A Survey of Medical Entity Recognition and Relation Extraction in Biomedical Text Mining

www.surveyx.cn

Abstract

In the rapidly evolving field of biomedical text mining, the integration of advanced computational methodologies, particularly deep learning (DL) and natural language processing (NLP), has significantly enhanced the extraction and analysis of complex medical information. This survey provides a comprehensive overview of these technologies, focusing on their application in medical entity recognition (MER), joint extraction, and relation extraction (RE) within biomedical texts. Key findings highlight the superiority of DL techniques over traditional methods in terms of generalizability and performance, with joint-learning models achieving notable improvements in RE while maintaining high accuracy in named entity recognition (NER). The integration of structured knowledge, as exemplified by models like DMNER, has demonstrated significant advancements in biomedical NER, underscoring the adaptability and effectiveness of these approaches. Furthermore, frameworks such as KRC and CERC have shown competitive performance in biomedical relation extraction and summarization tasks, addressing the challenges of information overload in clinical contexts. Despite these advancements, challenges remain, particularly concerning data availability, model interpretability, and ethical concerns. The survey emphasizes the importance of interdisciplinary collaboration and robust validation studies to ensure the safe and effective deployment of AI tools in healthcare. As DL and NLP techniques continue to advance, ongoing research focused on integrating domain-specific knowledge and developing sophisticated models is essential to further enhance data analysis, knowledge discovery, and ultimately improve healthcare outcomes.

1 Introduction

1.1 Structure of the Survey

This survey is structured into several key sections, each addressing critical facets of medical entity recognition and relation extraction within biomedical text mining. The introduction underscores the significance of these technologies in healthcare and medical research, laying the groundwork for a comprehensive exploration of relevant methodologies. The Background and Definitions section offers essential definitions and a historical context, referencing foundational works such as [1], which benchmarks Named Entity Recognition (NER) and Relation Extraction models, and [2], which examines NER techniques specific to biomedical texts.

The core of the survey is divided into segments focusing on Medical Entity Recognition, Joint Extraction of Entities and Relations, and Relation Extraction in Biomedical Texts. These sections delve into deep learning models and NLP approaches, the integration of external knowledge, and recent advancements, drawing insights from [3], which reviews NER advancements utilizing deep neural networks. The Joint Extraction section introduces methodologies and applications, emphasizing the importance of simultaneous extraction processes.

Generated by SurveyX.cn.

