1-score of 0.495, marking a significant improvement over the best existing method, CoType [RWH+17], which had an F1-score of 0.463. The results validate the effectiveness of the novel tagging scheme and the end-to-end model in extracting entities and their relations from unstructured text.

The study [ZLC+19b] presents a novel multi-task learning approach to address the challenge of extracting entity mentions and their relations from domain-specific biomedical texts. The framework consists of a shared transformer encoder for named entity recognition (NER) and relation extraction tasks, followed by separate mention recognition and relation extraction layers. The mention recognition layer uses an enhanced BIOHD tagging schema to handle disjoint and overlapping entity mentions. In contrast, the relation extraction layer predicts relations by considering Microorganism entities as central to all relations. The model is trained using a joint loss function combining both NER and relation extraction losses. The evaluation results underscore the exceptional performance of the proposed model on the Bacteria Biotope (BB) rel+ner subtask. It

not only surpasses traditional pipeline and other joint extraction methods but also demonstrates a significant improvement over baseline methods. The model achieves a Slot Error Rate (SER) of 0.947, precision of 0.493, and recall of 0.339, outperforming the Pipeline method by reducing the SER by 0.525. For specific relation types, the model excels with an SER of 0.954 for all relation classes, 0.982 for Exhibits, 1.318 for Lived in-geo, and 0.927 for Lived in-habitat.

2.5 Large Language Models for Medical Data

LLMs have found extensive applications in clinical trials, revolutionizing diverse processes with their advanced capabilities. One significant use of LLMs is in summarizing complex clinical trial reports, making it easier for researchers and practitioners to grasp essential findings and insights quickly. Many researchers, such as [LHL+24] and [MEMR+24], have utilized LLMs for summarization. Additionally, LLMs excel in information extraction, enabling the automated retrieval of relevant data from vast amounts of clinical trial documents. This includes identifying key variables, patient outcomes, and treatment effects. Notable studies, including those by [LPH+24] and [AAA+23], have demonstrated the effectiveness of LLMs in information extraction. This section describes the previous work in this area and our contribution.

The study [JWF+23] explores the use of LLMs like GPT-3 for clinical information extraction, leveraging prompt-based learning to adapt pre-trained LLMs for tasks such as span identification, token-level sequence classification, and relation extraction. By introducing new annotated datasets based on the CASI dataset[MPL+14], the authors evaluate LLMs on diverse tasks, including clinical sense disambiguation, biomedical evidence extraction, coreference resolution, and medication status extraction. They employ a variety of techniques, including crafting specific prompt templates and mapping LLM outputs to structured label spaces through resolver functions, significantly simplifying the engineering effort required for these tasks. Additionally, weak supervision and model distillation are used to train smaller, task-specific models, enhancing deployability and performance. The results show that GPT-3 [AAA+23], combined with resolvers, outperforms existing zero-shot methods and achieves an accuracy of 0.86 and a macro F1 score of 0.69 in clinical sense disambiguation. For biomedical evidence extraction, resolved GPT-3 demonstrates a token-level F1 score of 0.61 and an abstract-level accuracy of 0.85, significantly higher than supervised baselines. In coreference resolution, GPT-3 with guided prompts achieves a recall of 0.78 and precision of 0.58, surpassing traditional deep learning models. Similarly, GPT-3 excels in medication extraction with a recall of 0.87 and precision of 0.83 and medication status classification with a conditional ac-

curacy of 0.85 and a macro F1 score of 0.69. Adding guided one-shot examples further enhances performance, particularly in complex tasks involving multiple attributes.

The SEETrials system [LPH+24], utilizing GPT-4, was designed to automate the extraction of data from oncology clinical trial abstracts. It comprises four modules: pre-processing, where clinical trial abstracts from ASCO[oCO24], ASH [oH24] conferences, and PubMed are collected, and tables in abstracts are converted to text; knowledge ingestion, which integrates oncology clinical trial background knowledge into prompts to enhance the LLM's analytical capabilities; prompt modeling, which creates tailored prompts to guide the LLM in identifying and merging relevant trial outcomes; and post-processing, where extracted data is refined and structured for subsequent analysis tasks. The system was evaluated using a dataset of 245 multiple myeloma (MM) clinical trial abstracts and 115 abstracts from breast, lung, lymphoma, and leukemia cancers. Performance metrics included a precision of 0.958, a recall (sensitivity) of 0.944, and an F1 score of 0.951. Significant heterogeneity in study outcomes across different therapy phases was noted, with I2 heterogeneity index scores exceeding 75% in several cases. The SEETrials system demonstrated high accuracy and versatility, facilitating nuanced data comparisons and rapid dissemination of clinical insights, potentially enhancing clinical decision-making and evidence synthesis in oncology.

The study [LHL+24], conducted by Pfizer, focused on utilizing LLMs to automate the generation of safety-related table summaries in clinical study reports (CSRs). The methodology involved a challenge where multiple teams used generative pre-trained transformer (GPT) models with prompt engineering to generate summaries from the safety tables of CSRs. The challenge, conducted over six weeks, included training and testing phases using CSRs from a variety of clinical trials. Participants used different techniques for table extraction, prompt engineering, and text generation, and their outputs were evaluated on factual accuracy, lean writing, and provenance by both automated metrics and expert reviewers. Results showed variability in performance across teams, particularly in factual accuracy and semantic similarity, indicating different levels of success in accurately capturing and summarizing safety data. Some teams achieved high comprehension of table structures and generated accurate summaries, while others faced issues like parsing errors and factual inaccuracies. The evaluation highlighted areas for improvement, such as better table ingestion methods, context addition, and fine-tuning of models. The study concluded that while LLMs hold the potential for automating CSR summarization, human involvement remains crucial, and ongoing research is needed to optimize these technologies for broader application in clinical documentation.

In this dissertation, we introduce a novel method utilizing bipartite graphs to quantify the preservation of entity types in summaries generated by LLM compared to the original summaries. This innovative approach allows for a detailed analysis of how well the generated summaries maintain the integrity and accuracy of critical medical entities, such as diseases, genes, and treatments, which are crucial for clinical trials and medical research. We can systematically evaluate the degree of entity type preservation by mapping entities from the original and generated summaries onto a bipartite graph. This contribution not only provides a robust mechanism for assessing the quality of LLM-generated summaries but also offers valuable insights into the potential and limitations of LLMs in medical text summarization.

2.6 Metrics for Evaluation

In this section, we describe the metrics used in our dissertation to evaluate the performance of our models, as well as all other metrics mentioned throughout this document. We focus on crucial evaluation metrics, including Precision, Recall, F1-Score, Slot Error Rate (SER), and ROUGE. Each of these metrics offers a different perspective on the model's performance, ensuring a comprehensive evaluation of the model's effectiveness.

Precision is a measure of the accuracy of the positive predictions made by a model. It is defined as the ratio of correctly predicted positive observations to the total predicted positive observations. High precision indicates a low false positive rate.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \tag{2.4}$$

Recall: also known as sensitivity or true positive rate, measures the model's ability to identify all relevant instances. It is defined as the ratio of correctly predicted positive observations to the all observations in the actual class. High recall indicates a low false negative rate.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{2.5}$$

F1-Score: The F1-Score is the harmonic mean of Precision and Recall. It provides a single metric that balances the trade-off between Precision and Recall. The F1-Score is particularly useful when you need to take both false positives and false negatives into account.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.6}$$

Slot Error Rate (SER): is a metric used to evaluate the performance of information extraction systems, particularly those that extract structured data from unstructured text. In the context of entity and relation extraction tasks, SER measures the number of incorrect, missing, or spurious slots (i.e., pieces of information) extracted by the model. A "slot" typically refers to an entity or a relation between entities.

SER is calculated as follows:

$$\text{SER} = \frac{\text{Substitution Errors} + \text{Insertion Errors} + \text{Deletion Errors}}{\text{Total Number of Slots in the Reference}} \tag{2.7}$$

Where:

- **Substitution Errors**: Instances in which the extracted slot is incorrect or does not match the reference.
- **Insertion Errors**: Instances in which the model extracts an extra slot that should not be present.
- **Deletion Errors**: Instances in which the model fails to extract a slot that is present in the reference.

A lower SER indicates better performance, as the model has made fewer errors in extracting the relevant information. This metric is particularly useful for evaluating tasks like the Bacteria Biotope rel+ner subtask, where the precision and accuracy of extracting specific entities and their relations are critical.

ROUGE-1: measures the overlap of unigrams (individual words) between the model-generated summary and the reference summary. It evaluates how many unigrams from the reference summary are present in the generated summary.

$$\text{ROUGE-1} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{w \in S} \text{Count}_{\text{match}}(w)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{w \in S} \text{Count}(w)} \tag{2.8}$$

ROUGE-2: measures the overlap of bigrams (two consecutive words) between the model-generated summary and the reference summary. It evaluates how many bigrams from the reference summary are present in the generated summary.

$$\text{ROUGE-2} = \frac{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{b \in S} \text{Count}_{\text{match}}(b)}{\sum_{S \in \{\text{Reference Summaries}\}} \sum_{b \in S} \text{Count}(b)} \tag{2.9}$$

ROUGE-L: measures the longest common subsequence (LCS) between the model-generated summary and the reference summary. It evaluates the longest sequence of words that appear in both summaries in the same order, thus capturing sentence-level structure similarity.

$$\text{ROUGE-L} = \frac{LCS(C, R)}{\text{Length}(R)} \tag{2.10}$$

where $LCS(C, R)$ is the length of the longest common subsequence between candidate summary C and reference summary R.

These ROUGE metrics provide a comprehensive evaluation of the quality of generated summaries by considering both individual word matches and sequence-level matches, offering insights into the relevance and coherence of the summaries.

Chapter 3

Medical Databases

In this chapter, we describe the datasets used in this dissertation, including the ClinicalTrials.gov [cli24], BioCreative VI, and BioCreative VII datasets known as CHEMPROT [WCH18] and DrugProt [MML+21]. ClinicalTrials.gov provides a vast repository of electronic health reports from clinical trials, encompassing a wide range of medical research data. These reports are essential to understand various clinical studies' scope, methodology, and results. On the other hand, the BioCreative datasets, CHEMPROT (CPI) and DrugProt include PubMed data specifically curated for named entity recognition and relation extraction tasks. These datasets focus on extracting and identifying relationships between chemical compounds and proteins, as well as drugs and their interactions, respectively. Combining these datasets enables a comprehensive analysis of clinical and biomedical texts, supporting the development of robust models for information extraction and summarization in the medical domain.

3.1 ClinicalTrials.gov

ClinicalTrials.gov is a comprehensive database of privately and publicly funded clinical studies conducted worldwide, providing valuable information on diverse medical research. As of now, it hosts information on approximately 456,000 clinical trials, making it one of the most extensive repositories of clinical trial data globally. This platform is crucial for researchers, healthcare professionals, and patients, offering detailed descriptions of each study, including the objectives, methodologies, participant eligibility criteria, locations, and outcomes. By facilitating access to this wealth of information, ClinicalTrials.gov plays a pivotal role in advancing medical knowledge, promoting transparency in research, and helping patients find studies they may be eligible to participate in

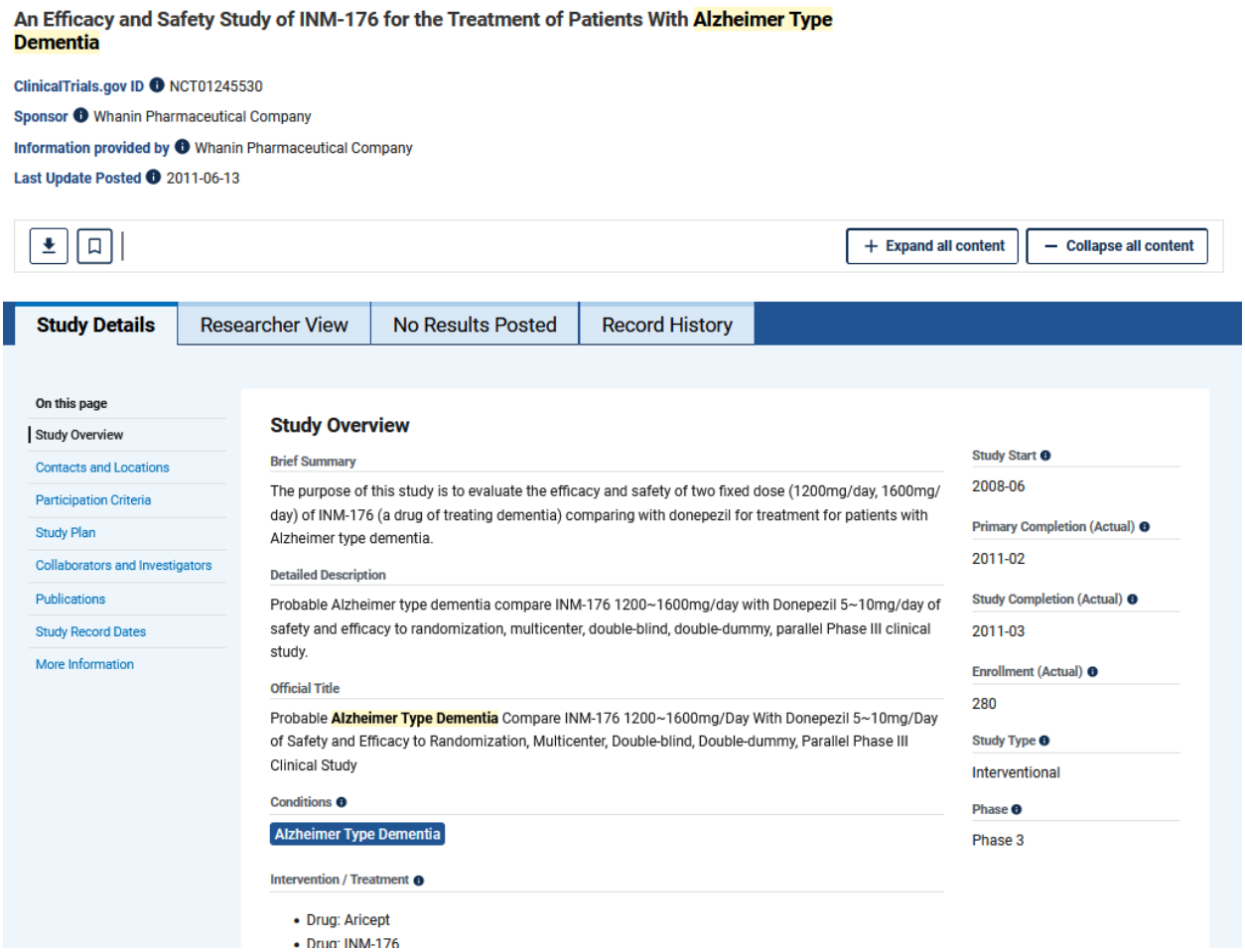
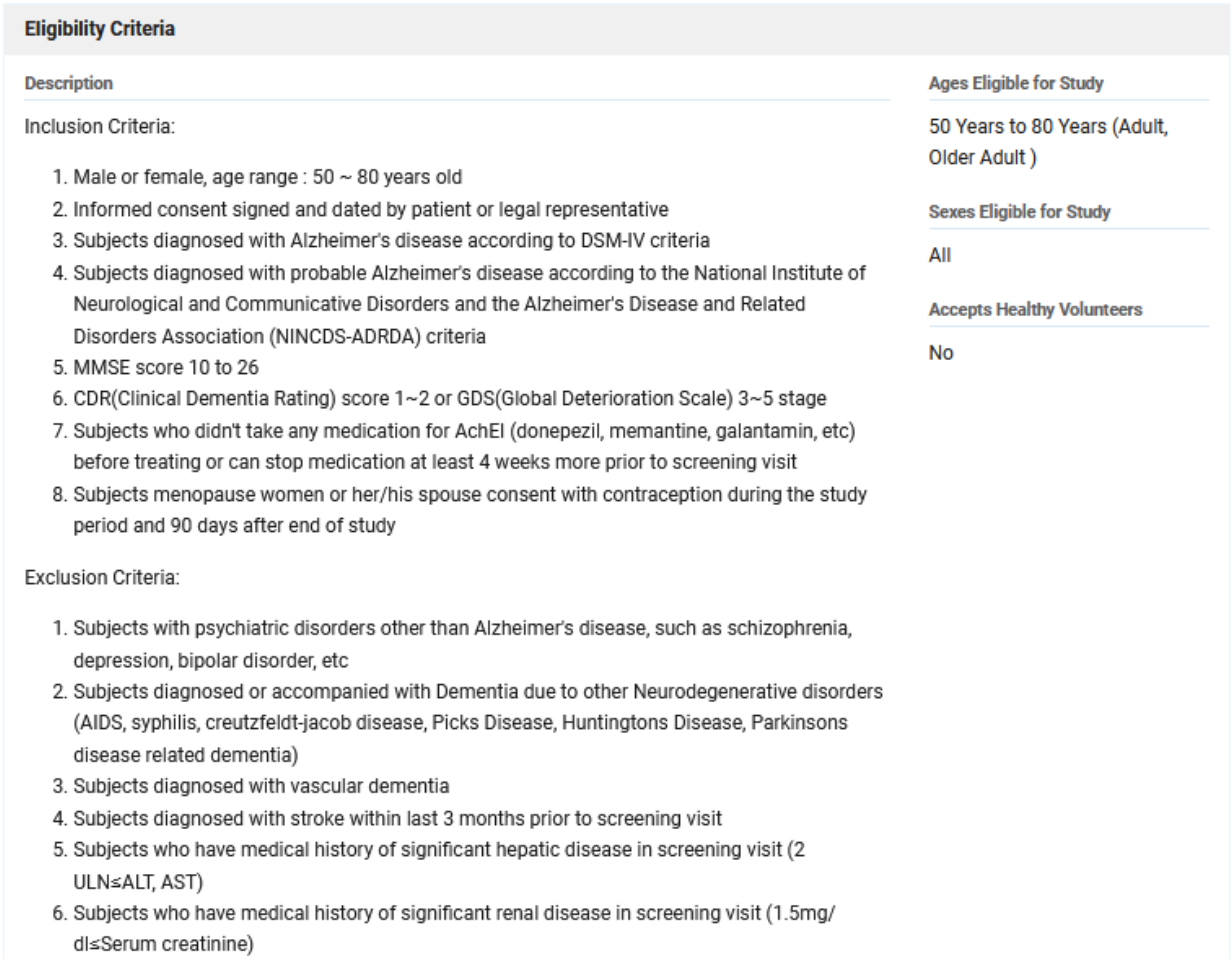


Figure 3.1: Clinicaltrials.gov Tabular Data Sample

[cli24] [ZTW+11]. Figures 3.1 and 3.2 indicates one data sample from this dataset. It has both tabular and large textual (eligibility criteria) data features. We describe each data field as follows:

- **Study Title**: A brief overview of the clinical study.
  - **Definition**: The title of the clinical study, often including information about the condition being studied and the intervention being tested.
  - **Description**: Provides a brief overview of the study's focus, helping users quickly understand the primary objective of the study.
- **NCT Number**: Unique identifier for the study.
  - **Definition**: A unique identifier assigned to each clinical study registered on ClinicalTrials.gov. The "NCT" stands for National Clinical Trial.



Eligibility Criteria

Description

Inclusion Criteria:

1. Male or female, age range : 50 ~ 80 years old
2. Informed consent signed and dated by patient or legal representative
3. Subjects diagnosed with Alzheimer's disease according to DSM-IV criteria
4. Subjects diagnosed with probable Alzheimer's disease according to the National Institute of Neurological and Communicative Disorders and the Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria
5. MMSE score 10 to 26
6. CDR(Clinical Dementia Rating) score 1~2 or GDS(Global Deterioration Scale) 3~5 stage
7. Subjects who didn't take any medication for AchEI (donepezil, memantine, galantamin, etc) before treating or can stop medication at least 4 weeks more prior to screening visit
8. Subjects menopause women or her/his spouse consent with contraception during the study period and 90 days after end of study

Exclusion Criteria:

1. Subjects with psychiatric disorders other than Alzheimer's disease, such as schizophrenia, depression, bipolar disorder, etc
2. Subjects diagnosed or accompanied with Dementia due to other Neurodegenerative disorders (AIDS, syphilis, creutzfeldt-jacob disease, Picks Disease, Huntingtons Disease, Parkinsons disease related dementia)
3. Subjects diagnosed with vascular dementia
4. Subjects diagnosed with stroke within last 3 months prior to screening visit
5. Subjects who have medical history of significant hepatic disease in screening visit (2 ULN≤ALT, AST)
6. Subjects who have medical history of significant renal disease in screening visit (1.5mg/dl≤Serum creatinine)

Ages Eligible for Study

50 Years to 80 Years (Adult, Older Adult)

Sexes Eligible for Study

All

Accepts Healthy Volunteers

No

Figure 3.2: ClinicalTrials.gov Textual Data Fields

  - **Description**: This number helps in uniquely identifying and referencing the study, ensuring precise tracking and retrieval of study information.