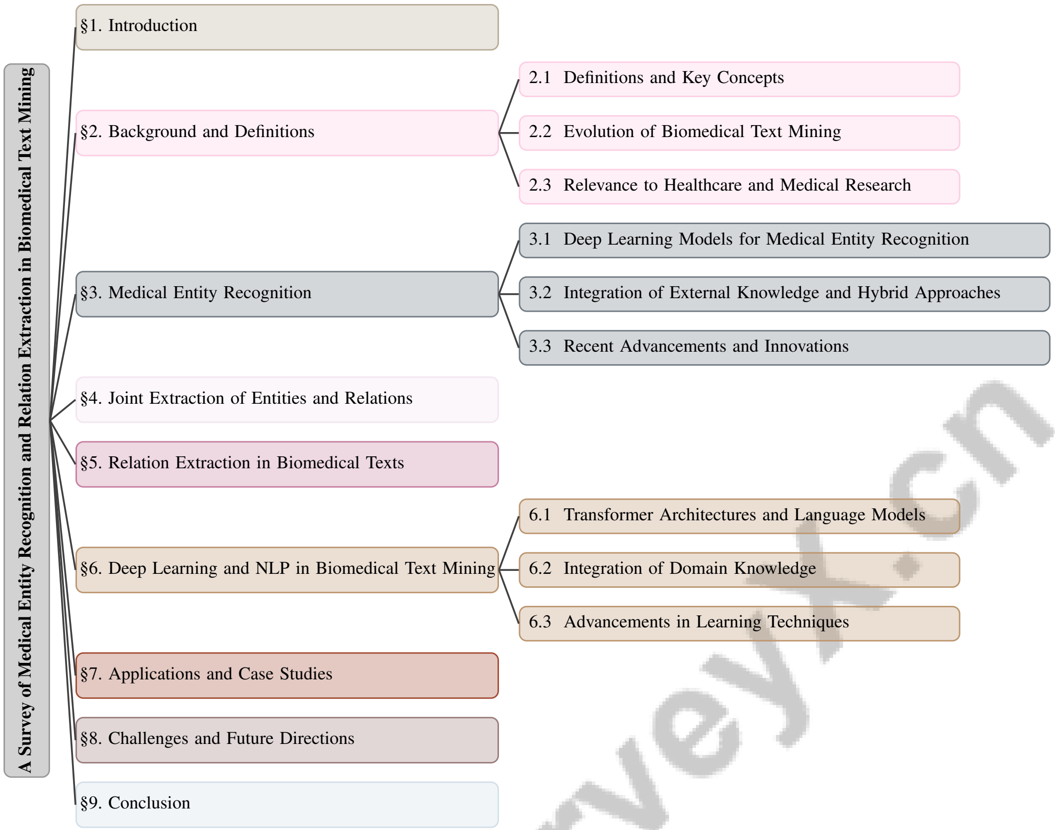

A Survey of Medical Entity Recognition and Relation Extraction in Biomedical Text Mining
§1. Introduction
§2. Background and Definitions
2.1 Definitions and Key Concepts
2.2 Evolution of Biomedical Text Mining
2.3 Relevance to Healthcare and Medical Research
§3. Medical Entity Recognition
3.1 Deep Learning Models for Medical Entity Recognition
3.2 Integration of External Knowledge and Hybrid Approaches
3.3 Recent Advancements and Innovations
§4. Joint Extraction of Entities and Relations
§5. Relation Extraction in Biomedical Texts
§6. Deep Learning and NLP in Biomedical Text Mining
6.1 Transformer Architectures and Language Models
6.2 Integration of Domain Knowledge
6.3 Advancements in Learning Techniques
§7. Applications and Case Studies
§8. Challenges and Future Directions
§9. Conclusion

Figure 1: chapter structure

Subsequently, the survey investigates the role of Deep Learning and NLP in biomedical text mining, discussing transformer architectures, domain knowledge integration, and advancements in learning techniques. The Applications and Case Studies section presents real-world examples, particularly in clinical contexts, to demonstrate the practical impact of these technologies.

The survey culminates in a discussion of Challenges and Future Directions, addressing issues such as data availability, model interpretability, and ethical concerns while identifying promising avenues for future research. This comprehensive structure aims to facilitate an in-depth examination of the evolving landscape and future potential of biomedical text mining technologies, particularly focusing on Literature Based Discovery (LBD) for hypothesis generation, the extraction of valuable information from biomedical abstracts using advanced deep learning models, and the pivotal role of Biomedical Information Extraction (BioIE) in processing extensive scientific literature to support clinical and research advancements [4, 5, 6].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Definitions and Key Concepts

Medical Entity Recognition (MER) plays a critical role in biomedical text mining by identifying and classifying entities such as diseases, treatments, and symptoms within medical texts [7]. This is essential for managing the information overload in biomedical literature and clinical notes, as generative models often struggle with specific medical concepts due to limited training data [8]. Named Entity Recognition (NER), a subset of MER, leverages advanced deep learning techniques to effectively identify and classify biomedical entities in unstructured healthcare texts, thereby enhancing biomedical NLP by accurately extracting entities like medications and diseases [9, 10]. NER is also crucial for evaluating large language models (LLMs) in terms of their clinical language understanding capabilities [11].

AI-generated, for reference only.

Relation Extraction (RE) involves the automatic identification of relationships between entities in unstructured texts, typically represented as triplets of head entity, relationship, and tail entity [12]. This process is fundamental for semantic understanding and knowledge summarization, particularly in extracting drug interactions from electronic health records [13]. Biomedical Text Mining employs computational techniques to transform unstructured data into structured insights, thereby enhancing decision-making in the biomedical field. The integration of Deep Learning (DL) techniques significantly improves NER and RE processes by modeling complex patterns in large datasets. Models like the Structural Bidirectional Long Short-Term Memory (SLSTM) capture interdependencies between words and sentences, enhancing precision and recall rates in these tasks. Frameworks such as DeepReader facilitate comprehensive information extraction from document images by identifying visual entities and establishing relational schemas [14, 15, 16].

Natural Language Processing (NLP) in the biomedical domain focuses on the interaction between computers and humans through natural language, converting unstructured clinical text into structured data to enhance clinical workflows and patient care. However, challenges arise from the limited availability of training data [17]. Information Extraction (IE) in this domain employs advanced NLP techniques, including NER and RE, to systematically extract and synthesize critical biological relationships from vast unstructured datasets, such as millions of PubMed abstracts. This process aids in identifying disease-related genes and predicting novel gene-disease associations, achieving a 20

2.2 Evolution of Biomedical Text Mining

Biomedical text mining has evolved from traditional rule-based and dictionary-based approaches, which relied heavily on extensive dictionaries and target-specific rules [18], to advanced computational techniques in information extraction. These traditional methods often neglected semantic information and prior knowledge, particularly in relation extraction tasks [19], and struggled with noise and overlapping entities as biomedical texts grew more complex [20].

The introduction of machine learning and deep learning methodologies marked a significant advancement, improving performance in information extraction by leveraging large datasets and sophisticated algorithms [21]. Deep learning has notably enhanced the accuracy and efficiency of NER and RE [22], while pre-trained language models (PLMs) have further refined relation extraction techniques to address the complexities inherent in biomedical texts [12].

Benchmarks have been crucial in the historical development of biomedical text mining, enabling the evaluation and comparison of various methods. However, traditional benchmarks often failed to capture the complexities of real-world biomedical data [23]. Recent initiatives have focused on creating more comprehensive benchmarks, such as Llama2-MedTuned, which assess the capabilities of large language models in biomedical tasks [10]. These benchmarks have been instrumental in refining extraction techniques and enhancing system performance.

The categorization of methods into dictionary-based, rule-based, machine learning, deep learning, and hybrid approaches underscores the evolution and effectiveness of each in clinical contexts [24]. This progression highlights the dominance of deep learning methods, which have excelled in managing complex biomedical texts. As the field continues to evolve, the integration of cutting-edge methodologies promises to enhance the efficiency and effectiveness of biomedical data analysis, ultimately improving healthcare outcomes.

2.3 Relevance to Healthcare and Medical Research

The integration of natural language processing (NLP) and deep learning techniques has profoundly transformed the extraction of medical information from unstructured clinical texts, significantly impacting healthcare and medical research. Accurately extracting social determinants of health (SDOH) from clinical narratives is essential, as these factors are crucial for health outcomes and are uniquely documented in clinical notes [25]. Moreover, effectively extracting and integrating diverse data types from electronic health records (EHRs) is vital, as advancements in chronic disease prediction can greatly improve healthcare delivery [26].

Effective relation extraction is crucial for understanding relationships between chemical-induced diseases, thereby influencing biomedical research [27]. Techniques such as ClinLinker, which enhance the linking of medical concepts in Spanish, show significant potential for improving data analysis and contributing to informed healthcare and research practices [28]. Furthermore, the extraction of relations between drugs and medication-related entities is critical for preventing adverse drug events (ADEs), impacting healthcare and medical research [29].

The exponential growth of biomedical literature necessitates more efficient automated extraction techniques to support researchers and healthcare professionals [30]. Benchmarks focusing on extracting structured, semantic representations of medical problems and drug information from clinical narratives, particularly in oncology, enhance healthcare by improving information extraction from clinical documents [31]. The application of deep learning in relation extraction has advanced healthcare and medical research by improving accuracy and adaptability in identifying relationships within biomedical texts [12].

Challenges such as data privacy, the need for high-quality labeled datasets, and the integration of AI systems into existing healthcare workflows persist [32]. Benchmarks designed to evaluate the performance of state-of-the-art large language models (LLMs) on clinical language understanding tasks are essential for enhancing the understanding of LLM capabilities and their impact on healthcare [11].

The application of deep learning in biomedical Named Entity Recognition (bNER) within clinical records emphasizes its significance for improving medical information extraction, illustrating the transformative potential of these technologies in healthcare [9]. However, the lack of sufficient training data remains a significant barrier, affecting the development of effective NLP methods in the biomedical domain [17].

Recent advancements in NLP and deep learning technologies have the potential to significantly transform medical research by streamlining the extraction of clinical information from unstructured sources, such as clinical trial reports and electronic health records. These innovations facilitate the efficient identification of medical entities and their relationships, enhancing clinical decision-making processes and integrating vast biomedical literature into practice. State-of-the-art models have demonstrated improved accuracy in tasks like named entity recognition and relation extraction, which are crucial for minimizing medical errors and optimizing treatment strategies. By leveraging these technologies, researchers and healthcare professionals can better utilize existing knowledge, ultimately advancing medical understanding and improving patient outcomes [33, 34, 35, 36].