- **Study Status**: Current phase of the study.
  - **Definition**: Indicates the current phase of the study, such as "Recruiting," "Completed," "Terminated," etc.
  - **Description**: This status helps users understand the current state of the study, including whether it is actively enrolling participants or has concluded.

- **Conditions**: Medical conditions or diseases being studied.
  - **Definition**: Lists the medical conditions or diseases being studied in the trial.
  - **Description**: Helps users identify studies related to specific health issues, providing

context about the scope of the research.

- **Interventions**: Treatments or therapies being tested.
  - **Definition**: Details about the treatments, drugs, procedures, or therapies being tested.
  - **Description**: Includes the type and nature of the intervention, outlining what is being investigated to assess its impact on the condition.
- **Sponsor/Collaborators**: Organizations responsible for the study.
  - **Definition**: The organization or individual responsible for the study and any additional organizations collaborating on the study.
  - **Description**: Includes pharmaceutical companies, research institutions, or government bodies that are overseeing or funding the study.
- **Study Design**: Structure and methodology of the study.
  - **Definition**: Provides information about how the study is structured, including the type of study (e.g., interventional, observational), the number of participants, the allocation, intervention model, and masking details.
  - **Description**: Outlines the methodological approach of the study, helping users understand how the research is conducted and its scientific rigor.
- **Eligibility Criteria**: Requirements for study participation.
  - **Definition**: Describes the characteristics that participants must have to be included or excluded from the study.
  - **Description**: Helps determine who can participate in the study, ensuring that the study population is appropriate for the research question.
- **Locations**: Geographic locations of the study sites.
  - **Definition**: Lists the geographic locations where the study is being conducted, often including the name of the institution and contact information.
  - **Description**: This helps potential participants find study sites near them and provides researchers with site details for collaboration.
- **Outcome Measures**: Goals of the study.

  - **Definition**: Defines what the study aims to measure to determine the effect of the intervention, including primary and secondary outcome measures.
  - **Description**: These outcomes help in evaluating the effectiveness and safety of the intervention, guiding the study's objectives and endpoints.

- **Study Dates**: Timeline of the study.
  - **Definition**: Provides the start date, estimated primary completion date, and estimated study completion date.
  - **Description**: These dates give an idea of the study timeline and duration, helping users understand the study's schedule and progress.

- **Contacts and Locations**: Contact information for the study.
  - **Definition**: Information about how to contact the study research staff and where the study is being conducted.
  - **Description**: Useful for potential participants and researchers looking to collaborate or get more information about the study.

- **Phase**: Phase of the clinical trial.
  - **Definition**: Indicates the phase of the clinical trial (e.g., Phase 1, Phase 2, Phase 3, Phase 4).
  - **Description**: Each phase has specific goals, such as assessing safety, efficacy, and side effects, providing context about the study's stage in the clinical trial process.

- **Researcher View**: Additional information for researchers.
  - **Definition**: This tab provides additional information useful for researchers, including detailed study protocol, statistical analysis plans, and other scientific data that may not be included in the general public view.
  - **Description**: Offers in-depth information for scientific scrutiny, aiding researchers in understanding the study's design and methodology.

- **Results Posted/Not Posted**: Status of study results.

- **Definition**: Indicates whether the results of the study have been posted on ClinicalTrials.gov.
- **Description**: If results are posted, this section includes data on the study's outcome measures, adverse events, and other findings. If not posted, it may indicate that results are pending or the study has not yet been completed.

- **Record History**: History of changes to the study record.
  - **Definition**: This feature tracks the history of changes made to the study record.
  - **Description**: Includes updates to the study status, amendments to the study protocol, changes in sponsor information, and other modifications since the study was first registered. This helps provide transparency and allows users to see how the study has evolved over time.

In Chapters 4 and 5 of this dissertation, we introduce our system designed to investigate data on ClinicalTrials.gov for visualization, drug Mechanism of Action (MoA) classification, and clinical trial summarization. Our approach leverages advanced data processing techniques to enhance the understanding and accessibility of clinical trial information.

3.2 BioCreative VI and VII Datasets

For over a decade, BioCreative—a critical evaluation of text mining methodologies in molecular biology—has been dedicated to extracting annotated biomedical triples, such as *(Chemical, Interaction-type, Gene)*, from PubMed articles [KRL$^+$15, KRA$^+$17]. This initiative has significantly advanced the field of biomedical text mining by providing robust datasets and challenges that drive innovation and improve the accuracy of information extraction methods.

In our research, we assess the Bio-RIFRE system on two prominent public datasets: CHEMPROT (CPI) [KRA$^+$17] and DrugProt [MML$^+$21]. These datasets are derived from the BioCreative VI Track 5 and VII Track 1 challenges, respectively, and are designed to explore the interactions between chemicals and genes/proteins. Both datasets are meticulously curated and manually annotated, facilitating the extraction of chemical-protein entity relations from PubMed abstracts, which is crucial for understanding complex biological processes.

The CHEMPROT dataset serves as a gold standard for chemical-protein interaction (CPI) extraction and comprises 2,420 PubMed abstracts. These abstracts are classified into five relation

Table 3.1: BioCreative VI and VII datasets statistics [PRKL18]

Dataset	**Set**	**# Abstracts**	**# Relations**	**# Entities**	
				Gene	**Chemical**
CHEMPROT	Training	1020	4157	12752	13017
	Development	800	2416	7568	8004
	Test	600	3458	10019	10810
DrugProt	Training	3500	17274	43255	46274
	Development	750	3761	9005	9853
	Test	750	3491	9515	9434

categories: CPR 3 (upregulator), CPR 4 (downregulator), CPR 5 (agonist), CPR 6 (antagonist), and CPR 9 (substrate). The dataset includes two main groups of entities: genes and chemicals, totaling 30,339 gene entities and 31,831 chemical entities. This detailed classification supports the development and evaluation of CPI extraction algorithms by providing clear and well-defined interaction types.

The DrugProt dataset, published in 2021, extends the scope of relation extraction with a broader set of interactions. It consists of 5,000 PubMed abstracts and encompasses thirteen distinct relation classes: INDIRECT DOWNREGULATOR, INDIRECT UPREGULATOR, DIRECT REGULATOR, ACTIVATOR, INHIBITOR, AGONIST, AGONIST-ACTIVATOR, AGONIST-INHIBITOR, ANTAGONIST, PRODUCT-OF, SUBSTRATE, SUBSTRATE PRODUCT-OF, and PART-OF. This dataset includes 61,775 gene entities and 65,561 chemical entities. The wide range of interaction types in DrugProt allows for a more comprehensive exploration of biochemical interactions, enhancing the ability to train and evaluate sophisticated models for biomedical relation extraction. Table 3.1 summarizes the statistics of both datasets, highlighting their extensive and varied nature.

Chapter 4

An Open-Source Platform for Clinical Trials Analysis

ClinicalTrials.gov is an extensive online repository managed by the National Library of Medicine. It provides information about clinical studies on various interventions, such as drugs, devices, and behavioral treatments. It is a crucial resource for scientists, medical researchers, pharmaceutical companies, and other institutions. However, managing the progression of clinical trials, extracting research-specific information, and utilizing AI assistants for decision-making in various domains, such as drug mechanism of action recognition, present substantial challenges when using this database. A significant issue is the dataset website's lack of visualization tools, which hinders practical data interpretation and analysis. Without visual aids, identifying patterns and trends within the data becomes exceedingly difficult. This deficiency complicates monitoring clinical trial advancements and integrating AI-driven insights into research workflows. Consequently, researchers and healthcare professionals need help fully leveraging this valuable resource's potential, impeding progress in critical areas of medical research [DCKH$^+$19] [LHS23].

To address these issues, we developed the open-source platform Tri-AL to enhance data visualization and analysis and facilitate comprehensive data extraction from textual fields. This platform supports medical experts in their research by maintaining historical data and offering detailed diversity statistics, making it a valuable tool for clinical trial analysis and improving the efficiency of drug development processes.

4.1 Introduction

The ClinicalTrials database is utilized by scientists, medical researchers, pharmaceutical companies, and other public and private institutions. One critical use case of this database is to create *drug development pipeline observatories* for various diseases. These reports summarize the current state of drug development, comment on drug effectiveness and side effects, and inform national drug policies. Example observatories include those for Alzheimer's Disease (AD) [CLN+22] [MCT22], COVID-19 [PPH+20], Amyotrophic Lateral Sclerosis (ALS) [New22], and Parkinson's Disease [Par22]. There are also other motivations for investigating this database. For example, one significant reason is to figure out the key indicators of the clinical trials problem in participant recruitment; it is estimated that 80% of trials cannot meet their patient recruitment timeline or fail to recruit a minimum number of participants required for completion [BMAZ+20]. The Text Retrieval Conference (TREC) runs an annual challenge that uses the ClinicalTrials.gov data to match participants to trials [Con22] to find a better way for better participant trial matching. Additionally, there are other motivations for investigating this database. For example, identifying key indicators of the clinical trials problem in participant recruitment is crucial, as it is estimated that 80% of trials cannot meet their patient recruitment timeline or fail to recruit the minimum number of participants required for completion [BMAZ+20]. The high failure rate of trials due to recruitment issues leads to substantial financial and resource losses. To address this challenge, the Text Retrieval Conference (TREC) runs an annual challenge using ClinicalTrials.gov data to match participants to trials, aiming to improve participant trial matching and enhance the efficiency and success rate of clinical trials [Con22].

Another critical concern is the recruitment of minorities. It is well known that sex, race, and ethnicity significantly influence disease incidence, drug responses, and treatment outcomes [GLCI+18, NMPO19]. However, as of 2014, less than 2 percent of cancer institute clinical trials were allocated to minority populations [CJLD+14]. Additionally, even though African Americans accounted for 21 percent of COVID-19 deaths, they only represented 3 percent of participants in vaccine clinical trials [WFHST20]. Addressing these disparities is crucial for ensuring that clinical trial results are representative and that all populations benefit equally from medical advancements.

4.2 Problem Statement

The ClinicalTrials.gov website, as the database's interface, has significant limitations in user customization, data accessibility, and advanced analytics. tracking clinical trials over time requires repeated manual data downloads, as most existing solutions lack robust mechanisms for historical data tracking. The ClinicalTrials.gov interface has several notable drawbacks when it comes to accessing and utilizing the data as follows:

Limited Filtering Options: The platform offers only basic filtering capabilities, which can make it difficult for users to narrow down their searches to find specific trials or data points that are most relevant to their needs.

Lack of Visualization Tools: ClinicalTrials.gov does not provide built-in visualization tools for data analysis. Users cannot easily create charts, graphs, or other visual aids directly from the database, which hinders the ability to quickly interpret and analyze the data.

Access to Latest Data Only: The platform only stores the most recent version of data for each trial. This limitation means that historical data is not readily accessible, making it challenging to perform longitudinal studies or track changes in trials over time.

Interface Usability: The user interface of ClinicalTrials.gov can be cumbersome and not particularly user-friendly. This can pose a barrier to efficiently navigating and extracting data, especially for users who are not highly familiar with the platform.

Inadequate Support for Advanced Analytics: The interface lacks support for advanced analytics and data manipulation, limiting the ability to perform complex data analysis directly within the platform.

Limited Reporting Features: The platform does not offer robust reporting features that allow users to easily generate comprehensive reports based on their specific criteria or research needs.

Insufficient Diversity Analysis Tools: There are limited tools available for analyzing and reporting on the diversity of trial participants, such as sex/gender and race/ethnicity statistics. This gap makes it harder to assess and address disparities in clinical trial representation.

Textual Data Extraction: Extracting specific information from textual fields within the database can be challenging, as the interface is not optimized for detailed text retrieval or analysis.

Addressing these drawbacks requires developing additional tools and features that enhance the usability, analytical capabilities, and data accessibility of ClinicalTrials.gov. We described related

work in this area in section 2.1. Previous research in this area has highlighted the importance of analyzing participant diversity, yet this often involves manual data analysis, which is time-consuming and prone to errors. Additionally, search and information retrieval from clinical databases rely heavily on advanced techniques like named entity recognition and transformer-based models, but these methods are not seamlessly integrated with ClinicalTrials.gov for real-time data analysis and reporting.

To address these drawbacks and challenges, we introduce *Tri-AL* (VisuAL ClinicAL TriALs), an open-source data platform for visualization and analysis of ClinicalTrials.gov. Our primary purpose is to provide automated tools that assist medical experts in developing reports and analyzing data. This includes creating charts and tables and extracting information from textual fields. Another key goal is maintaining historical data to examine trial changes over time concerning their phase, status, and study outcomes. Historical analysis is crucial for evaluating trial feasibility and identifying study design issues and intervention progression, as demonstrated in the recent analysis of SARS-CoV-2 trials [HKC+22] [CLRZ18] [CLN+22]. For example, the estimated completion date of a trial may change over time, making access to historical data vital for analyzing trials that take longer than initially expected. *Tri-AL* also maintains sex/gender and race/ethnicity statistics for US trials, facilitating diversity reports and trend analysis. Furthermore, *Tri-AL* is programmable, allowing users to extract disease-specific information from textual fields such as the description of the study design. Machine learning models can also be plugged into *Tri-AL* to generate additional information, such as predicted labels. For example, in our extension of *Tri-AL* to Alzheimer's Disease, we implemented a deep learning model to predict the Mechanism of Action (MoA) of drugs based on their descriptions extracted from trial data [NKC+22]. MoA reflects the physical and chemical processes through which drugs interact with the human body, but it is not yet clear for diseases like AD. Hence, this ML model is helpful for the researcher in deciding on the tested drugs' MoA. Notably, *Tri-AL* addresses several clinical trial data analysis technical challenges that previous research does not address. This includes historical data tracking and extendable modules for machine learning models and clinical information extraction. Moreover, *Tri-AL*'s open-source architecture is designed for flexibility and customization, enabling users to adapt the platform to their needs.

To summarize, the main contributions of *Tri-AL* are as follows:

- *Tri-AL* is an open-source interactive platform for ClinicalTrials.gov, featuring built-in charts and tables for comprehensive data visualization.

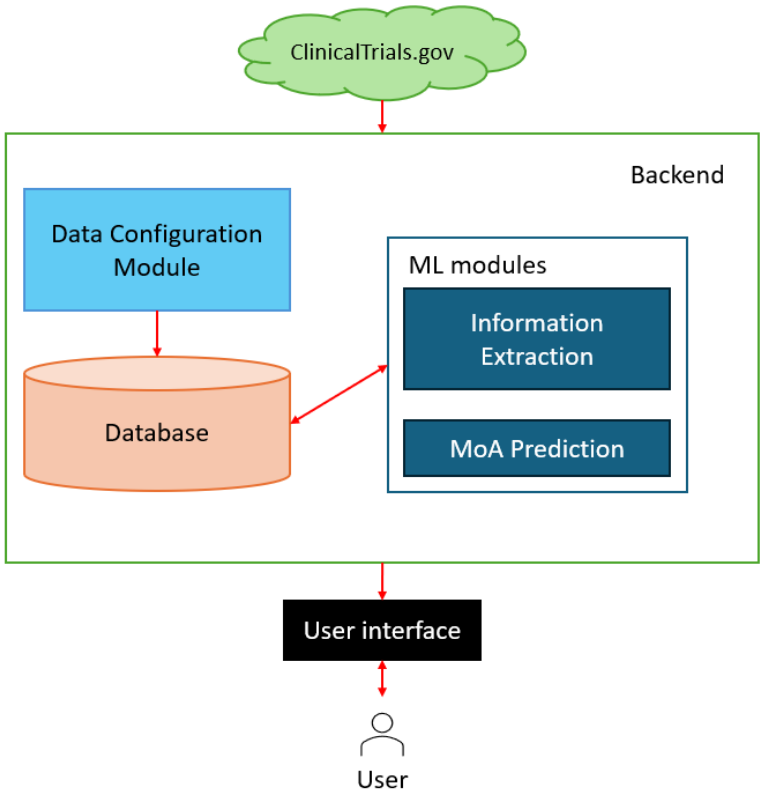


Figure 4.1: An Overview of the *Tri-AL* Architecture

- To facilitate historical analysis, *Tri-AL* captures the complete history of every field, including trial status and study outcomes.
- *Tri-AL* offers detailed reports and statistics on the sex/gender and race/ethnicity demographics of US trials.
- *Tri-AL* is designed to be customized for a specific disease. It can be programmed to focus on disease-specific information and extract related information from its textual fields. This functionality is demonstrated through a case study on Alzheimer's Disease.

4.3 System Overview

This section outlines the design of *Tri-AL*. The system architecture is depicted in Figure 4.1 and is explained in detail below.

4.3.1 Data Configuration

We configure the data in two phases.

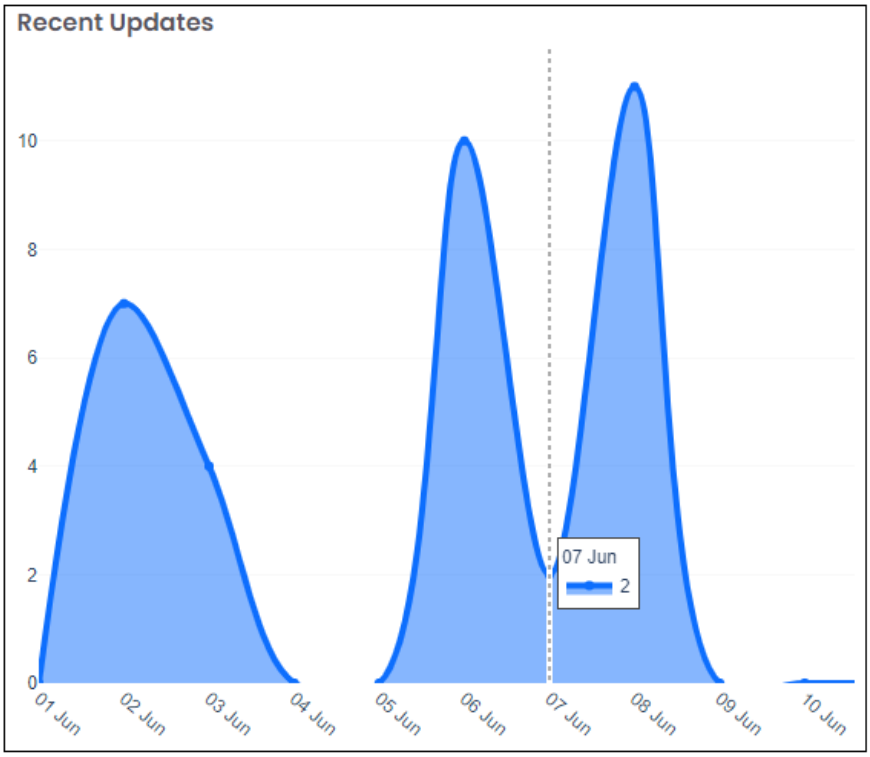


Figure 4.2: Number of Updated AD Trials

Phase One involves the initial data download. ClinicalTrials.gov periodically uploads a comprehensive data backup in a single zip file, which we use to initialize the *Tri-AL* database. This backup includes XML files for each trial, and to parse these XML files, we use the libxslt C libraries and the lxml Python interface. Notably, this zip file is 2 GB, and when unzipped, the data expands to 20 GB, highlighting the substantial initial setup required.

Phase Two manages the continuous import of new data into *Tri-AL*. While one method is to download the entire backup from ClinicalTrials.gov daily, this approach is inefficient when updating. Therefore, we utilize the ClinicalTrials.gov search tools and API for updates. The search tools allow us to retrieve updates in formats such as comma-separated values (CSV) within a specific period. However, these updates lack some fields present in the full data. To acquire these additional fields, we use the API, querying with the IDs of the updated trials. This combined method enables us to download complete copies of the updated trials, ensuring the *Tri-AL* database contains every trial version along with a timestamp for historical analysis. For instance, figure 4.2 indicates changes in the number of Alzheimer's trials over time.