3 Medical Entity Recognition

3.1 Deep Learning Models for Medical Entity Recognition

Deep learning models have revolutionized medical entity recognition (MER) by offering sophisticated frameworks for the precise identification and classification of entities in biomedical texts. As illustrated in Figure 2, the hierarchy of deep learning models for medical entity recognition categorizes these models into transformer-based models, large language models, and innovative frameworks. Transformer-based architectures, particularly from the BERT family like BioBERT and ClinicalBERT, excel in extracting intricate patterns from biomedical data, yielding substantial improvements in performance metrics. BioBERT, for example, achieves an F1-score of 0.89 in named entity recognition and 0.79 in relation extraction, outperforming traditional methods. Fine-tuning these models for specific biomedical applications has enhanced entity normalization accuracy by up to 1.17

The utilization of large language models (LLMs) such as GPT-3.5, GPT-4, LLaMA 2, and PMC LLaMA has further bolstered MER capabilities, especially in zero-shot and few-shot learning scenarios, critical in the biomedical domain where labeled data is limited [37, 10]. Innovative frameworks like DMNER, employing a two-step approach for entity boundary detection and matching, and RAMIE, focusing on dietary supplement entity identification, showcase significant contributions to MER [20, 38].

Models such as CoEx-Bert, which integrate entity and relation extraction tasks through multi-task learning, highlight deep learning's capacity to capture complex linguistic structures [39]. Lightweight Clinical Transformers (LCT), utilizing knowledge distillation and continual learning, illustrate the efficacy of deep learning in resource-constrained environments [40]. Ensemble deep learning methods like EDL-RE, which integrate weighted BiLSTM and transformer networks, enhance contextual understanding in electronic health records, indirectly benefiting MER [29]. The BERT-segMCNN architecture, combining BERT with multi-channel convolutional neural networks, shows promise in improving relation extraction performance [41]. Furthermore, the REMAP approach optimizes disease relation extraction by embedding partial knowledge graphs and medical language datasets into latent vector spaces, closely related to MER [42].

Comparative analyses reveal that deep learning approaches, particularly those employing LSTM and CRF, surpass traditional methods in accuracy and efficiency for Named Entity Recognition (NER) and related tasks [22]. The Adversarial Multi-task Learning Framework (Ad-MTL) utilizes adversarial training to enhance relationship extraction from biomedical texts, further illustrating deep learning's efficacy in this domain [30]. Benchmarking efforts, as described in [23], are crucial for evaluating state-of-the-art models against established baselines, driving the field forward by validating improvements in MER performance. The use of synthetic data generated through transformer-based methodologies has also been proposed to enhance training datasets for biomedical NLP tasks, providing a novel solution to data scarcity [17]. As neural architectures evolve, deep learning models' capabilities in biomedical text mining are set to improve, ultimately benefiting healthcare delivery and medical research.

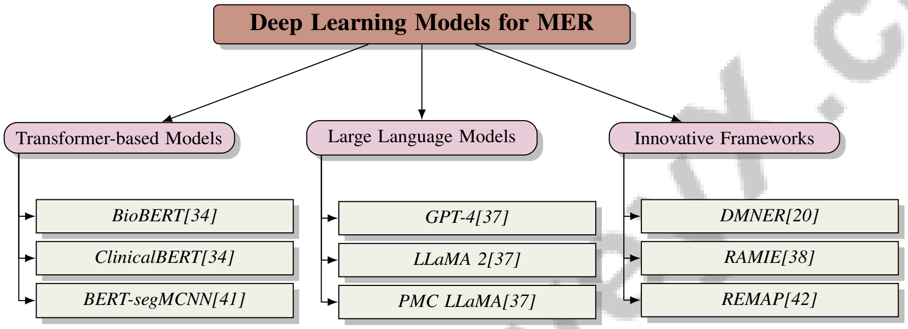


Figure 2: This figure illustrates the hierarchy of deep learning models for medical entity recognition, categorizing them into transformer-based models, large language models, and innovative frameworks.

3.2 Integration of External Knowledge and Hybrid Approaches

The integration of external knowledge bases and hybrid models is pivotal in advancing medical entity recognition (MER), enhancing accuracy and efficiency by leveraging domain-specific knowledge and combining computational techniques. Hybrid models that merge rule-based methods with data-driven deep learning frameworks offer robust solutions for refining model predictions. Knowledge-augmented frameworks, which incorporate structured knowledge, advocate for blending rule-based and data-driven methods to enhance entity recognition [43].

External knowledge bases significantly improve the contextual understanding and prediction accuracy of MER systems. The KECI framework exemplifies this integration by utilizing a structured graph approach for better contextual comprehension and improved predictions [44]. Similarly, entity linking methods that connect concepts to reference knowledge bases effectively incorporate external knowledge into the recognition process, as seen in various clinical applications [45].