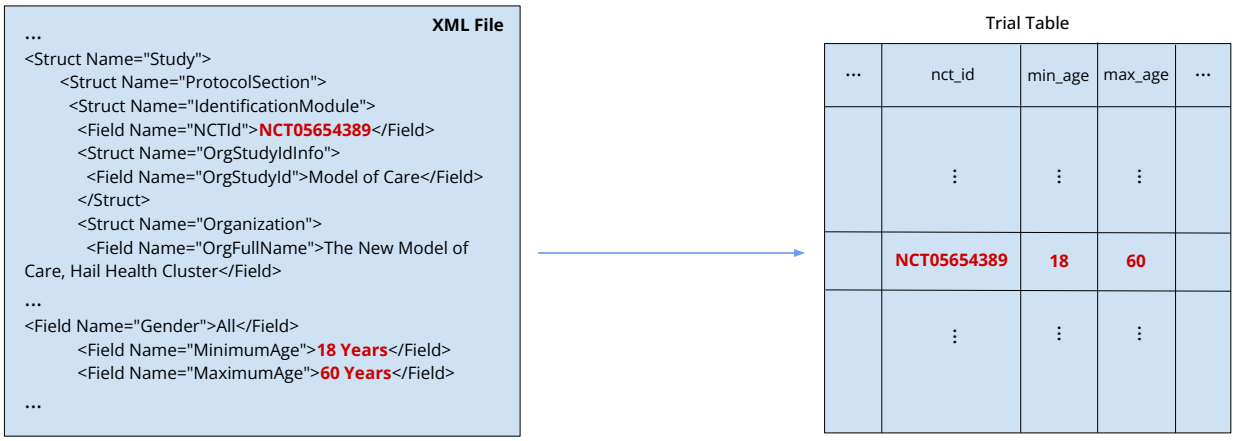


Figure 4.3: Inserting XML Data into SQL Tables

4.3.2 Database

The database underpinning the ClinicalTrials.gov website comprises 46 tables [Int24]. To manage complexity and streamline data handling and analysis, we consulted with medical researchers to identify the most crucial relations. For example, the original database includes tables, such as "Milestone" and "DropWithdrawal," which provide supplementary details about trial milestone periods and reasons for participant withdrawal in trials marked with a withdrawal status. Based on feedback from medical researchers during the requirements-gathering phase, we decided to exclude tables that contain non-essential data for the drug progression system's goals.

Our simplified schema comprises six key data tables: Agent (interventions), Condition, Biomarker, Sponsor, Country, and Trial. Additionally, the schema includes four system tables: Subscriber, Newsletter, UpdatesLog, and HistoricalTrial. The Subscriber and Newsletter tables track users who subscribe to receive our newsletter with information on newly added trials. For more details on the database schema, refer to the "`models.py`" file in the code [ALc24]. We use the Django framework and SQLite 3 as the database backend due to its lightweight architecture, simplicity, and ease of setup. Figure 4.3 indicates how the XML file is read from the data configuration module and is stored in the trial table.

4.3.3 ML Module: Information Extraction

Tri-AL can extract disease-specific information from textual data fields by utilizing a set of pre-defined named entities as parameters. The system employs a deep learning-based model, detailed in Chapter 6, and a rule-based model explained in this chapter. For example, Table 4.1 delineates

specific data features such as Biomarkers, Mini-Mental State Examination scores, CSF or PET data, and Subject Characteristics that need to be extracted from specified data fields. The information extraction module within *Tri-AL* allows users to define new columns in the *Tri-AL* database and implement functions and models to populate these new columns with relevant data, thereby enhancing the granularity and specificity of the extracted information for research and analysis purposes. Algorithm 1 is designed to extract specific information related to the MiniMental State Examination (MMSE) and the presence of Cerebrospinal Fluid (CSF) mentions from clinical trial criteria. The `Functions` class contains two key methods: `extract_mmse` and `is_csf`. To facilitate pattern matching, the `extract_mmse` method preprocesses the trial criteria by replacing specific Unicode characters and phrases with standardized symbols (line 3). It tokenizes the processed criteria text into sentences and filters these sentences to identify those containing mentions of 'mmse' or 'mini-mental state examination'(line 5). Using regular expressions, it extracts specific patterns from these target sentences, such as numerical ranges and comparisons, and joins these patterns into a single string, which is then returned (lines 7-12). The `is_csf` method checks if the term 'csf' is mentioned in the trial criteria text, returning a boolean value indicating its presence (lines 14-15). Additionally, a `Database` class is defined to specify the structure of database columns for storing extracted information, including a character field for `mmse` and a boolean field for `csf`. This structured approach enables efficient extraction and storage of disease-specific information from textual data, facilitating subsequent analysis and querying of clinical trial data related to AD.

Table 4.1: List of the parameters for AD pipeline

Data Feature	**Keywords Example**	**Field of Data in ClinicalTrials.gov**
Biomarkers	Amyloid PET, CSF amyloid, CSF Neurofilament light, CSF GPAP, CSF Neurogranin, CSF p-tau 181, CSF p-tau217, CSF p-tau 231, CSF total tau, FDG-PET, vMRI, Plasma Amyloid, Plasma Neurofilament light, Plasma p-tau 181, Plasma p-tau217, Plasma p-tau 231, Plasma total tau, Tau PET	Eligibility Criteria, Primary Outcome, Secondary Outcome, Other Outcome
Mini-mental State Examination	MMSE or mini-mental state examination	Eligibility Criteria
CSF or PET	CSF or PET	Eligibility Criteria
Subject Characteristics	Autosomal dominant AD mutation carriers, Alzheimer Dementia, preclinical-MCI due to AD, MCI-mild AD dementia, moderate AD dementia-severe AD dementia, MCI due to AD, mild Ad dementia, mild-moderate-severe AD dementia, severe AD dementia, MCI to Moderate Dementia, MCI, Healthy Volunteers, Severe AD, Mild-Moderate AD Dementia, Prodromal/Prodromal-Mild, Preclinical AD	Eligibility Criteria

Algorithm 1: Extract MMSE and CSF Information from Trial Criteria

Data: Trial dictionary

Result: Extracted MMSE information as a string, CSF presence as a boolean

```
Function extract_mmse(trial: dict) → str:
    criteria ← trial['criteria'];
    criteria ← criteria.replace('\u2212', '-').replace('greater than',
     '>').replace('less than', '<').replace('\u2264', '<=').replace('\u2265',
     '>=');
    sentences ← sent_tokenize(criteria);
    target_sentences ← { sent | 'mmse' in sent.lower() or 'mini-mental state
     examination' in sent.lower() };
    results ← [ ];
    foreach sentence in target_sentences do
        results.extend(re.findall(r'\d+\s?to\s?\d+', sentence));
        results.extend(re.findall(r'[¡¿]\s?\d+', sentence));
        results.extend(re.findall(r'¿=s?\d+', sentence));
        results.extend(re.findall(r'\d+ and \d+', sentence));
    end
    return ' | '.join(results);
Function is_csf(trial: dict) → bool:
    return 'csf' in trial['criteria'].lower();
Class Database;
columns ← { 'mmse':  models.CharField(max_length=50, null=True),
'csf':  models.BooleanField(default=False) };
```

4.3.4 ML Module: MoA Prediction

The urgent need for a system to predict the Mechanism of Action (MoA) of Alzheimer's Disease (AD) drugs stems from the chaotic and fragmented nature of existing medical data repositories. With vast amounts of unstructured, unlabeled, and sparse data spread across multiple sources, researchers face significant challenges in efficiently organizing and accessing the information necessary for effective analysis and development. A robust prediction system would streamline the process of labeling and categorizing crucial data, such as MoA, thereby enhancing the accuracy and speed

of research efforts. This system's importance is further underscored in its ability to consolidate disparate information into a cohesive framework, facilitating more targeted and efficient drug development. The design and implementation of this MoA prediction module are detailed in Chapter 5 of this document.

4.3.5 System Dashboard

The *Tri-AL* user interface is meticulously crafted to meet the diverse needs of researchers, clinicians, and biomedical scientists specializing in computer science. It includes advanced filtering options, map-based exploration for geographical insights, and time-series visualizations for monitoring temporal trends. The system leverages the *plotly* Python library to create dynamic plots, which are then exported as HTML code and embedded as interactive visualizations within the dashboard. The Tri-AL interface is organized into the following sections:

- **Home Page**: Figure 4.4 showcases the *Tri-AL* home page, which provides summary statistics such as the total number of trials, the number of conditions or diseases studied, the specific interventions and drugs tested, the countries with at least one trial, and a list of countries with the highest number of trials. The home page also offers a breakdown of trial statuses over a given period.

- **Search Page**: This section provides a comprehensive list of all trials in the database and enables users to search for specific trials using trial IDs. Figure 4.5 illustrates this section in the context of our extension of *Tri-AL* for analyzing Alzheimer's disease trials. Each trial is displayed with its status, phase, and the date of the last update. The search page also includes filtering options by status, trial phase, and date of the last update.

- **Detail Page**: *Tri-AL* includes an information extraction module that allows users to extract additional data from textual fields, which are then stored in new columns in the database. Expert users are recommended to verify the automatically collected data to ensure accuracy. The data extraction page provides a graphical interface where users can edit the data for any given trial. Figure 4.6 demonstrates this with extracted data for Alzheimer's disease trials, including biomarkers and drug mechanisms of action (MoAs).

- **Demographics Page**: This page displays statistics on the race/ethnicity and sex/gender of trial participants. Users can examine the proportion of trials that included participants iden-

tified as White, Black, American Indian, Asian, and Latin. Furthermore, this page features charts that illustrate the trends in race/ethnicity reporting over time.

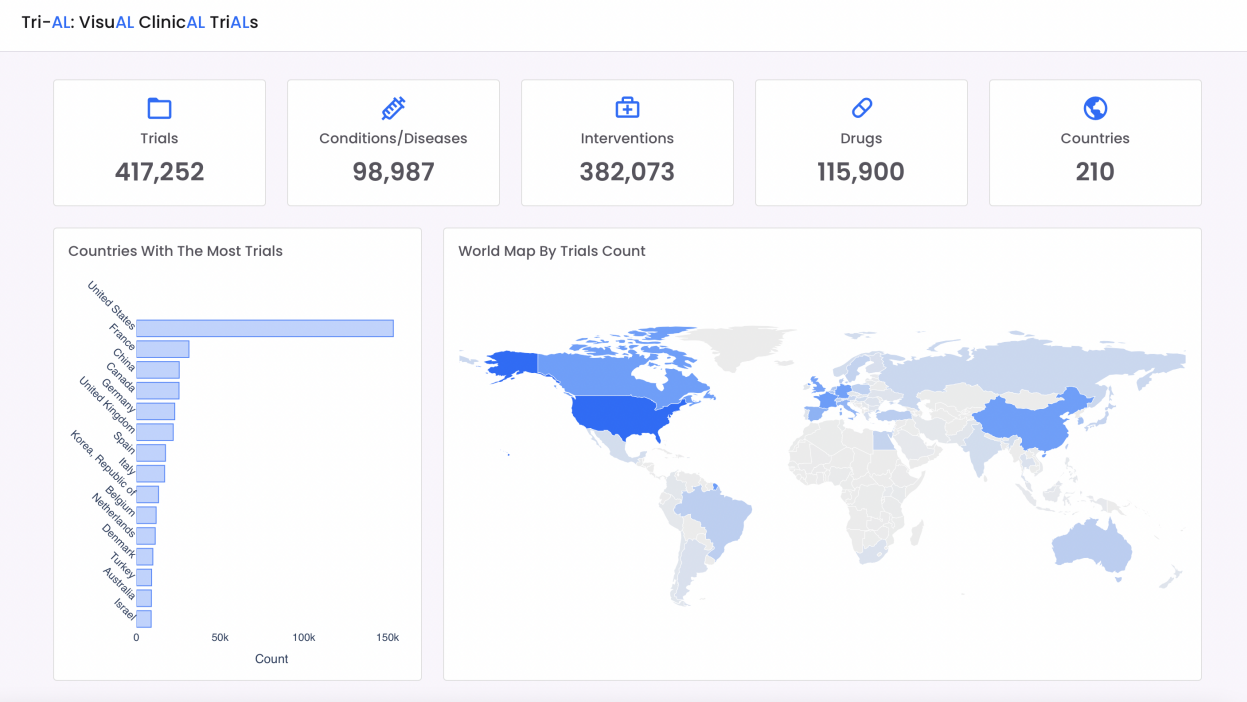


Figure 4.4: The *Tri-AL* Dashboard

4.3.6 Diversity Reporting

Clinical trials with diverse participant pools are essential for identifying effective treatments for minority populations. Consequently, numerous studies and regulations emphasize the importance of evaluating the inclusion and diversity of trial participants. This section demonstrates *Tri-AL*'s capabilities for analyzing race/ethnicity and gender/sex data, comparing the results with previous studies to highlight changes in the reporting of minority data on "ClinicalTrials.gov" over the past two years.

Race and ethnicity definitions can vary across different organizations and regulatory bodies. According to the NIH/OMB policy, race and ethnicity are categorized into five groups: American Indian or Alaska Native, Asian (including Native Hawaiian or other Pacific Islander), Black or African American, White, and Hispanic or Latino. On "ClinicalTrials.gov," the Baseline Measurements field contains statistics about the race/ethnicity and sex/gender of participants. However, this field is free-text, leading to a lack of standardization with 920 distinct values for race and ethnic-

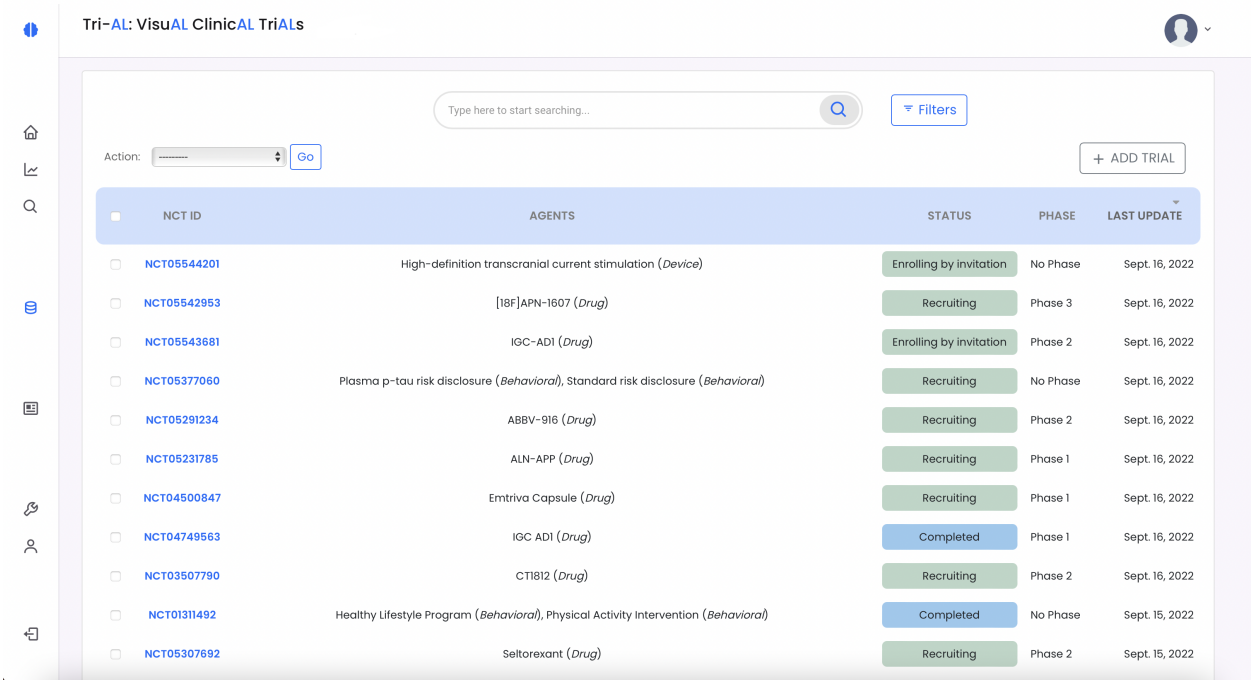


Figure 4.5: A Searchable List of AD Trials

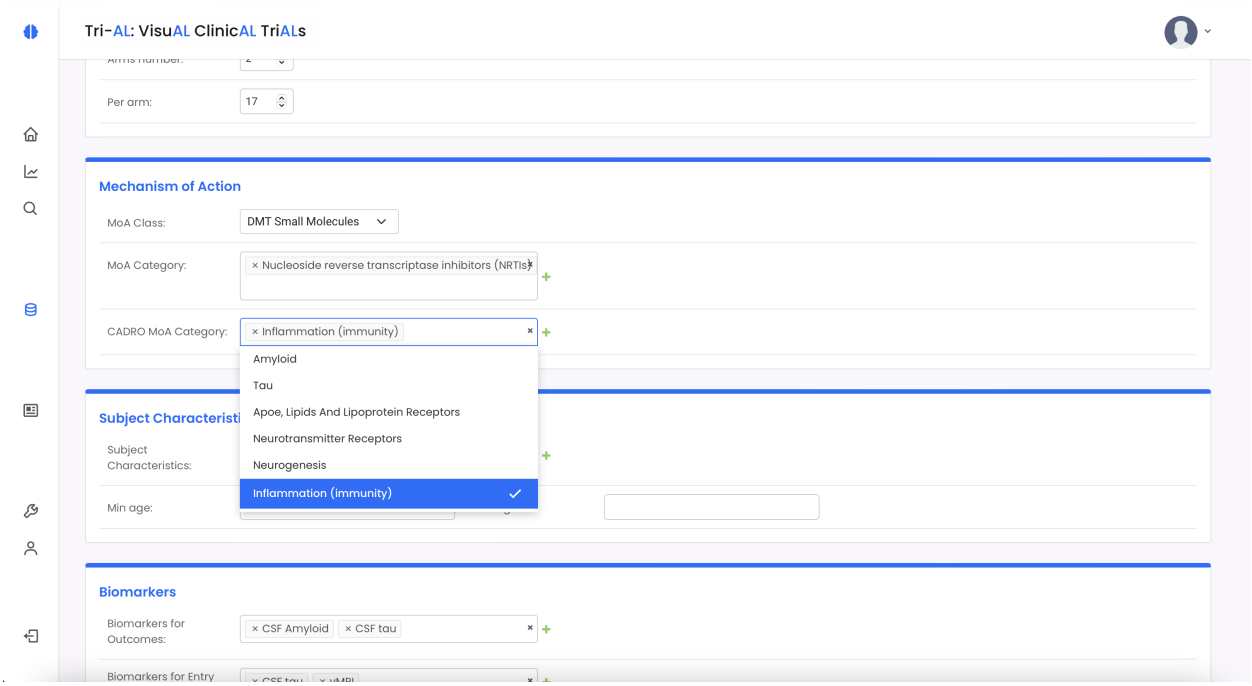


Figure 4.6: Data Extracted from AD Trial Descriptions and Eligibility Criteria

ity. For instance, variations such as "American Indian or Alaska native" versus "American Indian / Alaska Native" occur due to differences in capitalization and separators. *Tri-AL* includes data cleaning processes to standardize these entries according to NIH definitions, ensuring consistency

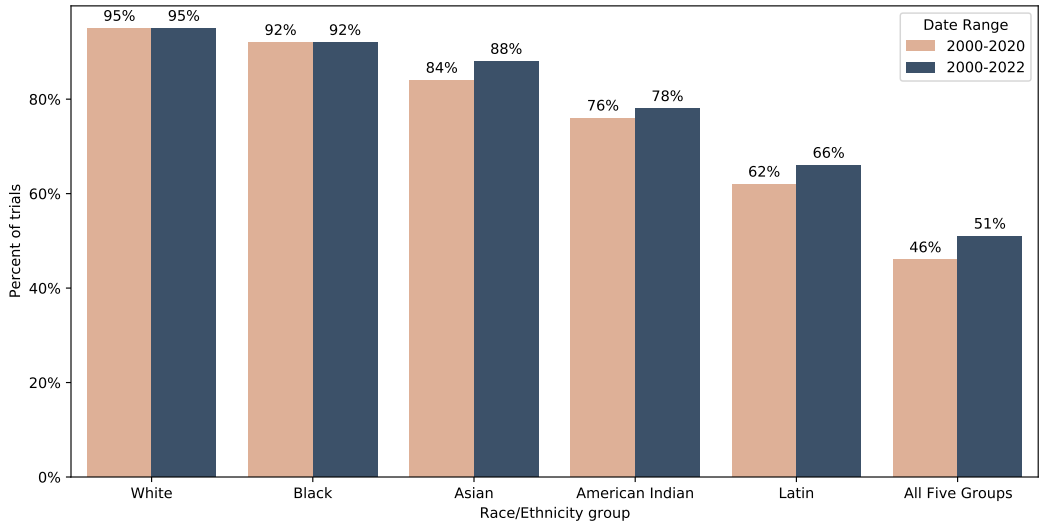


Figure 4.7: Trial Race/Ethnicity Percentages in Years 2000-2020 and the Tri-AL Output for Years 2000-2022

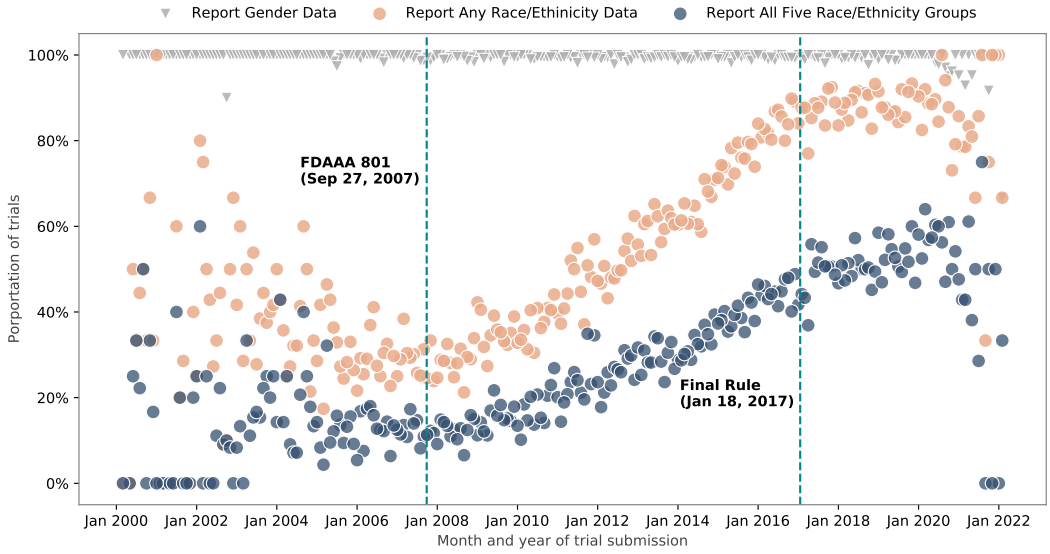


Figure 4.8: Fraction of Trials Reporting Race/Ethnicity and Sex/Gender per Month from 2000-2022

and accuracy in the analysis.

Previous work analyzed the trials reporting ethnicity and race data up to 2020 [TSW+22]. To provide a comparative analysis, *Tri-AL* was employed to generate similar statistics for the period extending to 2022. This analysis focused on interventional trials conducted in the United States,

totaling 30,405 trials. Of these, 16,532 trials (54%) reported at least one of the five race and ethnicity groups, while 13,875 trials (46%) did not provide such information.

Figure 4.7 shows the percentage of trials that reported results for participants from each of the five race-ethnicity groups. The bars on the left depict the results from previous work up to 2020, whereas the bars on the right extend up to 2022, as computed by *Tri-AL*. It is important to note that many trials submitted to "ClinicalTrials.gov" from 2000 to 2022 were updated between 2020 and 2022. Therefore, instead of only considering the period from 2020 to 2022, a comparison is made between the 2000-2020 totals from [TSW$^+$22] and the 2000-2022 totals computed by *Tri-AL*. The data reveal that most clinical trials report results for White, Black, and Asian participants. Additionally, diversity has recently improved, with Asian, American Indian, and Latin race and ethnicity groups being represented in a more significant fraction of trials. However, as shown in the two right-most bars, only about half of the trials include all five groups.

In total, 5.5 million participants were in all trials with reported race/ethnicity data. Of these participants, 58% were White, 22.5% Black, 11% Latin, 5.8% Asian, and 1% American Indian.

Figure 4.8 displays a time series chart showing the proportion of trials reporting participant gender data (gray), any race/ethnicity data (orange), and data for all five race/ethnicity groups (blue). This figure extends similar work from previous research [TSW$^+$22] up to 2020, with *Tri-AL* extending the series to include data from 2020-2022. The top line of gray points reveals that 99% of trials submitted results between 2000 and 2022 included participant sex/gender distribution. However, race and ethnicity reporting was not mandated on ClinicalTrials.gov until the implementation of the Food and Drug Administration Amendments Act (FDAAA 801) in September 2007 and its final ruling in January 2017. The figure shows a significant increase in reporting following these regulations.

For a final comparison with [TSW$^+$22], Table 4.2 categorizes the number of trials by funding type, primary purpose, phase, size, and study status. The first three columns correspond to data from the previous work, covering the period from September 2007 to March 2020. The next three columns present data from *Tri-AL* for the period from September 2007 to March 2022. Table 4.3 show the percentage difference between 2007-2020 and 2007-2022. Columns labeled T indicate the total number of trials, while columns labeled *Yes* and *No* count the trials with and without reported race/ethnicity, respectively. Each cell in Table 4.2 contains two numbers: the number of trials and, in parentheses, the percentage within the given category. Each cell in Table 4.3 reports the change in the number of trials ($\Delta n\%$), with the number in parentheses indicating the percentage difference

Table 4.2: Comparing study [TSW+22] and Tri-AL's result from 2007-2022

	[TSW+22] n(%)			Tri-AL n(%)		
Trial Feature	T	No	Yes	T	No	Yes
Funding						
Industry	7,717 (46.0)	4,345 (48.3)	3,372 (43.3)	7819 (29.7)	3311 (30.1)	4508 (29.4)
Academic	5,669 (33.8)	3202 (35.6)	2476 (31.7)	16508 (62.8)	7127 (64.9)	9381 (61.3)
US Government	3394 (20.2)	1441 (16.0)	1953 (21.1)	1931 (7.3)	532 (4.8)	1399 (9.1)
Primary Purpose						
Treatment	11877 (70.8)	6361 (70.8)	5516 (70.8)	18360 (69.9)	7733 (70.4)	10627 (69.5)
Basic Science	791 (4.7)	403 (4.5)	388 (5.0)	1218 (4.6)	474 (4.3)	744 (4.8)
Prevention	1289 (7.7)	688 (7.7)	601 (7.7)	2112 (8.0)	845 (7.7)	1267 (8.2)
Other	2319 (13.8)	1196 (13.3)	1123 (14.4)	4094 (15.5)	1596 (14.5)	2498 (16.3)
Missing	504 (3.0)	340 (3.8)	164 (2.1)	474 (1.8)	322 (2.9)	152 (0.9)
Phase						
N/A	5660 (33.7)	3130 (34.8)	2528 (32.4)	9353 (35.6)	4077 (37.1)	5276 (34.5)
Phase 1	1316 (7.8)	563 (6.3)	753 (9.7)	1964 (7.4)	683 (6.2)	1281 (8.3)
Phase 1/2-2	5623 (33.5)	2913(32.4)	2710 (34.8)	8722 (33.2)	3270 (29.8)	5452 (35.6)
Phase 2/3-3	1844 (11.0)	979 (10.9)	865 (11.1)	2762 (10.5)	1178 (10.7)	1584 (10.3)
Phase 4	2337 (13.9)	1401 (15.6)	936 (12.0)	3413 (13)	1750 (15.9)	1663 (10.8)
Enrollment						
0-9	2243 (13.4)	1341 (14.9)	902 (11.6)	3096 (11.7)	1388 (12.6)	1708 (11.1)
10-49	7280 (43.4)	4033 (44.9)	3247 (41.7)	11106 (42.2)	4799 (43.7)	6307 (41.2)
50-99	2958 (17.6)	1541 (17.1)	1417 (18.2)	4866 (18.5)	2055 (18.7)	2811 (18.3)
100-499	3499 (20.9)	1698 (18.9)	1801 (23.1)	5802 (22)	2214 (20.1)	3588 (23.4)
500-999	465 (2.8)	219 (2.4)	246 (3.2)	826 (3.14)	317 (2.8)	509 (3.3)
≥ 1000	335 (2.0)	156 (1.7)	179 (2.3)	562 (2.14)	197 (3.3)	365 (2.3)
Study Status						
Completed	13358 (79.6)	7093 (78.9)	6265 (80.4)	21100 (80.3)	8812 (80.3)	12288 (80.3)
Ongoing	338 (2.0)	76 (0.8)	262 (3.4)	466 (1.7)	39 (0.3)	427 (2.7)
Stopped Early	3073 (18.3)	1812 (20.2)	1261 (16.2)	4663 (17.7)	2110 (19.2)	2553 (16.6)
Unknown	11 (0.1)	7 (0.1)	4 (0.1)	29 (0.1)	9 (0.08)	20 (0.1)

(Δ%). Each value in the Table 4.3 is calculated by comparing the values from Table 4.2, the last three columns in 2022 (num_{2022}) with its corresponding value in 2020 from the first three columns (num_{2020}), using Equation 4.1 below.

$$\Delta n = \frac{(num_{2022} - num_{2020})}{num_{2020}} \tag{4.1}$$

Table 4.3 shows that the number of academically funded trials reporting race/ethnicity has increased by 93.4%, while the number of such trials with industry and US government funding has decreased by 32.0% and 56.9%, respectively. There are no significant changes (more than 20%) in the distribution of trials across different phases, primary purposes, and sizes. Regarding study status, although the number of trials reporting race/ethnicity has increased by 63%, the proportional percentage has decreased by 20.6% over the last two years.

4.4 System Performance

The performance evaluation of the XML parser selected for the Import Module is initiated by comparing it with a popular Python parser from the "BeautifulSoup" package, referred to as the

Table 4.3: Differences between study [TSW+22] and Tri-AL's result from 2007-2022

	$\Delta n\%$ ($\Delta\%$)		
Trial Feature	T	No	Yes
Funding			
Industry	1.3 (-35.4)	-23.8 (-37.7)	**33.7 (-32.1)**
Academic	191.2 (85.8)	122.6(82.3)	**278.9 (93.4)**
US Government	-43.1 (-63.9)	-63.1 (-70.0)	**-28.4 (-56.9)**
Primary Purpose			
Treatment	54.6 (-1.3)	21.6 (-0.6)	92.7 (-1.8)
Basic Science	54.0 (-2.1)	17.6 (-4.4)	91.8 (-4.0)
Prevention	63.8 (3.9)	22.8 (0.0)	110.8 (6.5)
Other	76.5 (12.3)	33.4 (9.0)	122.4 (13.2)
Missing	-6.0 (-40.0)	-5.3 (-23.7)	-7.3 (-57.1)
Phase			
N/A	65.2 (5.6)	30.3 (6.6)	108.7 (6.5)
Phase 1	49.2 (-5.1)	21.3 (-1.6)	70.1 (-14.4)
Phase 1/2-2	55.1 (-0.9)	12.3 (-8.0)	101.2 (2.3)
Phase 2/3-3	49.8 (-4.5)	20.3 (-1.8)	83.1 (-7.2)
Phase 4	46.0 (-6.5)	24.9 (1.9)	77.7 (-10.0)
Enrollment			
0-9	38.0 (-12.7)	3.5(-15.4)	89.4 (-4.3)
10-49	52.6 (-2.8)	19.0 (-2.7)	94.2 (-1.2)
50-99	64.5 (5.1)	33.4 (9.4)	98.4 (0.5)
100-499	65.8 (5.3)	30.4 (6.3)	99.2 (1.3)
500-999	77.6 (12.1)	44.7 (16.7)	106.9 (3.1)
$\geq$ 1000	67.8 (7.0)	26.3 (94.1)	103.9 (0.0)
Study Status			
Completed	58.0 (0.9)	24.2 (1.8)	96.1 (-0.1)
Ongoing	37.9 (-15.0)	-48.7 (-62.5)	**63.0 (-20.6)**
Stopped Early	51.7 (-3.3)	16.4 (-5.0)	102.5 (2.5)
Unknown	163.6 (0.0)	28.6 (-20.0)	400.0 (0.0)

"Baseline." The chosen parser is written in the C language. Figure 4.9 presents the results, displaying the data size on the x-axis (representing the number of trials that need to be parsed) and the running time in seconds on the y-axis. The results indicate that the selected parser scales significantly better than the baseline when handling larger data volumes.

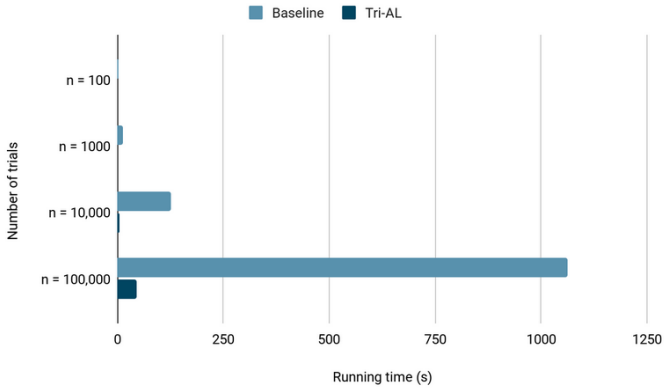


Figure 4.9: Performance of *Tri-AL* Parser

.

Chapter 5

Predicting the Mechanism of Action for Alzheimer's Disease Drugs

5.1 Introduction

The rise of online medical databases and reports has created a chaotic repository of unstructured, unlabeled, and sparse data on diseases and treatments. This information is often scattered across various locations due to clinical trials by different organizations or remains inaccessible due to sensitivity concerns. The sheer volume of data poses significant challenges, highlighting the importance of a system to predict labels for medical data, ensuring efficient organization and accessibility for analysis and research. Consequently, experts must implement processing, classification, and extraction pipelines to obtain information pertinent to their research, such as adverse drug reactions [SG14] and cancer stage detection [AGR$^+$18, CRH$^+$14]. Furthermore, researchers often need to gather additional information from other sources when developing reports on treatment and drug development for specific diseases. For instance, in the Alzheimer's Disease (AD) Drug Development Pipeline reports [CMZ14, CLR$^+$20], information from "ClinicalTrials.gov" is useful for cataloging details such as sponsors, disease phases, agents' criteria, and key trial dates. However, it only sometimes provides other crucial information, such as a drug's Mechanism of Action (MoA) or therapeutic purpose. Consequently, researchers must consult multiple resources, such as drug companies' websites, to fill in these gaps. Even then, they must use their expertise to decide how to label some columns, like MoA or the therapeutic purpose of a drug.

This chapter aims to automate the analysis and classification of AD-drug text for information not found on "ClinicalTrials.gov". Specifically, we focus on the MoA of a drug, which constantly

evolves based on experimental data from various sources beyond "ClinicalTrials.gov" and drug manufacturer websites. To identify the best classifier for Alzheimer's disease drug mechanisms of action (AD-drug MoA), we explored various machine learning algorithms, including Random Forest (RF), XGBoost, Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), and a Multi-layered Neural Network (NN) based on the BioBERT Encoder. Preliminary results on AD drugs' free-text indicate that the BioBERT-based NN achieved the highest F1 score of 0.97. However, the Decision Tree algorithm, with an F1 score of 0.92, also performed well. Considering the complexity trade-off, the Decision Tree's output is reasonable.

The main contributions of our study can be summarized as follows: First, we recognize that different information related to AD drugs is available across various resources. Therefore, we collect and merge this data to create a contiguous dataset. Second, we evaluate different machine-learning methods to identify the best model for classifying AD-drug texts. Our results indicate that the Decision Tree algorithm is the most effective model for this classification task. This model can play a crucial role in aiding the medical community in generating annual reports on AD-drug progression [CLR$^+$19].

5.2 Methods

Five different machine learning models, Random Forest (RF), XGBoost, Logistic Regression (LR), Support Vector Machines (SVM), and Decision Tree (DT), were applied to classify the Mechanism of Action (MoA) of AD drugs. This section provides an explanation of the dataset, a detailed description of the models used, and a discussion of the experimental results achieved by these methods.

5.2.1 Dataset

Background and finding texts for each drug were collected from the ALZFORUM Therapeutics dataset (http://www.alzforum.org). Using the drugs' targets from the website and the MOA classes of medications provided by [CLR$^+$19], [CLRZ18], a total of 233 labeled records were obtained. The MOA of AD drugs is classified into two main categories: **small molecules** and **Disease-Modifying Therapies (DMT) Biologics**. The dataset was divided into 75% (186 records) for training and 25% (47 records) for testing. The dataset is imbalanced, with 175 records in the first class and 59 in the second. To address this, the Synthetic Minority Over-sampling Technique

(SMOTE) [CBHK02] was applied to the training set, resulting in 141 records for each class. The overall text portion of the dataset comprises 124,180 words and 7,693 sentences.

5.2.2 Selecting the Best ML Algorithm

To achieve the best results, several machine learning algorithms were tested to determine which one performs best. In this study, various machine learning algorithms were employed for document classification. Decision Tree (DT) demonstrated high accuracy and practicality in selecting important words. XGBoost, a Gradient Boosting-based Decision Tree algorithm, was noted for its speed and performance. Random Forest (RF), which aggregates predictions from multiple Decision Trees, effectively addressed overfitting issues. Logistic Regression (LR) modeled event occurrence as a linear function of predictor variables, providing probabilistic outcomes. Lastly, the Support Vector Machine (SVM) with an RBF kernel offered robust performance in text data classification, ensuring reliable separation of data in higher dimensions while resisting overfitting.

Data preparation for ML models: First, we cleaned the data by eliminating stop words, punctuation, extra white spaces, undesired characters, and words lacking meaningful information. We then applied the Term Frequency-Inverse Document Frequency (TF-IDF) method to create a matrix representation of the text. This method leverages two components: Term Frequency (TF), which measures how often a term appears in a document; and Inverse Document Frequency (IDF), which assesses how common or rare a term is across all documents. The TF-IDF value is computed as follows:

$$tf_{ij}idf_i = tf_{ij} \times \log_2\left(\frac{N}{df_i}\right) \tag{5.1}$$

where N is the total number of documents, tf_{ij} is the frequency of term i in document j, and df_i is the number of documents containing term i. Using this method, the five machine learning algorithms in our study showed promising results.

5.2.3 BioBERT-based NN Model

We opted for the BERT model to classify the Mechanisms of Action (MoA). BERT is a text encoding model that has achieved state-of-the-art results across various tasks. It operates as a bidirectional transformer network pre-trained on extensive language modeling tasks using large datasets. For our specific needs, we employ BioBERT [LYK+20], a pre-trained model fine-tuned on biomedical

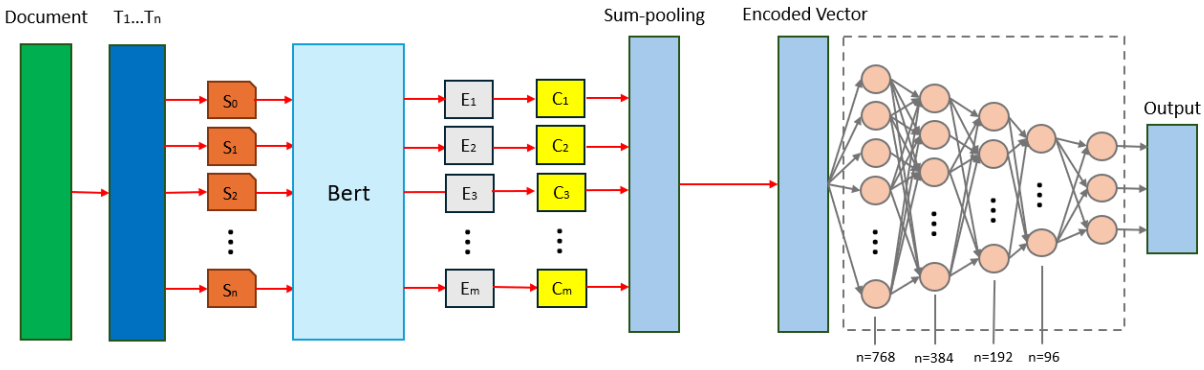


Figure 5.1: Overview of BioBERT-based NN Model

texts from PubMed and PMC. BioBERT excels in tasks such as text mining, classification, and other natural language processing (NLP) applications within the biomedical domain. Algorithm 25 indicates the step of our BioBERT-based NN model for MoA text classification. This model is a fine-tuned version of BERT_BASE, featuring 12 hidden layers and an embedding size of 768. Transformers like BERT process words in a sentence simultaneously, which means the sequence of words is not inherently preserved. They utilize Position Embeddings to retain positional information. Combined with Word Embeddings, these embeddings create a single representation that the model can process. Initially, BERT's WordPiece tokenization breaks the input sentence into subword tokens. For instance, the sentence "Elevated blood glucose levels are linked to diabetes and heart disease" is tokenized into Elev, ##ated, blood, gl, ##ucose, levels, are, linked, to, diabetes, and, heart, disease. Each subword token is then converted into a 768-dimensional vector. The first token in any sequence is a unique [CLS] token, representing the entire sequence. BioBERT has a limitation of processing a maximum of 512 tokens per input sequence. To handle longer documents, we use a sliding window approach, dividing the document into smaller segments. For example, using a window size of 4, the sentence is split into Elev, ##ated, blood, gl, ##ucose, levels, are, linked, to, diabetes, and, heart, disease. We then add [CLS] and [SEP] tokens at the beginning and end of each window and input these segments into the BioBERT model. The output of BioBERT is a list of embeddings, as shown in the following equation:

$$e_i = BioBERT(t_i) \tag{5.2}$$

Where input text T into tokens: $T = \{t_1, t_2, \ldots, t_n\}$ and $e_i \in \mathbb{R}^{768}$.

To classify these embeddings, we need a single [CLS] representation for each document. This

representation is obtained by summing the [CLS] vectors from each window, as shown below:

$$CLS = \sum_{i=1}^{m} C_i \tag{5.3}$$

where C_i represents the [CLS] token embedding from each window. Figure 5.1 illustrates the detailed layers of the model. The BioBERT output is processed by a five-layer fully connected sequential Neural Network to classify Mechanisms of Action (MoA). The architecture starts with an initial layer of 768 neurons. Each subsequent layer has half the number of neurons as the preceding one, with each neuron utilizing a Rectified Linear Unit (RELU) activation function. This results in layers containing 768, 384, 192, and 96 neurons. The final layer, responsible for classification, features two neurons and employs a softmax activation function to assign scores to the classes.