Hybrid approaches also combine traditional machine learning techniques with syntactic information to enhance entity recognition. The Shortest Dependency Path Global Alignment Support Vector Machine (SDP-GA-SVM) exemplifies such a hybrid approach, utilizing classic machine learning techniques alongside syntactic information to improve relation extraction [46]. This method highlights the potential of diverse methodologies in addressing complex biomedical text mining challenges.

The integration of relational models in information extraction frameworks, such as DeepReader, captures relationships between extracted entities, paralleling the use of external knowledge bases in hybrid approaches [15]. These models effectively utilize external knowledge to enhance the semantic understanding of biomedical texts, improving the reliability of extracted information.

Moreover, the RAMIE framework emphasizes the integration of multi-task learning and retrieval-augmented generation to enhance dietary supplement information extraction [38]. This approach illustrates the effectiveness of leveraging external knowledge and hybrid models in complex information extraction tasks.

AI-generated, for reference only.

The categorization of current methods into gene-disease relations, gene-protein interactions, protein-protein interactions, and point mutation extraction further illustrates the diverse applications of hybrid approaches in biomedical text mining [47]. By integrating external knowledge and employing hybrid models, these methodologies present comprehensive solutions to the challenges faced in medical entity recognition, ultimately contributing to better healthcare outcomes and advancements in medical research.

3.3 Recent Advancements and Innovations

Method Name	Technological Integration	Domain-Specific Adaptation	Performance Challenges
LLMMs[26]	Llmms Framework	Chronic Disease Semantics	Dataset Limitations
CL[28]	-	Sapbert IN Clinlinker	Terminology Variability
LCT[40]	-	Sapbert IN Clinlinker	Dataset Limitations
Ad-MTL[30]	-	-	Terminology Variability
GPT[13]	Soft Prompting	Gatortrongpt	Hallucinations
FAM[7]	Bert Integration	Task Representation Encoder	Terminology Variability

Table 1: Table 1 presents a comparative analysis of recent methodologies in medical entity recognition and relation extraction, focusing on technological integration, domain-specific adaptations, and performance challenges. The table highlights the diverse approaches, including Large Language Multimodal Models and domain-adapted language models, and their respective limitations in dataset availability and terminology variability.

Recent advancements in medical entity recognition (MER) and relation extraction (RE) are significantly driven by the integration of large language models and innovative frameworks, enhancing both accuracy and efficiency. Table 1 provides a comprehensive overview of these recent advancements, illustrating the integration of large language models and the challenges faced in improving model performance. The introduction of Large Language Multimodal Models (LLMMs) exemplifies this trend, utilizing attention mechanisms to integrate clinical and laboratory data, thereby improving predictive capabilities in clinical settings [26]. This approach emphasizes the trend of leveraging multimodal data for comprehensive analysis.

In-domain adapted language models, such as SapBERT used in ClinLinker, represent substantial improvements over existing multilingual models, providing enhanced performance in medical entity linking tasks [28]. These advancements highlight the significance of domain-specific adaptations in improving model performance.

Lightweight clinical transformers have emerged as breakthroughs in clinical NLP tasks, delivering superior performance with reduced latency and computational resource requirements [40]. This innovation is particularly valuable in resource-constrained environments where efficiency is critical.

Despite these advancements, challenges persist, particularly regarding the variability of biomedical terminology and the lack of comprehensive datasets, which can lead to inaccuracies in entity recognition [22]. The ongoing evolution of neural network models continues to address these issues, with deep learning approaches demonstrating significant improvements in Named Entity Recognition (NER) accuracy over traditional methods [3].

In relation extraction, recent advancements driven by deep learning and pre-trained language models (PLMs) have enhanced both performance and adaptability [12]. Multi-task learning and adversarial training frameworks have shown superior performance across various biomedical tasks, highlighting their benefits in improving relation extraction [30].

Generative large language models have also contributed to state-of-the-art performance across multiple clinical NLP tasks, further advancing medical entity recognition [13]. The focused attention model has achieved remarkable accuracy in both entity and relation extraction, with F1-scores of 96.89% for NER and 88.51% for relation classification [7].

The integration of deep learning techniques has substantially improved the efficiency and accuracy of biomedical Named Entity Recognition (bNER) systems, facilitating better extraction of relevant medical information from electronic health records (EHRs) [9]. As these technologies evolve, they have the potential to significantly enhance data analysis and knowledge discovery in biomedical text mining, ultimately contributing to improved healthcare delivery and medical research.