5.2.4 Results

The previously mentioned dataset, comprising 141 records for training and 47 for testing, was utilized in our experiments. We employed Python libraries such as Gensim for cleaning clinical text [RS10], and Scikit-learn [PVG+11], along with Pandas [McK10] for model implementation.

Table 5.1 summarizes the results for each model, measuring Precision, Recall, F1-Score, Accuracy, and ROC Area. A visual comparison of the models' performance based on the ROC area metric is provided in Figure 5.2. The findings indicate that, although the Decision Tree (DT) model achieves higher accuracy compared to the Support Vector Machine (SVM) and Logistic Regression (LR) models, SVM and LR demonstrate superior precision. This indicates that SVM and LR are more cautious and accurate in their class predictions. The higher precision of SVM and LR shows that their predictions are more aligned with the actual classes than those of DT in various classifications. The application of TF-IDF enhanced performance, highlighting the models' sensitivity to word presence or absence rather than contextual meaning. Figure 5.3 depicts the DT process on the dataset, utilizing the Gini Index as the splitting criterion. This index measures overall variance within the tree classes, making it a suitable criterion for node purity [BK18]. As shown in the trained Decision Tree (DT) in Figure 5.3, the height of the tree is 4, and the word "antibody" is chosen by the algorithm to separate the classes at the first level. Using TF-IDF for feature extraction, when the score of "antibody" is more than 0.017, the DT was able to correctly classify 34 out of 40 total samples of the second class. Other words selected by the DT include "delivered," "mild," "mimics," and "intracerebral."

Algorithm 2: BioBERT-NN Based Model

Data: Input text T
Result: Output probabilities from the neural network

Step 1: Tokenization;
Split the input text T into tokens: $T \rightarrow \{t_1, t_2, \ldots, t_n\}$;

Step 2: Initial Embedding;
for *each token* t_i *in* $\{t_1, t_2, \ldots, t_n\}$ **do**
    Convert t_i into an embedding vector e_i of size 768: $t_i \rightarrow e_i \in \mathbb{R}^{768}$;
end

Step 3: Transformation through Layers;
for *each layer* l *in* $1, 2, \ldots, 12$ *(for BioBERT model)* **do**
    for *each embedding vector* e_i **do**
        Pass e_i through the transformer layer l to obtain a new representation h_i^l;
    end
end

Step 4: Final Hidden States;
for *each token* t_i **do**
    The final representation of t_i is the hidden state h_i^{12}: $h_i^{12} \in \mathbb{R}^{768}$;
end

Step 5: Neural Network;
Feed *outputs* into a five-layer fully connected sequential Neural Network;
Define the sequential model as follows:;
$h^{(1)} \leftarrow \text{RELU}(W^{(1)} \cdot \text{CLS} + b^{(1)})$;
$h^{(2)} \leftarrow \text{RELU}(W^{(2)} \cdot h^{(1)} + b^{(2)})$;
$h^{(3)} \leftarrow \text{RELU}(W^{(3)} \cdot h^{(2)} + b^{(3)})$;
$h^{(4)} \leftarrow \text{RELU}(W^{(4)} \cdot h^{(3)} + b^{(4)})$;
output $\leftarrow \text{Softmax}(W^{(5)} \cdot h^{(4)} + b^{(5)})$;

return *output*

Table 5.1: MoA classification results

Algorithms	Precision	Recall	F1-score	Accuracy
Random Forest	0.857	0.46	0.6	0.82
XGBoost	0.76	0.76	0.76	0.87
Logistic Regression	0.9	0.62	0.78	0.89
SVM	0.9	0.69	0.78	0.89
Decision Tree	0.86	1.00	0.92	0.95
BioBERT-based NN	0.95	1.00	0.97	0.96

In our evaluation, the BioBERT-based NN achieved the highest F1 score of 0.97, while the Decision Tree algorithm, with an F1 score of 0.92, also showed promising results. Considering the complexity trade-off, the Decision Tree's performance is suitable.

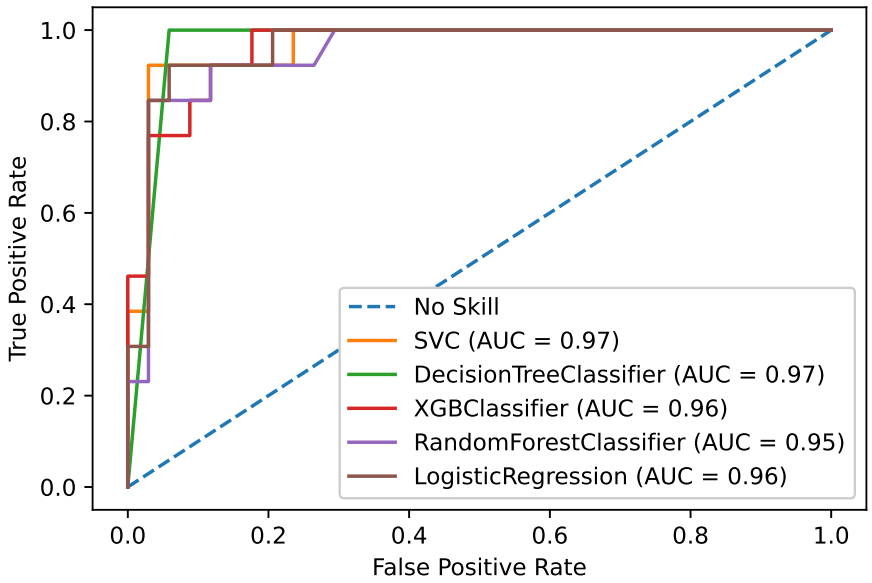


Figure 5.2: ROC Curve for ML Algorithms

5.2.5 Conclusion

This chapter makes an effort to automate information retrieval from "ClinicalTrials.gov," aiming to generate potential candidates for expert verification, such as identifying the Mechanism of Action (MoA) for a drug. The developed models will assist in predicting various categories typically evaluated by experts. The intention is to support, rather than replace, the experts by providing ideal candidates for their review, akin to the functionality of a spell checker. The success of this

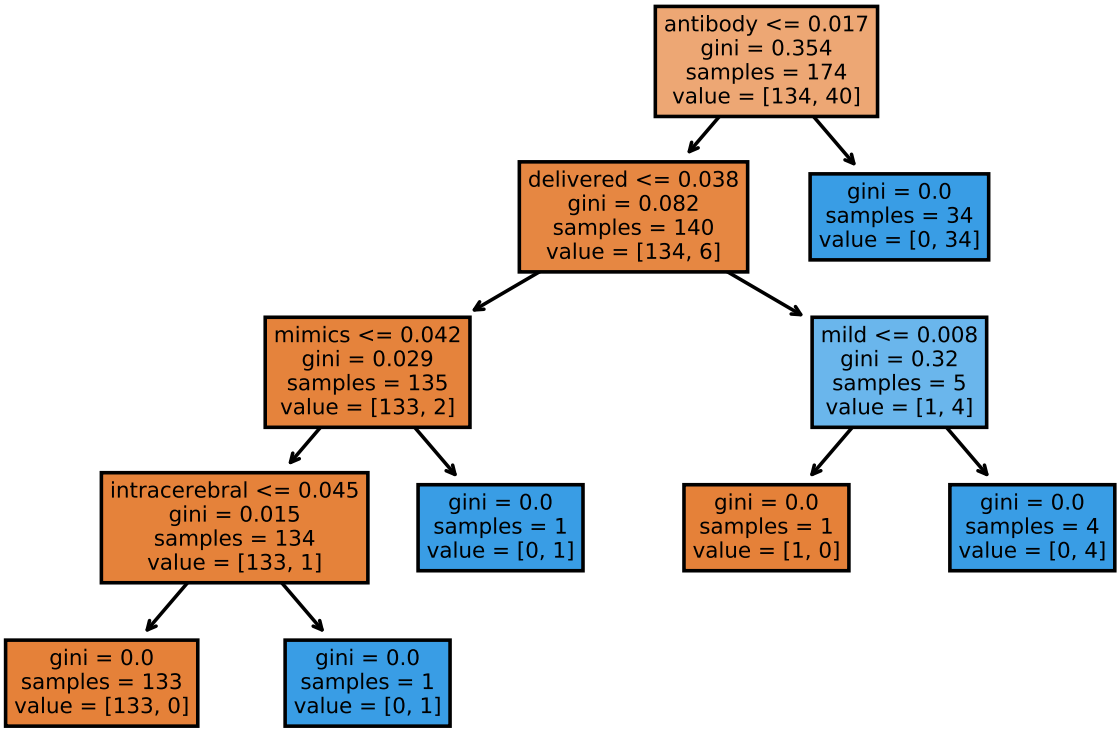


Figure 5.3: Decision Tree for Classifying MoA in AD Texts

work will be measured by its ability to highlight links that might have been overlooked or to save experts' time in suggesting MoAs for drugs. Their time can then be redirected toward making significant advances in discovering new treatments or even a cure.

Chapter 6

Joint Named Entities and Relation Extraction

The development of automated systems for knowledge extraction from unstructured texts significantly reduces the time experts spend on manual data processing [KNC+22]. This automation primarily involves two pivotal tasks utilizing natural language processing (NLP): Named Entity Recognition (NER) and Relation Extraction (RE). NER Named Entity Recognition (NER) is a task in NLP that involves identifying and classifying proper nouns in text into predefined categories. These categories typically include names of people, organizations, locations, dates, numbers, and sometimes more specialized categories like product names, events, or expressions of time. In a biomedical context, it identifies specific entities such as genes, chemicals, and diseases within texts. The goal of NER is to automatically scan entire documents and extract fundamental data points, helping machines understand the text by highlighting which words (entities) carry essential information. This task is crucial for applications such as information retrieval, content classification, and data extraction for further processing or analysis.

Relation Extraction (RE) involves identifying and classifying semantic relationships between entities within a text. This process helps construct a structured understanding of textual data by linking identified entities (people, organizations, locations) through specific relationships. For example, in the sentence "Barack Obama was born in Hawaii," an RE system would identify "Barack Obama" and "Hawaii" as entities and classify the relationship between them as "born in." The primary aim of relation extraction is to convert unstructured text into a structured format that can be used in databases, knowledge graphs, or directly for tasks such as information retrieval,

question answering, and summarization. RE frames these associations as triples in the form of *Entity-1, Relation-type, Entity-2* [ZSLW20]. This is critical in biomedical research, where understanding the relationships between diseases, drugs, and treatments in literature can lead to new insights and discoveries. In addition, this function is particularly beneficial in clinical trials, as it facilitates the extraction of vital information from textual data, potentially leading to novel scientific breakthroughs.

Types of Relational Tuples: Documents often contain multiple triples, categorized based on entity overlap into three groups: 'normal' sequences with no overlapping entities, 'single entity overlap' (SEO) where at least one triple shares an entity with another, and 'entity pair overlap' (EPO) where at least two triples share the same entities but differ in their relation types [ZZH$^+$18] [EEH22]. EPO sequences, in particular, present intricate multi-label classification challenges. For example, consider a biomedical text stating, "Aspirin increased the expression of P53 and decreased the expression of MDM2, significantly altering the P53/MDM2 ratio." Here, the Chem-GENE pair ("Aspirin," "P53") is classified under both CPR:3 and CPR:4 relation types in the BioCreative dataset [KRA$^+$17].

Joint Named Entity and Relation Extraction (JNERE) is an NLP technique that simultaneously identifies named entities and their relationships within a text. Unlike pipeline methods that perform NER and RE as separate tasks in sequential steps, JNER integrates both tasks into a single model. This approach significantly enhances accuracy and efficiency by leveraging the interdependencies between entity recognition and relation extraction, capturing contextual information more effectively and reducing the propagation of errors that often occur in pipeline methods. By jointly modeling these tasks, JNERE can provide a more coherent and contextually accurate extraction of information, particularly beneficial in complex text analysis scenarios, such as biomedical text mining, when understanding the interactions between entities is as essential as identifying the entities [LFM$^+$19] [ZWB$^+$17]. In Chapter 2, Section 2.4, we provide a comprehensive explanation of this area.

In this chapter, we describe the application of the BioRIFRE model, which utilizes Graph Neural Network (GNN) for joint-named entity and relation extraction [ZXC$^+$21], specifically for gene and chemical components named entities. This approach enhances the representation of words and relationship nodes by iteratively augmenting them within the graph structure. We conducted extensive tests using various encoders and discovered that BioBERT provides the best performance, significantly outscoring other models. BioBERT's advanced understanding of biomedical texts

allowed it to achieve superior results in the relation extraction task compared to related work. The model achieves a 0.69 F1-score state-of-art for the relation extraction task.

6.1 Method

This section provides a detailed overview of the Bio-RIFRE model's architecture, thoroughly explaining each layer and component. The architecture, illustrated in Figure 6.1, consists of three primary layers: the representation layer, the Graph Attention Network (GAN) layer, and the final taggers. The word embedding layer transforms input words into continuous vector representations, capturing semantic meanings and syntactic roles, and serves as the foundation for further processing. The GAN layer, at the core of the model, processes these word embeddings using attention mechanisms to focus on the most relevant parts of the input text and augment the words and relation classes' representations. It constructs a graph where two types of nodes represent words and relations. It dynamically assigns an edge between two node types if an entity has that relation in the text, resulting in a comprehensive and augmented representation of the input text [ZC20]. The final layer consists of specialized chemical and gene taggers that use the enhanced representations from the GAN layer to accurately identify and classify chemical and gene entities within the text by applying domain-specific knowledge. The trained model outputs hidden relation embedding vectors from these layers, encapsulating the intricate relationships between paired entities within a sentence, such as interactions between chemicals and genes. The model can effectively classify and interpret the relationships by leveraging these embeddings, providing valuable insights for tasks like information extraction and data mining in biomedical texts.

6.1.1 Words and Relation Representation

The initial layer of the model employs two distinct word embedding components to convert the sentence and its relations into vector representations. For a sentence with N words, we use the BioBERT model [LYK$^+$19], which is an adaptation of the BERT, to produce word representation. The BERT model, introduced by [DCLT18a], utilizes the encoder block from the transformer architecture [VSP$^+$17] to derive deep bidirectional representations by considering the context from both sides of each word in a sequence [DCLT18a][ET21][HJ20].

To elaborate, BioBERT is specifically fine-tuned for biomedical text, leveraging the strengths of the original BERT model. The BERT model itself revolutionized natural language processing

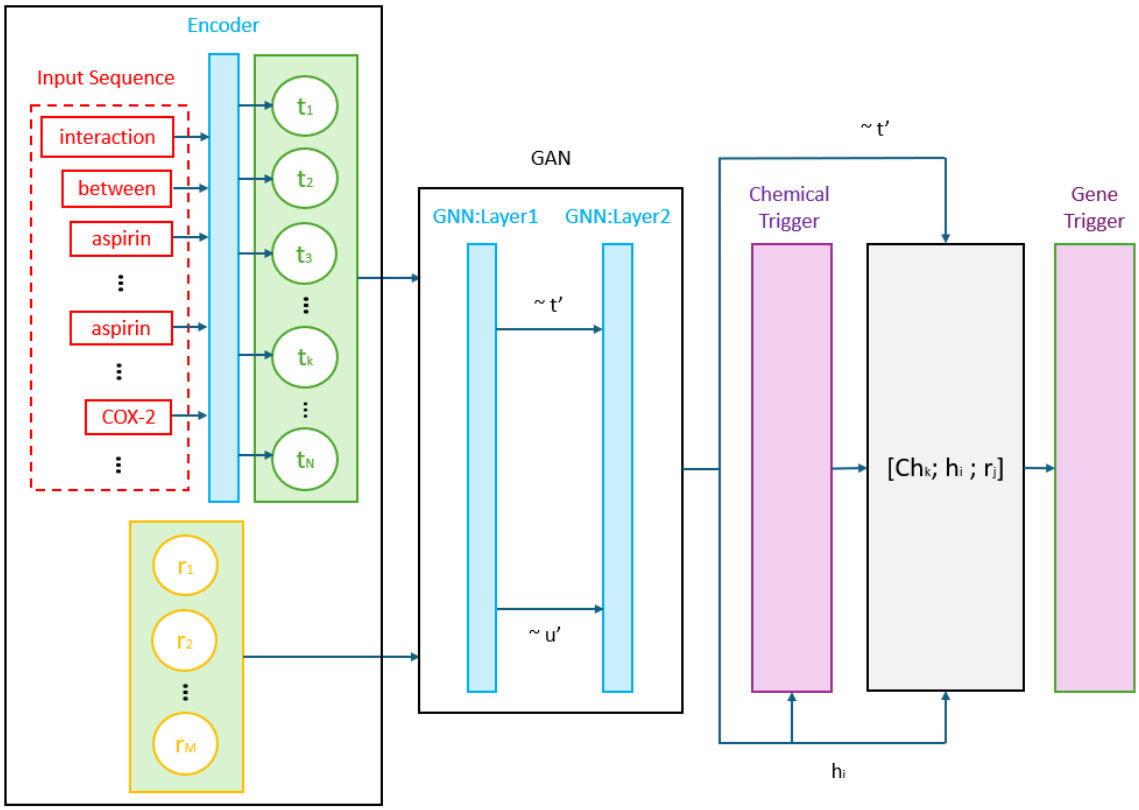


Figure 6.1: An Overview of Bio-RIFRE Model

by enabling deep contextual understanding through its bidirectional approach, where it reads the text in both directions (left-to-right and right-to-left) to better capture the meaning of each word based on its context. This method significantly enhances the model's ability to understand the nuances and complexities of language, particularly useful in specialized fields such as biomedical text processing.

The BioBERT model has demonstrated superior performance compared to earlier unidirectional models, and its pre-trained architecture enables us to use it without additional training. For relation representations, for a sentence with M relations, we initialize one-hot vector representations for each relation by random numbers. These relation vectors are intended to be trained in subsequent layers, where their representations will be enriched with word embeddings.

As depicted in Figure 6.1, the representation layer outputs word tensors $[t_1, t_2, \ldots, t_N]$ and relation tensors $[r_1, r_2, \ldots, r_M]$, where N is the length of the sequence, and M is the number of relation classes. These tensors are subsequently fed into the GAN layer, which refines them and incorporates prior knowledge of potential relationships into both the word and relation representations.

6.1.2 Graph Attention Neural Networks

Graph Attention Networks (GANs) were introduced by [VCC+18] to apply a self-attention mechanism specifically to graph-structured data. This innovation marked a significant advancement in how graphs are processed, allowing for more flexible and context-aware data representations within a graph. Before the advent of GAN models, Convolutional Neural Networks (CNNs) [LBBH98] had achieved remarkable success across various domains, most notably in image processing, where their ability to capture spatial hierarchies in grid-like data structures proved highly effective. Inspired by the success of CNNs, researchers sought to generalize convolutional techniques to handle graph-structured data [HR20], which do not possess a regular grid-like structure.

Traditional CNNs, while powerful for structured data like images, struggle with the irregularity and complexity of graph data. Graphs often represent entities and their relationships in a non-Euclidean space, making direct application of CNNs impractical [VCC+18]. The challenge lies in the inherent heterogeneity and the variable nature of graph node connections, contrasting with the fixed uniform connectivity in image pixels.

GANs address this challenge by leveraging a self-attention mechanism that dynamically weighs the importance of neighboring nodes when aggregating information. This approach allows GANs to focus on the most relevant parts of the graph, enhancing the representation of each node based

on its context within the network. The self-attention mechanism thus provides a more nuanced and adaptable way of processing graph data, capturing intricate dependencies and relationships that are often missed by traditional CNNs.

The introduction of GANs has opened new avenues for applying deep learning to graph-structured data, including node classification, link prediction, and graph classification. By overcoming the limitations of CNNs in handling non-grid-like data, GANs have enabled significant advancements in various applications, ranging from social network analysis to biological network modeling, where understanding complex relationships and interactions is crucial [WJS+19].

In this study, we utilize the model proposed by [ZXC+21]. Specifically, we consider two types of nodes: word nodes and relation nodes, which are connected if they participate in at least one triple. Each GAN layer of the model follows three steps to update the vector representations of nodes using an attention mechanism, as detailed below.

Calculate Attention Weights:

Given a sentence containing N words and M relations, we denote the words as t_i and the relations as r_i. Each word embedding t_i and relation embedding r_i are treated as nodes within a graph, with connections established between them if they are neighbors. Figure 6.2 visually depicts the relationships and connections between these nodes. This graph structure enables the model to effectively capture and utilize the contextual interactions between words and their corresponding relations.

For each pair of word-relation nodes in the graph, we calculate the attention weights, denoted as α, using the following equations [VCC+18]:

$$a_{ij} = W_a[W_q t_i : W_k r_j] \tag{6.1}$$

$$\alpha_{ij} = \frac{\exp(a_{ij})}{\sum_i^N \exp(a_{ij})} \tag{6.2}$$

Here, W_a, W_q, and W_k are trainable parameters within the neural network, and [:] represents the concatenation of two high-dimensional vectors. These attention weights help the model focus on the most relevant word-relation pairs by assigning higher importance to certain connections, thereby enhancing the contextual understanding of the graph structure.

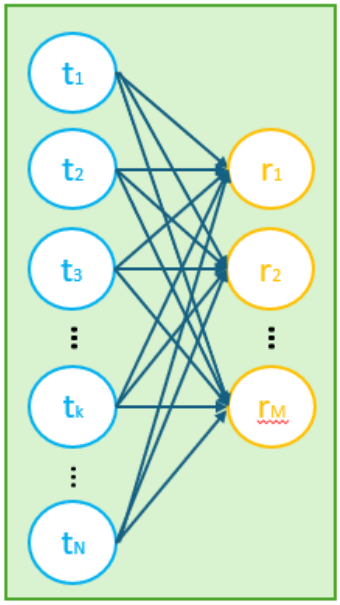


Figure 6.2: Graph Neural Network architecture

6.1.3 Augmenting the Representations:

After determining the attention weights for every pair of word-relation nodes, we proceed to update the vector representation of each node within the graph. For each node t_i, its vector representation is updated by considering all the relation nodes and their corresponding significance to t_i. This update is performed using the following equation:

$$t'_i = t_i + \sum_j \alpha_{ij} r_j \tag{6.3}$$

In this equation, α_{ij} represents the attention weight between node t_i and relation node r_j. Figure 6.3 illustrates the updating process for the first-word node ($i = 1$) within the GAN. This method ensures that each node's updated vector representation incorporates the influence of its connected relation nodes, weighted by relevance.

We perform the same calculation to update each r_i node, taking into account their attention to all the other u_i nodes:

$$r'_i = r_i + \sum_j^N \alpha_{ij} W_r t_j \tag{6.4}$$

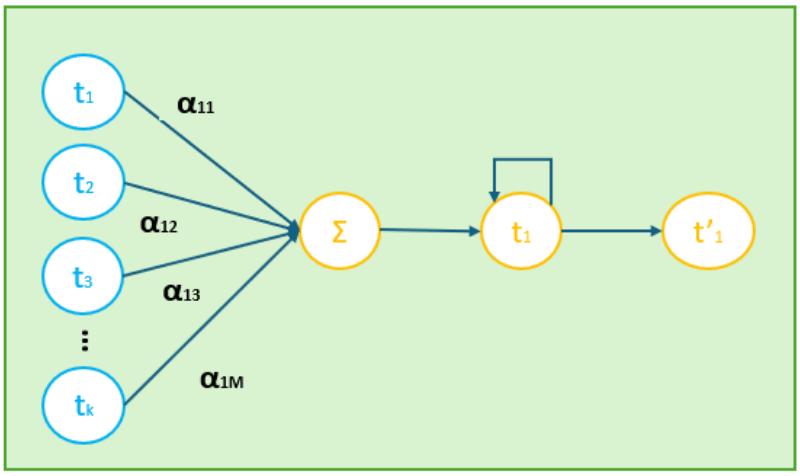


Figure 6.3: Schema of Word and Relation Node Updates in GAN

6.1.4 RE and Taggers

After updating the representations of words and relations in the final Graph Neural Network (GNN) layer, we utilize the CasRel framework, as proposed by [ZXC+21], to detect all potential chemical and gene entities within a given sequence. This model employs two sequential processes, the Chemical Tagger and the Gene Tagger, to methodically identify and classify these entities.

First, the Chemical Tagger is applied to the sentence, identifying and tagging chemical entities using a combination of the updated word and relation representations. The Chemical Tagger focuses on detecting chemical names' start and end positions within the text. Once the chemical entities are tagged, their information is fed into the Gene Tagger.

Next, the Gene Tagger uses the tagged chemical entities as context to identify gene entities. It operates similarly to the Chemical Tagger but is specifically tuned to recognize gene names. The Gene Tagger also considers the relationships between the previously identified chemical entities and the potential gene entities, ensuring accurate classification of both entity types.

This two-step tagging process, with each step leveraging the outputs of the previous one, allows the model to identify and classify chemical and gene entities systematically, ensuring a comprehensive understanding of the relationships within the sequence [ZXC+21].

The equations guiding these processes are crucial for transforming the initial word and relation representations into meaningful entity tags. For instance, the Chemical Tagger utilizes the updated vector representations from the GNN layer, applying specific rules and thresholds to determine the presence of chemical entities. Similarly, the Gene Tagger builds upon this foundation, incorporating

the identified chemical entities to enhance its tagging accuracy for gene names.

This cascading approach, enabled by the CasRel framework, enhances the model's ability to handle complex biomedical texts, accurately extracting and classifying critical information about chemical and gene entities [ZXC+21].

Chemical Tagger

We use the updated t_i embeddings from the final layer of the GAN network as input for the chemical tagger. This tagger consists of two binary classifiers that are designed to pinpoint the first and last word positions of chemical named entities within a given input sentence.

$$p_i^{start_{ch}} = \sigma(W_{start} t_i' + b_{start}) \tag{6.5}$$

$$p_i^{end_{ch}} = \sigma(W_{end} t_i' + b_{end}) \tag{6.6}$$

Firstly, using Equations 6.5 and 6.6, we calculate the probability of each token in the input being a named entity. $p_i^{start_{ch}}$ indicates the probability that the i^{th} token in the sentence is the starting position of the chemical, while $p_i^{end_{ch}}$ represents the probability that the i^{th} token is the ending position of the chemical. The parameters W_{start}, W_{end}, b_{start} and b_{end} are trainable, and σ is the activation function [WSW+20]. The goal of the chemical tagger is to minimize the following likelihood function for the binary classifier:

$$p_\theta(ch|x) = \prod_{t \in \{start_{ch},\ end_{ch}\}} \prod_{i=1}^{N} (p_i^t)^{I\{y_i^t=1\}} (1 - p_i^t)^{I\{y_i^t=0\}} \tag{6.7}$$

where θ represents the parameter set of the chemical tagger. For a given sentence x, Equation 6.7 computes the probability of each word being a chemical component, with t defining the start or end tags as $y_i^{start_{ch}}$ or $y_i^{end_{ch}}$, and $I\{y_i^t = 1\}$ indicating if the binary classifier tag is true (1) or false (0).

We merge the outputs from the GAN layer for word nodes (h) and relation nodes (r) with the output from the chemical tagger (ch). This combined result is then processed through the tanh activation function, resulting in the embedding vector w_{ijk}'. This vector is subsequently used as the input for the Gene tagger's binary classifier. The Gene tagger then identifies the start and end positions of gene-named entities within the sentence and determines their relationships with chemical-named entities. Equation (6.8) demonstrates the formulation of the input for the Gene

tagger. The high-dimensional vector w'_{ijk} captures detailed information about the chemical-named entities and their potential interactions with gene entities.

$$w'_{ijk} = \tanh(W_h[ch_k; r_j; h_i] + b_h) \tag{6.8}$$

Gene Tagger

The gene tagger identifies a list of potential gene candidates for a given chemical named entity by iteratively considering all relations in the sentence. Equations 6.9 and 6.10 calculate the probability of the i^{th} position being the start or end of the gene span. W_{start}, W_{end}, b_{start}, b_{end} are trainable parameters.

$$p_i^{start_g} = \sigma(W_{start} w'_{ijk} + b_{start}) \tag{6.9}$$

$$p_i^{end_g} = \sigma(W_{end} w'_{ijk} + b_{end}) \tag{6.10}$$

The likelihood for the gene binary classifiers follows a similar structure to that of the chemical tagger described earlier. It utilizes the same definitions and principles outlined in the chemical tagger subsection. The gene binary classifiers use this function to determine the probability of each word being part of a gene named entity, ensuring consistency in the tagging approach.

$$p_\theta(g|x, ch, r) = \prod_{t \in start_s, end_s} \prod_{i=1}^{L} (p_i^t)^{I\{y_i^t=1\}} (1 - p_i^t)^{I\{y_i^t=0\}} \tag{6.11}$$

Finally, the loss function for the entire model can be defined using the method described in [ZXC+21] by combining the likelihood functions of the chemical and gene taggers. It is equal to taking log from equation 6.12:

$$\begin{aligned} L = \log \prod_{(ch,r,g) \in T} p((ch, r, g)|x) = \sum_{ch \in T_j} \log p_{\theta_{ch}}(ch|x) + \sum_{r \in T_j|ch} \log p_{\theta_g}(g|x, ch, r) + \\ \sum_{r \in R \setminus T_j|ch} \log p_{\theta_0}(g_\emptyset|x, ch, r) \end{aligned} \tag{6.12}$$

6.2 Method Evaluation

Table 6.1 presents a comparison of different pre-trained BERT models based on their performance metrics: F1 score (F1), Recall (R), and Precision (P). We tested several models, including SciBERT,

SapBERT, BERT-based-cased, and BioBERT, to determine which model performs best in our specific application. Based on our evaluation, BioBERT delivered the best results across all metrics, making it the most suitable model for our application.

Table 6.1: Comparison of various pre-trained BERT models

Model	**F1**	**R**	**P**
SciBERT [BLC19]	0.41	0.46	0.37
SapBERT [JCL$^+$20]	0.41	0.40	0.42
BERT-based-cased [DCLT18b]	0.60	0.59	0.61
BioBERT [LYK$^+$19]	0.69	0.70	0.68

For the task of JNERE in biomedical texts, the model must accurately predict the start and end positions of chemical-gene spans, ensuring that substrings are not considered valid named entities. Additionally, it must correctly identify the relationships between chemical and gene entities. The model's performance is evaluated using Precision (P), Recall (R), and the Micro F1-score. Our experiments with various pre-trained models indicated that the BioBERT model [LYK$^+$19], pre-trained on PubMed abstract datasets, is the most suitable for our needs.

We utilize the BioBERT-cased model pre-trained in PyTorch, with an embedding layer that supports a maximum token size of 512. Consequently, the maximum sequence length is also 512, and only triples where the chemical-gene entities fall within this length are considered, excluding all inter-sequence triples.

For optimization, we employ Stochastic Gradient Descent (SGD) with early stopping set to 20 iterations to prevent overfitting. The learning rate is set at 0.1, and the batch size is 10. We compare Bio-RIFRE with six baseline models that have evaluated their joint methods on the CPI dataset. All the performance metrics are extracted from the recent studies by [ZZ21] and [SYW$^+$22].

- **Att-BiLSTM-CRF+ELMO** [LYC$^+$20] is an attention-based model designed for named entity recognition and relation extraction. It employs a Convolutional Neural Network (CNN) to capture character-level features and integrates a pretrained ELMO model in the embedding layer to enhance word representations. These features are concatenated and fed into two Bidirectional Long Short-Term Memory (BiLSTM) networks, followed by a Conditional Random Field (CRF) layer to predict the final tags.

- **Dygiepp** [ZLC$^+$19a] consists of two main layers built on top of the word embedding layer: the mention recognition layer and the relation extraction layer. This model uses a CRF

to calculate scores for identifying microorganism entities and determining their relationships. The architecture is designed to effectively handle overlapping entities and complex relationship types.

- **SpanMB Bert** [ZZ21] proposes a novel span-based framework that fine-tunes BERT to derive contextual embeddings. It then calculates span scores to classify both entities and relation types. This approach allows the model to consider spans of text as potential entities, improving its ability to capture and classify complex biomedical relationships.

- **MRC4BioER** [SYW+22] transforms the task of joint named entity and relation extraction into a Machine Reading Comprehension (MRC) task. In this model, sentences are treated as the context, relations as the target query, and entity spans as the answer. The model uses BERT for embedding representations and includes a tagging algorithm that addresses the problem of overlapping entities by iteratively evaluating each span pair for different relation classes.

- **MRC4BioER (Zheng's tagging)** is a variation of the MRC4BioER model that utilizes Zheng's tagging schema [ZWB+17]. This schema provides a systematic approach to tagging entities and relations, enhancing the model's accuracy in recognizing complex biomedical interactions.

- **MRC4BioER (Lou's tagging)** is another variation of the MRC4BioER model that employs Lou's tagging schema [LYC+20]. This schema incorporates attention mechanisms and CRF layers, further refining the model's capability to accurately tag and classify biomedical entities and their relationships.

6.2.1 Result

Table 6.2 showcases the performance of various models on the Chemical-Protein Interaction (CPI) dataset. The results highlight that incorporating BERT in the embedding layer leads to a significant improvement in model performance. Specifically, the SpanMB BERT model achieves a state-of-the-art F1-score of 0.88 for the Named Entity Recognition (NER) task, indicating its exceptional ability to identify chemical and gene entities within biomedical texts accurately.

Among the baseline models, MRC4BioER demonstrates the best performance for the Relation Extraction (RE) task, with an F1-score of 0.66. This model converts the joint named entity

Table 6.2: Evaluation of Bio-RIFRE on the CPI dataset

Model	NER			RE		
	F1	**R**	**P**	**F1**	**R**	**P**
Att-BiLSTM-CRF+ELMo	0.811	0.798	0.825	0.551	0.512	0.595
Dygiepp	0.887	0.876	0.897	0.629	0.605	0.654
SpanMB Bert	**0.888**	**0.883**	**0.893**	0.646	0.615	0.680
MRC4BioER (Zheng's tagging)	-	-	-	0.615	0.539	0.717
MRC4BioER (Lou's tagging)	-	-	-	0.624	0.556	0.711
MRC4BioER	-	-	-	0.660	0.617	0.70
Bio_RIFRE	0.870	0.851	0.860	**0.690**	**0.70**	**0.681**

and relation extraction task into a Machine Reading Comprehension (MRC) task, which helps in effectively identifying relationships between entities.

Although Bio-RIFRE does not achieve the highest F1-score in the NER task compared to the SpanMB BERT model, it excels in the RE task with an F1-score of 0.69, surpassing all other baseline models. This superior performance in the RE task suggests that Bio-RIFRE is particularly effective at accurately classifying the relationships between chemical and gene entities once the spans are correctly identified. This indicates that Bio-RIFRE has a robust mechanism for understanding and categorizing the interactions between entities, which is crucial for detailed biomedical text analysis.

6.3 Discussion

The baseline models were compared based on their ability to manage sequences with overlapping chemical-gene named entities and sequences with varying triples per sequence. The evaluation demonstrated that Bio-RIFRE maintained its robustness even when applied to datasets with more relation classes, as evidenced by its performance on the DrugProt dataset. The results of this evaluation are presented in Table 6.2.

6.3.1 Chemical-Gene Overlapping

The Bio-RIFRE model was compared with various baseline models on different types of sequences, including Normal, SEO, and EPO sequences. The MRCBioER variation models were the only baseline models that provided results on these different sequence types. Figure 6.4 vividly illustrates the F1-scores of the Bio-RIFRE model and three MRCBioER models. The results clearly show that Bio-RIFRE significantly improved the F1-score by 3%, 2%, and 32% for Normal, SEO, and EPO sequences, respectively. This notable improvement for EPO sequences underscores the model's

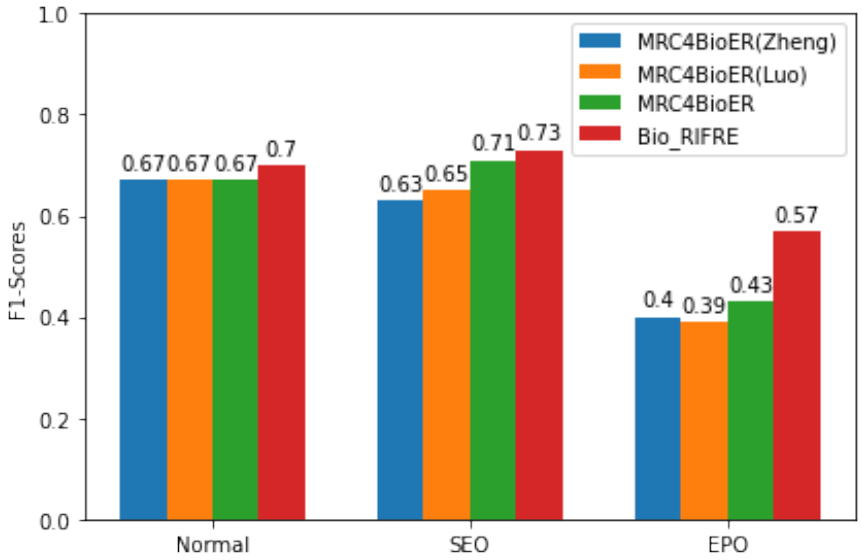


Figure 6.4: F1-score in Extracting Various Sequences

proficiency in the multi-labeling task, where it assigns different relation classes to triples with the same chemical-gene named entities.

6.3.2 Different Numbers of Triples per Sequence

To evaluate the robustness of Bio-RIFRE compared to various baseline models, the model was tested on a CPI test set containing different numbers of triples per sequence. For instance, $N = 3$ indicates that the model was evaluated on sequences with exactly three triples. Figure 6.5 presents the results, demonstrating that the Bio-RIFRE model maintains robustness as the number of triples per sequence increases, as evidenced by the model's F1-score remaining relatively stable. Additionally, when sequences contain more than five triples, Bio-RIFRE outperforms other baseline models in identifying entities and their relationships.

6.3.3 More Relation Classes

Table 6.3 provides a detailed overview of the Bio-RIFRE model's performance on the DrugProt dataset. As this dataset is relatively new, there is limited research available for comparison. However, the results clearly indicate that the Bio-RIFRE model performs admirably, even with the thirteen different relation classes present in the DrugProt dataset. Furthermore, the model's performance remains consistent and does not significantly change as the number of triples per sequence increases, further highlighting its robustness.

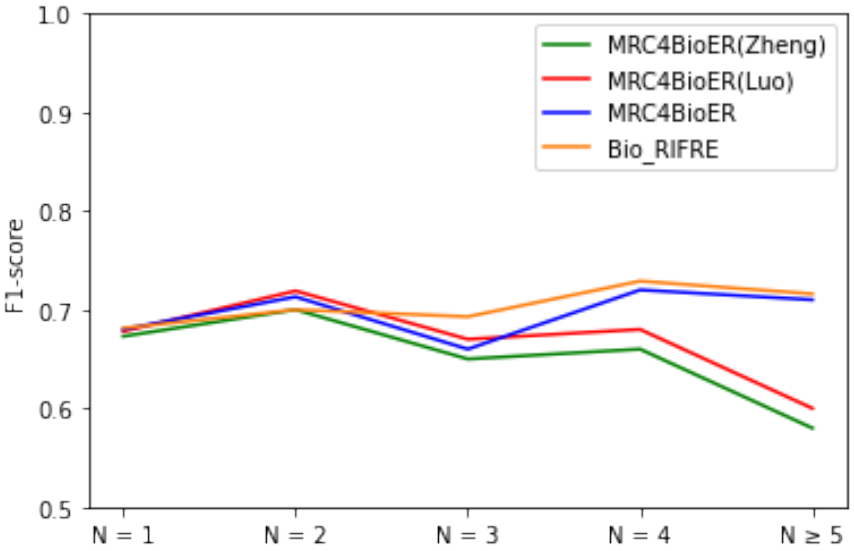


Figure 6.5: F1-score for Extracting Different Numbers of Triples (N) from Sequences

Table 6.3: Evaluation of Bio-RIFRE on DrugProt [PRKL18]

Number of Triples	**F1**	**R**	**P**
N = 1	0.727	0.811	0.6595
N = 2	0.716	0.758	0.752
N = 3	0.7163	0.701	0.732
N = 4	0.7281	0.6985	0.7602
N ≥ 5	0.7144	0.6624	0.7752

Chapter 7

Evaluation of Large Language Models in Clinical Data summarization

7.1 Introduction

In NLP area, the ability of LLMs to generate summaries has indicated significant advancements in automated content extraction and understanding. Summarization techniques broadly fall into two categories: extractive and abstractive. Extractive methods involve selecting and rearranging existing sentences or phrases from the source text to form a concise summary, while abstractive methods generate novel sentences that convey the core information in a more condensed and coherent manner.

The proliferation of abstractive summarization, mainly driven by deep learning models like transformer-based architectures, introduces the challenge of effectively evaluating the quality and reliability of these generated summaries. Unlike extractive methods, where overlap metrics like ROUGE are widely used, abstractive summarization necessitates metrics that assess semantic coherence, informativeness, and syntactic correctness in generated text [Lin04].

The development of robust evaluation metrics tailored to abstractive summarization is crucial for several reasons. Firstly, it ensures the reliability and comparability of summary generation models across different datasets and tasks. Secondly, it facilitates the iterative improvement of these models by providing actionable feedback on their outputs. Moreover, effective metrics can help researchers and practitioners identify strengths and weaknesses in current approaches, guiding future advancements in NLP and summarization technology.

For instance, recent studies [CY04] [ZKW+19] have proposed metrics such as BERTScore, which

leverages contextual representation from transformer models to measure the similarity between generated and reference summaries more effectively than traditional n-gram overlap measures or other metrics relying solely on surface-form overlaps like BLEU or METEOR [Gra15]. BERTScore computes the cosine similarity between the embeddings of tokens, enhancing the ability to recognize these semantic relationships. Compared to existing metrics, this method shows a higher correlation with human judgment across tasks such as machine translation and image captioning. As another example, research conducted by [CCP22] introduces InfoLM, an innovative metric developed to evaluate the quality of text summarization and data-to-text generation. InfoLM leverages untrained metrics utilizing a pre-trained masked language model (PMLM) to evaluate texts by comparing discrete probability distributions of tokens generated from the candidate and reference texts. InfoLM is distinguished by its use of information theory measures, which allows it to adapt the metric to different evaluation criteria without extensive retraining.

These advancements underscore the evolution towards more sophisticated evaluation frameworks suited to the complexities of abstractive summarization tasks. While extractive summarization benefits from established evaluation metrics, the maturation of abstractive methods demands innovative approaches reflecting content synthesis's nature. To bridge the gap in summarization metrics, we proposed a new metric that employs a novel evaluation approach using a bipartite knowledge graph. This method examines the named entity class types and their placement within the summaries by tracing their corresponding sentence numbers. This approach ensures that the summaries not only capture essential information but also faithfully maintain the logical and sequential flow of the original narrative. Our assessment strategy enhances our understanding of the summarization process by guaranteeing that the generated summaries preserve linguistic coherence and semantic integrity.

7.1.1 Graph-based Summary Quality Metric

In our enhanced summarization evaluation method, we utilize bipartite knowledge graphs to analyze the structure and order of named entities within summaries. Below is a detailed breakdown of the process:

- **Node Classification:** Our bipartite graphs consist of two distinct types of nodes. The first type represents named entity classes, categorizing entities according to their semantic roles such as 'Person', 'Location', or 'Organization'. The second type corresponds to the sentence

numbers within the text, indexing each sentence where entities appear.

- **Entity Extraction Process:** We perform named entity recognition in two passes to ensure accuracy and completeness. The first pass scans the text to identify preliminary entity mentions and their classes. In the second pass, we refine these results, correcting misclassifications and confirming each entity's boundaries.

- **Selection of Significant Entities:** After identifying the entities and their classifications, we focus on the top 20 percentile of entities based on their frequency and relevance to the text's central themes. For each of these entities, we record the sentence numbers where they appear.

- **Graph Construction:** We then construct a bipartite graph for both the original text and its summary. In this graph, we create edges between nodes representing entity classes and sentence numbers. An edge is formed between an entity class node and a sentence number node if that particular entity class is present in the corresponding sentence. Figure 7.1 indicates the graph overview.

- **Graph Comparison for Evaluation:** To assess the summary's quality, we compare the bipartite graph of the original text with that of the generated summary. This comparison focuses on the similarity of the graphs in terms of structure and entity distribution. The key aspect we measure is whether the summary preserves the order and structural context of the named entity classes as they appear in the original text.

- **Preservation of Order and Structure:** By analyzing the correspondence between the graphs' configurations, we can quantitatively evaluate how well the summary maintains the logical and sequential flow of entities. This provides a metric for assessing the preservation of narrative structure and information integrity within the summarized content.

This methodology allows for a precise and structured evaluation of text summarization algorithms, focusing particularly on their ability to maintain coherence and fidelity to the original text's structure. Through this approach, we gain deeper insights into the summarization process's effectiveness, ensuring that the essential elements of the original narrative are accurately reflected in the summaries.

7.2 Information-Theoretic Formulation of Bipartite Graphs Metric

We redefine the elements in the context of a bipartite graph where:

- x_i represents a pair indicating a named entity class and its sentence number in the original text T.
- y_j similarly represents pairs of named entity classes and sentence numbers in the summary S.

Entropy and Mutual Information

The entropy calculations now focus on the presence of specific entity-sentence pairs rather than individual words or sentences, capturing the structured information:

$$H(T) = -\sum_i p(x_i) \log p(x_i) \tag{7.1}$$

$$H(S) = -\sum_j p(y_j) \log p(y_j) \tag{7.2}$$

Here, $p(x_i)$ and $p(y_j)$ are the probabilities of each entity-sentence pair occurring in the original text and the summary, respectively, estimated based on their frequencies.

Mutual Information Calculation

Mutual information measures how much information about the entity-sentence pairs in the original text is preserved in the summary:

$$I(T;S) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{7.3}$$

Here, $p(x,y)$ represents the joint probability that an entity-sentence pair x from the original text and an entity-sentence pair y from the summary are the same.

Information Loss Evaluation

The information loss, particularly in terms of the structure and semantic integrity, is assessed by:

$$\text{Information Loss} = H(T) - I(T;S) \tag{7.4}$$

This calculation reveals how much entity-related structural information is lost in the summarization process.

By redefining these calculations to focus on bipartite graphs, this approach evaluates not only the preservation of content but also how well the summary maintains the logical structure and order of the original text, as represented by the entity-sentence relationships. This methodology is particularly useful where the structural integrity of information is as crucial as the content itself.

Case Study: Enhancing Clinical Trial Summaries on ClinicalTrials.gov

In clinical research, the efficient summarization of trial details is crucial in sharing information concisely with researchers, practitioners, and the public. On platforms such as ClinicalTrials.gov, each clinical trial is accompanied by a "Brief Summary" data field, which summarizes the trial's objectives, methodologies, and other key elements. However, not all entries on ClinicalTrials.gov are complete, with some trials lacking a concise brief description. Recognizing this gap, the study cited in [GKS+19] employs extractive summarization techniques to generate these brief summaries from a more extensive "Detailed Description" data field. This approach not only aims to provide greater information in a more summarized form but also addresses the issue of missing brief summaries. The main motivation behind this task is to enhance the accessibility and usability of clinical trial data by condensing detailed narratives into informative, easy-to-digest summaries.

In an effort to improve the summarization of clinical trials listed on ClinicalTrials.gov, we have employed the BART (Bidirectional and Auto-Regressive Transformers) model, a state-of-the-art method known for its effectiveness in generating coherent and contextually relevant text summaries. Recognizing that the structure of the summary is crucial for its utility and readability, we have implemented a meticulous evaluation process for the summaries generated by the BART model. This evaluation involves several key steps: First, we perform an initial quality check to assess the factual accuracy and relevance of the content in the generated summaries compared to the original detailed descriptions. This ensures that no critical information is misrepresented or omitted in the summarization process. Second, we analyze the coherence and flow of the summaries, examining how well the BART model maintains logical sequencing and connectivity between ideas, which is vital for the reader's comprehension.

Bipartite Graph Generation

In our study, we structured the nodes of our bipartite graphs into two primary classes derived from the data available on ClinicalTrials.gov. The entities we analyzed are classified into ten specific types, collectively labeled as $E = \{e_1, e_2, \ldots, e_{10}\}$. These types encompass essential components of clinical trials, including 'disease,' 'medical condition,' 'drug,' 'device,' 'dose or measurements,' 'clinical trial phase,' 'population,' 'time,' 'medical procedure,' and 'biomarker.'

- **First phase**: Employing the *llama 2-70 billion* model parameters [TMS+23] facilitated by vLLM [KLZ+23], we extracted named entities from 1000 randomly chosen brief summaries. This phase prioritized the extraction based on the frequency of occurrence of the entities.

- **Second phase**: Building on the identified key entity types, we then processed summaries generated by the BART model for this dataset. We broke down each summary into individual sentences for further analysis using *llama 2-70 billion*. A tailored prompt (Listing 7.1) was applied to check for the presence or absence of each entity type across sentences. The result for each document d_i was composed of two binary matrices—one representing the original text T as the ground truth (matrix size: $m_i \times 10$), and the other for the BART-generated summary (matrix size: $n_i \times 10$), with m_i and n_i indicating the number of sentences in the respective summaries of document i.

Each document's relationship between entity types and sentence numbers was mapped using binary matrices, forming the basis for bipartite graph construction. We assigned G_{1i} to represent the graph of the original text and G_{2i} for the graph of the generated summary.

Our analysis concentrated on documents in which the original and the BART-generated summaries contained an equal number of sentences ($n = m$). We employed the Jaccard similarity metric (outlined in Formula 7.6) to evaluate these graphs over $D = 1000$ trials. The resulting average Jaccard similarity of 0.71 indicates that about 71% of the entity types and their sequential placement in the text are preserved in the summaries [IH98]. This finding underscores the effectiveness of our summarization method in maintaining essential details in the structured format of clinical trial descriptions.

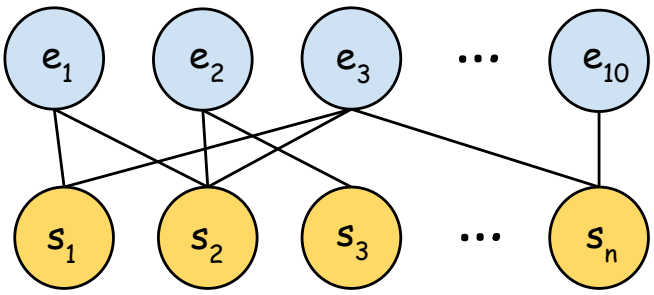


Figure 7.1: Overview of Bipartite Graph

Query 7.1: Prompt for Entity Types Extraction

```
# List of entity types for query
entity_types = [
    "disease", "medical condition", "drug", "device",
    "dose or measurements", "clinical trial phase", "population",
    "time", "medical procedure", "biomarker"
]

# Template for constructing the question
question_template = """Does the following sentence include any named entities of type '{entity_type}'?:
    ```{sentence}```"""

# Example of how to format the question with a specific entity type and sentence
question = question_template.format(entity_type=entity_types[i], sentence=sentence)
```

Proposition 7.2.1 *Consider that the count of sentences in both the original and generated summaries is equal, represented by the equality $n = m$. In this scenario, $E(G)$ symbolizes the collection of all edges within the graph G.*

$$\frac{\sum_{i=1}^{D} J(G_{i1}, G_{i2})}{D} \quad \text{for } 1 \leq i \leq D \tag{7.5}$$

$$J(G_1, G_2) = \frac{|E(G_1) \cap E(G_2)|}{|E(G_1) \cup E(G_2)|} \tag{7.6}$$
```

Chapter 8

Conclusion

In this dissertation, we present a series of innovative computational models and systems designed to enhance the analysis and interpretation of biomedical data and clinical trials. Each contribution not only stands as an achievement in its field but also sets the stage for further research and application development.

8.1 Overview of Work

This concluding chapter summarizes the key contributions and innovations presented throughout our work, emphasizing their impact on medical informatics and clinical research. The provided models not only address current challenges but also start numerous possibilities for future exploration and application. Here, we reflect on our achievements and cast a forward-looking view on the potential avenues for future work, aiming to expand further the capabilities and applications of our developed technologies. This discussion sets the stage for ongoing enhancements and exploring new frontiers in medical and pharmaceutical sciences.

8.1.1 Tri-AL: An Open-source System for Tracking Clinical Trials

Our development of Tri-AL marks a significant advancement in the monitoring and analyzing clinical trials. This open-source system, capable of tracking clinical trials over time on ClinicalTrials.gov, integrates a module for analyzing the race and ethnicity of participants, which is crucial for ensuring diversity and equity in clinical research. The system's interactive interface and robust data visualization tools greatly facilitate the exploration of data features, allowing researchers to uncover trends and disparities in trial participation. Future enhancements to Tri-AL could include

real-time data tracking and integration with additional biomedical databases, which would broaden its applicability and utility. Furthermore, incorporating machine learning models to predict trends in clinical trial data based on historical patterns could enhance its predictive capabilities, providing early insights into emerging research areas or potential biases in trial recruitment.

8.1.2 Predicting Drug Mechanisms of Action

The supervised predictive model we developed for determining drug mechanisms of action utilizes advanced Ml and DL language models to enhance our understanding of drug functions and their interactions. This model not only accelerates drug development processes but also holds the potential to revolutionize therapeutic strategies by predicting adverse drug reactions and optimizing drug combinations for personalized medicine. Expanding this model to cover more drug categories and incorporating more diverse datasets, including real-world patient data, could improve its accuracy and applicability in real-world scenarios. Additionally, integrating this model with electronic health records (EHRs) could facilitate personalized drug recommendations, enhancing patient outcomes and treatment efficiencies.

8.1.3 Heterogeneous Graph Neural Network for Gene-Chemical Entity Relation Extraction

Implementing a heterogeneous Graph Neural Network (GNN) for extracting gene-chemical relationships augments word representations using message-passing techniques, enhancing the accuracy of identifying gene-chemical named entities and their relationships. Future work could extend this model to encompass additional entity types and deeper relational contexts, incorporating more complex layers of interaction, such as gene-environment interactions, which are crucial for understanding complex diseases. Exploring the application of this GNN model to other areas of bioinformatics, such as protein-protein interactions and cellular pathway analysis, could also yield significant benefits. This could involve developing dynamic GNNs that adapt to new discoveries in genomics and proteomics, providing a continually evolving tool for biomedical research.

8.1.4 Bipartite Graph Model for Evaluating Summarization Performance

The bipartite graph model proposed for evaluating the performance of large language models in summarizing clinical trials provides a robust framework to assess the accuracy and effectiveness of

automated summarization tools. This model's ability to evaluate summaries' coherence and context preservation can significantly contribute to clinical knowledge management systems, in which precise and reliable summarization of vast amounts of literature is critical. Future research could explore incorporating reinforcement learning to dynamically adjust summarization strategies based on feedback, enhancing the model's adaptability and accuracy. Testing this model across different types of medical literature and clinical guidelines could validate its versatility and effectiveness in various medical contexts, potentially leading to its integration into clinical decision support systems.

Each of these projects demonstrates our commitment to pushing the boundaries of current methodologies and highlights the potential for significant advances in the field of medical informatics. As we continue to refine these models and systems, they will play a crucial role in enhancing the accuracy and efficiency of biomedical research and healthcare delivery.

Bibliography

[AAA+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[ABCED23] Scott Askin, Denis Burkhalter, Gilda Calado, and Samar El Dakrouni. Artificial intelligence applied to clinical trials: opportunities and challenges. *Health and Technology*, 13(2):203–213, 2023.

[AGR+18] Abdulrahman Aalabdulsalam, Jennifer Garvin, Andrew Redd, Marjorie Carter, Carol Sweeny, and Stephane Meystre. Automated extraction and classification of cancer stage mentions fromunstructured text fields in a central cancer registry. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2017:16–25, 05 2018.

[ALc24] Tri-al: Visual clinical trials, 2024.

[BK18] Rahul Baboota and Harleen Kaur. Predictive analysis and modelling football results using machine learning approach for english premier league. *International Journal of Forecasting*, 35, 03 2018.

[BLC19] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. In *EMNLP*, 2019.

[BMAZ+20] Mette Brøgger-Mikkelsen, Zarqa Ali, John R Zibert, Anders Daniel Andersen, Simon Francis Thomsen, et al. Online patient recruitment in clinical trials: systematic review and meta-analysis. *Journal of medical Internet research*, 22(11):e22179, 2020.

[Bod04] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270, 2004.

[BSM+23] Ryan CL Brewster, Jecca R Steinberg, Christopher J Magnani, Jasmyne Jackson, Bonnie O Wong, Nishma Valikodath, Justin MacDonald, Anna Li, Paula Marsland, Steven N Goodman, et al. Race and ethnicity reporting and representation in pediatric clinical trials. *Pediatrics*, 151(4):e2022058552, 2023.

[BTC+12] Mary Regina Boland, Samson W Tu, Simona Carini, Ida Sim, and Chunhua Weng. Elixr-time: a temporal knowledge representation for clinical research eligibility criteria. *AMIA summits on translational science proceedings*, 2012:71, 2012.

[Car22] Benjamin Gregory Carlisle. Analysis of clinical trial registry entry histories using the novel r package cthist. *medRxiv*, 2022.

[CBHK02] N. Chawla, K. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *ArXiv*, abs/1106.1813, 2002.

[CCP22] Pierre Jean A Colombo, Chloé Clavel, and Pablo Piantanida. Infolm: A new metric to evaluate summarization & data2text generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 10554–10562, 2022.

[CJLD+14] Moon S Chen Jr, Primo N Lara, Julie HT Dang, Debora A Paterniti, and Karen Kelly. Twenty years post-nih revitalization act: Enhancing minority participation in clinical trials (empact): Laying the groundwork for improving minority clinical trial accrual: Renewing the case for enhancing minority participation in cancer clinical trials. *Cancer*, 120:1091–1096, 2014.

[cli24] clinicaltrials.gov. clinicaltrials.gov, 2024. Accessed: 2024-06-17.

[CLN+22] Jeffrey Cummings, Garam Lee, Pouyan Nahed, Mina Esmail Zadeh Nojoo Kambar, Kate Zhong, Jorge Fonseca, and Kazem Taghva. Alzheimer's disease drug development pipeline: 2022. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 8(1):e12295, 2022.

[CLR+19] Jeffrey Cummings, Garam Lee, Aaron Ritter, Marwan Sabbagh, and Kate Zhong. Alzheimer's disease drug development pipeline: 2019. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 5:272–293, 07 2019.

[CLR+20] Jeffrey Cummings, Garam Lee, Aaron Ritter, Marwan Sabbagh, and Kate Zhong. Alzheimer's disease drug development pipeline: 2020. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 6, 07 2020.

[CLRZ18] Jeffrey Cummings, Garam Lee, Aaron Ritter, and Kate Zhong. Alzheimer's disease drug development pipeline: 2018. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 4:195–214, 2018.

[CMZ14] Jeffrey Cummings, Travis Morstorf, and Kate Zhong. Alzheimer's disease drug-development pipeline: Few candidates, frequent failures. *Alzheimer's research therapy*, 6:37, 07 2014.

[Con22] Text Retrieval Conference. 2021 clinical trials track, 2022.

[CR11] Yee Seng Chan and Dan Roth. Exploiting syntactico-semantic structures for relation extraction. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, editors, *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 551–560, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[CRH+14] Vivien Chen, Bernardo Ruiz, Mei-Chin Hsieh, Xiao-Cheng Wu, Lynn Ries, and Denise Lewis. Analysis of stage and clinical/prognostic factors for lung cancer from seer registries: Ajcc staging and collaborative stage data collection system. *Cancer*, 120 Suppl 23:3781–92, 12 2014.

[CY04] Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004.

[DAN18] Ahmed Wahib Dahouk and Samy S Abu-Naser. A proposed knowledge based system for desktop pc troubleshooting. 2018.

[DBC+16] Louise Deléger, Robert Bossy, Estelle Chaix, Mouhamadou Ba, Arnaud Ferré, Philippe Bessieres, and Claire Nédellec. Overview of the bacteria biotope task at bionlp shared task 2016. In *Proceedings of the 4th BioNLP shared task workshop*, pages 12–22, 2016.

[DBR19] Surabhi Datta, Elmer V Bernstam, and Kirk Roberts. A frame semantic overview of nlp-based information extraction for cancer-related ehr notes. *Journal of biomedical informatics*, 100:103301, 2019.

[DCKH+19] Alan Davies, Marisa Cunha, Kamilla Kopec-Harding, Paul Metcalfe, James Weatherall, and Caroline Jay. Biomarker data visualisation for decision making in clinical trials. *International journal of medical informatics*, 132:104008, 2019.

[DCLT18a] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[DCLT18b] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[DME+19] Ibtesam M Dheir, Alaa Soliman Abu Mettleq, Abeer A Elsharif, Mohammed N Abu Al-Qumboz, and Samy S Abu-Naser. Knowledge based system for diabetes diagnosis using sl5 object. 2019.

[EEH22] Mina Esmail Zadeh Nojoo Kambar, Armin Esmaeilzadeh, and Maryam Heidari. A survey on deep learning techniques for joint named entities and relation extraction. In *2022 IEEE World AI IoT Congress (AIIoT) (IEEE AIIOT 2022)*, virtual, USA, June 2022.

[ET21] Armin Esmaeilzadeh and Kazem Taghva. Text classification using neural network language model (nnlm) and bert: An empirical comparison. In *Proceedings of SAI Intelligent Systems Conference*, pages 175–189. Springer, 2021.

[FNTW21] Kevin M Fain, Julianne T Nelson, Tony Tse, and Rebecca J Williams. Race and ethnicity reporting for clinical trials in clinicaltrials. gov and publications. *Contemporary Clinical Trials*, 101:106237, 2021.

[GGS+21] Allison Gates, Michelle Gates, Shannon Sim, Sarah A Elliott, Jennifer Pillay, and Lisa Hartling. Creating efficiencies in the extraction of data from randomized trials: a prospective evaluation of a machine learning and text mining tool. *BMC medical research methodology*, 21:1–12, 2021.

[GGS23] Aldren Gonzales, Guruprabha Guruswamy, and Scott R Smith. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082, 2023.

[GKS+19] Christian Gulden, Melanie Kirchner, Christina Schüttler, Marc Hinderer, Marvin Kampf, Hans-Ulrich Prokosch, and Dennis Toddenroth. Extractive summarization of clinical trial descriptions. *International journal of medical informatics*, 129:114–121, 2019.

[GLCI+18] Santiago Guerrero, Andrés López-Cortés, Alberto Indacochea, Jennyfer M García-Cárdenas, Ana Karina Zambrano, Alejandro Cabrera-Andrade, Patricia Guevara-Ramírez, Diana Abigail González, Paola E Leone, and César Paz-y Miño. Analysis of racial/ethnic representation in select basic and applied cancer research studies. *Scientific reports*, 8(1):1–8, 2018.

[Gra15] Yvette Graham. Re-evaluating automatic summarization with bleu and 192 shades of rouge. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 128–137, 2015.

[HJ20] Maryam Heidari and James H Jones. Using bert to extract topic-independent sentiment features for social media bot detection. In *2020 11th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, pages 0542–0547, 2020.

[HKC+22] Nora Hutchinson, Katarzyna Klas, Benjamin G Carlisle, Jonathan Kimmelman, and Marcin Waligora. How informative were early sars-cov-2 treatment and prevention trials? a longitudinal cohort analysis of trials registered on clinicaltrials. gov. *PloS one*, 17(1):e0262114, 2022.

[HLW16] Tianyong Hao, Hongfang Liu, and Chunhua Weng. Valx: a system for extracting and structuring numeric lab test comparison statements from text. *Methods of information in medicine*, 55(03):266–275, 2016.

[HO20] Udo Hahn and Michel Oleynik. Medical information extraction in the age of deep learning. *Yearbook of medical informatics*, 29(01):208–220, 2020.

[HR20] Maryam Heidari and Setareh Rafatirad. Semantic convolutional neural network model for safe business investment by using bert. In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–6, 2020.

[HZSBMD13] María Herrero-Zazo, Isabel Segura-Bedmar, Paloma Martínez, and Thierry Declerck. The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *Journal of biomedical informatics*, 46(5):914–920, 2013.

[IH98] GI Ivchenko and SA Honov. On the jaccard similarity test. *Journal of Mathematical Sciences*, 88:789–794, 1998.

[Int24] Clinical Trials Transformation Intiative. Aact database schema. https://aact.ctti-clinicaltrials.org/schema, 2024.

[JCL+20] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[JWF+23] Qiao Jin, Zifeng Wang, Charalampos S Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun, and Zhiyong Lu. Matching patients to clinical trials with large language models. *ArXiv*, 2023.

[KKB+16] Jens Kringelum, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I Oprea, and Olivier Taboureau. Chemprot-3.0: a global chemical biology diseases mapping. *Database*, 2016:bav123, 2016.

[KLZ+23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

[KNC+22] Mina Esmail Zadeh Nojoo Kambar, Pouyan Nahed, Jorge Ramón Fonseca Cacho, Garam Lee, Jeffrey Cummings, and Kazem Taghva. Clinical text classification of alzheimer's drugs' mechanism of action. In *Proceedings of Sixth International Congress on Information and Communication Technology*, pages 513–521. Springer, 2022.

[KRA+17] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, Martın Pérez Pérez, Jesús Santamaría, Gael Pérez Rodríguez, Georgios Tsatsaronis, Ander Intxaurrondo, José Antonio López, Umesh Nandal, et al. Overview of the biocreative vi chemical-protein interaction track. In *Proceedings of the sixth BioCreative challenge evaluation workshop*, volume 1, pages 141–146, 2017.

[KRL+15] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(1):1–17, 2015.

[KSMB21] Wan Yee Kong, Hamidreza Saber, Rohit Marawar, and Maysaa Merhi Basha. Racial and ethnic trends in antiseizure medications trial enrolment: A systematic review using clinicaltrials. gov. *Epilepsy Research*, 173:106613, 2021.

[Kum17] Shantanu Kumar. A survey of deep learning methods for relation extraction. *arXiv preprint arXiv:1705.03645*, 2017.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LFM+19] Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. A unified mrc framework for named entity recognition. *arXiv preprint arXiv:1910.11476*, 2019.

[LHL+24] Rogier Landman, Sean P Healey, Vittorio Loprinzo, Ulrike Kochendoerfer, Angela Russell Winnier, Peter V Henstock, Wenyi Lin, Aqiu Chen, Arthi Rajendran, Sushant Penshanwar, et al. Using large language models for safety-related table summarization in clinical study reports. *JAMIA open*, 7(2):ooae043, 2024.

[LHS23] M Leela, K Helenprabha, and L Sharmila. Prediction and classification of alzheimer disease categories using integrated deep transfer learning approach. *Measurement: Sensors*, 27:100749, 2023.

[Lin04] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[LPH+24] Kyeryoung Lee, Hunki Paek, Liang-Chin Huang, C Beau Hilton, Surabhi Datta, Josh Higashi, Nneka Ofoegbu, Jingqi Wang, Samuel M Rubinstein, Andrew J Cowan, et al. Seetrials: Leveraging large language models for safety and efficacy extraction in oncology clinical trials. *medRxiv*, pages 2024–01, 2024.

[LYC+20] Ling Luo, Zhihao Yang, Mingyu Cao, Lei Wang, Yin Zhang, and Hongfei Lin. A neural network-based joint learning approach for biomedical entity and relation extraction from biomedical literature. *Journal of biomedical informatics*, 103:103384, 2020.

[LYK+19] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 09 2019.

[LYK+20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[MB16] Makoto Miwa and Mohit Bansal. End-to-end relation extraction using LSTMs on sequences and tree structures. In Katrin Erk and Noah A. Smith, editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1105–1116, Berlin, Germany, August 2016. Association for Computational Linguistics.

[MBR+23] Kevin McFarthing, Susan Buff, Gary Rafaloff, Brian Fiske, Leah Mursaleen, Rosie Fuest, Richard K Wyse, and Simon RW Stott. Parkinson's disease drug therapies in the clinical trial pipeline: 2023 update. *Journal of Parkinson's Disease*, 13(4):427–439, 2023.

[McK10] Wes McKinney. Data structures for statistical computing in python. pages 56–61, 01 2010.

[MCT22] Anthony Muchai Manyara, Oriana Ciani, and Rod S Taylor. A call for better reporting of trials using surrogate primary endpoints. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 8(1), 2022.

[MEMR+24] Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper van Langen, and Christoph Meier. From rags to riches: Using large language models to write documents for clinical trials. *arXiv preprint arXiv:2402.16406*, 2024.

[MML+21] Antonio Miranda, Farrokh Mehryary, Jouni Luoma, Sampo Pyysalo, Alfonso Valencia, and Martin Krallinger. Overview of drugprot biocreative vii track: quality evaluation and large scale text mining of drug-gene/protein relations. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, 2021.

[MMMG23] Cecilia Monge, J Alberto Maldonado, Katherine A McGlynn, and Tim F Greten. Hispanic individuals are underrepresented in phase iii clinical trials for advanced liver cancer in the united states. *Journal of Hepatocellular Carcinoma*, pages 1223–1235, 2023.

[MPL+14] Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307, 2014.

[New22] William Newton. Als: drug trial results to watch in 2022, 2022.

[NIR+23] David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. *International Journal of Medical Informatics*, page 105122, 2023.

[NKBA19] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy, August 2019. Association for Computational Linguistics.

[NKC+22] Pouyan Nahed, Mina Esmail Zadeh Nojoo Kambar, Jorge Ramón Fonseca Cacho, Garam Lee, Jeffrey Cummings, and Kazem Taghva. A recommendation model for predicting alzheimer's drugs' mechanism of action. In *Intelligent Sustainable Systems - Selected Papers of WorldS4*. Springer, 2022.

[NMPO19] Bassel Nazha, Manoj Mishra, Rebecca Pentz, and Taofeek K Owonikoko. Enrollment of racial minorities in clinical trials: old problem assumes new urgency in the age of immunotherapy. *American Society of Clinical Oncology Educational Book*, 39:3–10, 2019.

[oCO24] Jornal of Clinical Onthology. Jco meeting abstracts, 2024. Accessed: 2024-06-17.

[oH24] American Society of Hematology. American society of hematology, 2024. Accessed: 2024-06-17.

[Par22] Cure Parkinson's. The international linked clinical trials (ilct) programme, 2022.

[PJC23] Kevin Pei, Ishan Jindal, and Kevin Chang. Abstractive open information extraction. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6146–6158, Singapore, December 2023. Association for Computational Linguistics.

[PPB17] Sachin Pawar, Girish K Palshikar, and Pushpak Bhattacharyya. Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*, 2017.

[PPH+20] Krishna Pundi, Alexander C Perino, Robert A Harrington, Harlan M Krumholz, and Mintu P Turakhia. Characteristics and strength of evidence of covid-19 studies registered on clinicaltrials. gov. *JAMA internal medicine*, 180(10):1398–1400, 2020.

[PPS+23] Dipti Pawar, Shraddha Phansalkar, Abhishek Sharma, Gouri Kumar Sahu, Chun Kit Ang, and Wei Hong Lim. Survey on the biomedical text summarization techniques with an emphasis on databases, techniques, semantic approaches, classification techniques, and similarity measures. *Sustainability*, 15(5):4216, 2023.

[PRKL18] Yifan Peng, Anthony Rios, Ramakanth Kavuluru, and Zhiyong Lu. Extracting chemical–protein relations with ensembles of SVM and deep learning models. *Database*, 2018, 07 2018. bay073.

[PVG+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[RS10] Radim Rehůřek and Petr Sojka. Software framework for topic modelling with large corpora. pages 45–50, 05 2010.

[RWH+17] Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th international conference on world wide web*, pages 1015–1024, 2017.

[SASK23] Rajesh Kumar Singh, Saurabh Agrawal, Abhishek Sahu, and Yigit Kazancoglu. Strategic issues of big data analytics applications for managing health-care sector: a systematic literature review and future research agenda. *The TQM Journal*, 35(1):262–291, 2023.

[SCH23] Qianmin Su, Gaoyi Cheng, and Jihan Huang. A review of research on eligibility criteria for clinical trials. *Clinical and experimental medicine*, 23(6):1867–1879, 2023.

[SG14] Abeed Sarker and Graciela Gonzalez. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. *Journal of Biomedical Informatics*, 53, 11 2014.

[SMD+19] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani, et al. Natural language processing of clinical notes on chronic diseases: systematic review. *JMIR medical informatics*, 7(2):e12239, 2019.

[SSM22] Robert A Smith, Paul P Schneider, and Wael Mohammed. Living hta: automating health technology assessment with r [version 1; peer review: 1 approved with reservations]. *Wellcome Open Research*, 7, 2022.

[SYW+22] Cong Sun, Zhihao Yang, Lei Wang, Yin Zhang, Hongfei Lin, and Jian Wang. Mrc4bioer: Joint extraction of biomedical entities and relations in the machine reading comprehension framework. *Journal of biomedical informatics*, 125:103956, 2022.

[TMS+23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[TPC+11] Samson W Tu, Mor Peleg, Simona Carini, Michael Bobak, Jessica Ross, Daniel Rubin, and Ida Sim. A practical method for transforming free-text eligibility criteria into computable criteria. *Journal of biomedical informatics*, 44(2):239–250, 2011.

[TSM+20] Yitong Tseo, MI Salkola, Ahmed Mohamed, Anuj Kumar, and Freddy Abnousi. Information extraction of clinical trial eligibility criteria. *arXiv preprint arXiv:2006.07296*, 2020.

[TSW+22] Brandon E Turner, Jecca R Steinberg, Brannon T Weeks, Fatima Rodriguez, and Mark R Cullen. Race/ethnicity reporting and representation in us clinical trials: A cohort study. *The Lancet Regional Health-Americas*, page 100252, 2022.

[TTE+23] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

[VCC+18] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. accepted as poster.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[VVVUB+23] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerova, et al. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square*, 2023.

[WCH18] Neha Warikoo, Yung-Chun Chang, and Wen-Lian Hsu. LPTK: a linguistic pattern-aware dependency tree kernel approach for the BioCreative VI CHEMPROT task. *Database*, 2018, 10 2018. bay108.

[WFHST20] Rueben C Warren, Lachlan Forrow, David Augustin Hodge Sr, and Robert D Truog. Trustworthiness before trust—covid-19 vaccine trials and the black community. *New England Journal of Medicine*, 383(22):e121, 2020.

[WJS+19] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *The world wide web conference*, pages 2022–2032, 2019.

[WSW+20] Zhepei Wei, Jianlin Su, Yue Wang, Yuan Tian, and Yi Chang. A novel cascade binary tagging framework for relational triple extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1476–1488, 2020.

[WWL+11] Chunhua Weng, Xiaoying Wu, Zhihui Luo, Mary Regina Boland, Dimitri Theodoratos, and Stephen B Johnson. Elixr: an approach to eligibility criteria extraction and representation. *Journal of the American Medical Informatics Association*, 18(Supplement_1):i116–i124, 2011.

[XVL+23] Hong Xiao, Riha Vaidya, Fang Liu, Ximing Chang, Xiaoqian Xia, and Joseph M Unger. Sex, racial, and ethnic representation in covid-19 clinical trials: a systematic review and meta-analysis. *JAMA Internal Medicine*, 183(1):50–60, 2023.

[YRT+19] Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, et al. Criteria2query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305, 2019.

[ZC20] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. *arXiv preprint arXiv:2010.12812*, 2020.

[ZKW+19] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[ZLC+19a] Qi Zhang, Chao Liu, Ying Chi, Xuansong Xie, and Xiansheng Hua. A Multi-Task Learning Framework for Extracting Bacteria Biotope Information. *BioNLP-OST@EMNLP-IJNCLP 2019 - Proceedings of the 5th Workshop on BioNLP Open Shared Tasks*, pages 105–109, 2019.

[ZLC+19b] Qi Zhang, Chao Liu, Ying Chi, Xuansong Xie, and Xiansheng Hua. A multi-task learning framework for extracting bacteria biotope information. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 105–109, 2019.

[ZSLW20] Sendong Zhao, Chang Su, Zhiyong Lu, and Fei Wang. Recent advances in biomedical literature mining. *Briefings in Bioinformatics*, 22(3), 05 2020. bbaa057.

[ZTW+11] Deborah A Zarin, Tony Tse, Rebecca J Williams, Robert M Califf, and Nicholas C Ide. The clinicaltrials. gov results database—update and key issues. *New England Journal of Medicine*, 364(9):852–860, 2011.

[ZTWC16] Deborah A Zarin, Tony Tse, Rebecca J Williams, and Sarah Carr. Trial reporting in clinicaltrials. gov—the final rule. *New England Journal of Medicine*, 375(20):1998–2004, 2016.

[ZWB+17] Suncong Zheng, Feng Wang, Hongyun Bao, Yuexing Hao, Peng Zhou, and Bo Xu. Joint extraction of entities and relations based on a novel tagging scheme. *arXiv preprint arXiv:1706.05075*, 2017.

[ZXC+21] Kang Zhao, Hua Xu, Yue Cheng, Xiaoteng Li, and Kai Gao. Representation iterative fusion based on heterogeneous graph neural network for joint entity and relation extraction. *Knowledge-Based Systems*, 219:106888, 2021.

[ZZ21] Mei Zuo and Yang Zhang. A span-based joint model for extracting entities and relations of bacteria biotopes. *Bioinformatics*, 2021.

[ZZ22] Mei Zuo and Yang Zhang. A span-based joint model for extracting entities and relations of bacteria biotopes. *Bioinformatics*, 38(1):220–227, 2022.

[ZZH+18] Xiangrong Zeng, Daojian Zeng, Shizhu He, Kang Liu, and Jun Zhao. Extracting relational facts by an end-to-end neural model with copy mechanism. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–514, 2018.

Curriculum Vitae

Graduate College
University of Nevada, Las Vegas

Mina Esmail Zadeh Nojoo Kambar
esmailzadeh.mina@gmail.com

Education

Master of Science in Information Technology 2017
Urmia University, Urmia, Iran

Work Experience

- Intern on Private Brand Intelligence Team, Amazon, Seattle 2024 — Present
  - Developing causal image and title quality scoring models using pairwise learning techniques to assess the impact of product images on customer satisfaction.
  - Conducting in-depth analyses to identify key factors and collaborating with cross-functional teams to implement data-driven strategies influencing customer satisfaction and improve image quality standards.
- Graduate Research Assistant, University of Nevada, Las Vegas 2019 — 2024
  - Fine-tuned Llama 2 for Electronic Health Records (EHRs) summarization using LoRA: a case study on ClinicalTrials.gov, achieved ROUGE-L Score: 0.38.
  - Developed a cutting-edge application leveraging Llama 2 and LangChain for structuring output, resulting in 65 informative binary features extracted from EHRs. Using vLLM for serving.

- Developed a graph neural network model in PyTorch for relation extraction, achieving 32% improvement in F1-score on the CHEMPROT dataset.
- Utilized Apache Spark and MLlib to predict trials' outcomes, differentiating between failed and successful studies using identified association rules.
- Designing Alzheimer's Drug Development Pipeline with the following features:
  * Implemented real-time data integration from clinicaltrials.gov and ALZFORUM.org, resulting in a 300% increase in data parsing speed through parallelism.
  * Developed an analytic dashboard,'Tri-AL', for the clinicaltrials.gov dataset using Python and JavaScript, enhancing data visualization by 45%.
  * Disease-related information extraction, and knowledge graph creation.
  * Developed decision tree model for sentiment analysis of drug descriptions, achieving a classification accuracy of 0.92.
  * Developed a neural network model with TensorFlow, fine-tuned the BioBert model for researchers of Brain Health Department, resulting in a 3% accuracy improvement.

- Data Scientist, PidaTech, Iran 2017 — 2019
  - Implemented demand forecasting models for grocery applications with a 20% improvement in accuracy.
  - Conducted market basket analysis on grocery logistics big data using the Apriori algorithm with Apache Spark, leading to a 14% increase in sales.
  - **Technologies**: Python (fbprophet), Java, Tableau, Azure, Apache Spark Streaming.

Publications

- M. Esmail Zadeh Nojoo Kambar, A. Esmaeilzadeh, and K. Taghva, *"Chemical-gene relation extraction with graph neural networks and BERT encoder,"* in The International Conference on Innovations in Computing Research. Springer, 2022, pp. 166–179.
- M. E. Z. N. Kambar, P. Nahed, J. R. F. Cacho, G. Lee, J. Cummings, and K. Taghva, *"Clinical Text Classification of Alzheimer's Drugs' Mechanism of Action,"* in Lecture Notes in Networks and Systems, 2022, vol. 235, doi: 10.1007/978-981-16-2377-6_48.

- A. Esmaeilzadeh, M. E. Z. N. Kambar, and M. Heidari, *"Graph attention neural network distributed model training,"* in 2022 IEEE World AI IoT Congress (AIIoT). IEEE, 2022, pp. 447–452
- J. Cummings, G. Lee, P. Nahed, M. E. Z. N. Kambar, K. Zhong, J. Fonseca, and K. Taghva *Alzheimer's disease drug development pipeline: 2022,"* Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 8, no. 1, p. e12295, 2022.
- M. Esmail Zadeh Nojoo Kambar, A. Esmaeilzadeh, Y. Kim, and K. Taghva, *"A Survey on Mobile Malware Detection Methods Using Machine Learning,"* in 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC) (IEEE CCWC 2022), Jan. 2022.
- M. E. Z. N. Kambar, A. Esmaeilzadeh, and M. Heidari, *"A survey on deep learning techniques for joint named entities and relation extraction,"* in 2022 IEEE World AI IoT Congress (AIIoT). IEEE, 2022, pp. 218–224.
- A. Esmaeilzadeh, J. R. F. Cacho, K. Taghva, M. E. Z. N. Kambar, and M. Hajiali, *"Building wikipedia n-grams with apache spark,"* in Science and Information Conference. Springer, 2022, pp. 672–684.
- P. Nahed, M. E. Z. N. Kambar, J. R. F. Cacho, G. Lee, J. Cummings, and K. Taghva, *"A recommendation model for predicting Alzheimer's's drugs' mechanism of action,"* in Sixth World Conference on Smart Trends for Systems Security and Sustainability, 2022. *In press.*

Awards

- **Best Paper Award** IEEE AIIOT 2022
  - Graph Attention Neural Network Distributed Model Training
- **Fellowships**
  - UNLV Graduate College Doctoral Research Fellowship: Awarded in 2023 for outstanding summer doctoral research fellowship.