4 Joint Extraction of Entities and Relations

The joint extraction of entities and relations has become increasingly significant in biomedical informatics, driven by the complexity of biomedical data and the demand for efficient extraction methodologies. This section explores the foundational aspects of joint extraction, beginning with an overview of its conceptual framework and relevance to contemporary information extraction challenges. Subsequently, it delves into the methodologies underpinning joint extraction, highlighting innovative approaches and techniques emerging in this evolving domain.

4.1 Introduction to Joint Extraction

Joint extraction marks a paradigm shift in information extraction, particularly in the biomedical domain, where the complex interactions among entities like diseases, treatments, and biological processes necessitate a comprehensive analytical approach. This methodology enables concurrent identification of entities and their interrelations, significantly enhancing the efficiency and precision of information extraction tasks [7]. By integrating Named Entity Recognition (NER) and Relation Extraction (RE) into a unified framework, joint extraction methods streamline processes and mitigate error propagation common in sequential pipeline approaches.

The importance of joint extraction lies in its ability to extract relations across entire documents, encompassing complex interactions such as coreference [12]. This capability is vital for constructing large-scale biomedical knowledge bases, where automated systems can discover and rank patterns, facilitating expert annotation and rapid knowledge base development [48].

Frameworks like RAMIE exemplify joint extraction by identifying dietary supplement entities and their relationships, enhancing model accuracy in discerning biomedical entities [38]. Reformulating joint extraction tasks into a text-to-text problem, as seen in generative approaches to clinical NLP tasks, underscores its potential in managing complex biomedical data [13]. This approach aligns with the intuition that while GPT models capture nuanced contextual information, they may overlook specific medical terminologies documented in knowledge bases like UMLS [8].

The Bidirectional Tree Tagging (BiTT) scheme further illustrates methodological advancements by organizing relation triples into binary trees and converting them into sequences of word-level tags, facilitating efficient extraction of medical triples [49]. Such innovative frameworks, along with advanced entity representation models and external knowledge integration, have consistently outperformed existing methods, setting new benchmarks for accuracy in biomedical entity normalization tasks [39]. Challenges faced by large language models (LLMs) in accurately performing RE and NER in biomedical contexts are addressed through rigorous benchmarking efforts, emphasizing the need for continued innovation in joint extraction methodologies [10].

Joint extraction methodologies significantly enhance the understanding and utilization of complex biomedical information. By integrating NER and RE within a unified framework, these advanced approaches improve the extraction and semantic linking of entities in unstructured biomedical data. This dual focus addresses the growing volumes of biomedical literature and clinical records, leading to significant improvements in data analysis and knowledge discovery. They achieve state-of-the-art performance on multiple benchmark datasets, facilitating the construction of biomedical knowledge graphs and enhancing the accuracy of clinical code mapping, crucial for optimizing text mining techniques in healthcare and medical research, ultimately driving better insights and outcomes in biomedical text mining [50, 35].

4.2 Methodologies for Joint Extraction

Joint extraction methodologies have evolved to incorporate advanced neural architectures and innovative learning paradigms, significantly enhancing the simultaneous extraction of entities and their relationships. A notable approach involves framing relation extraction as a multi-head selection problem, allowing the identification of multiple relations for each entity, improving the overall extraction process [51]. This method aligns with leveraging contextual embeddings and transformer architectures, capturing complex dependencies within biomedical texts to enhance accuracy and efficiency [12].

AI-generated, for reference only.

The integration of knowledge graph embeddings with deep language models, as demonstrated by the REMAP framework, addresses challenges of missing data types and improves disease relation extraction by fusing structured knowledge with contextual language understanding [42]. This approach exemplifies the potential of combining external knowledge with deep learning techniques to refine extraction accuracy.

Recent advancements include ensemble deep learning methods, integrating multiple models to enhance the joint extraction of relations between drugs and medication entities [29]. These ensemble approaches leverage the strengths of various models to improve robustness and reliability.

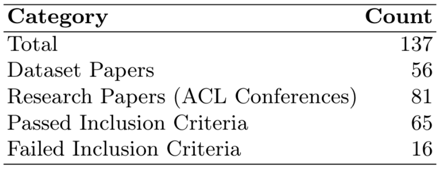
Frameworks like BiOnt, utilizing multiple biomedical ontologies, improve relation extraction performance by incorporating domain-specific knowledge into the learning process [52]. This integration aids in the disambiguation of entities and the accurate identification of relationships within complex biomedical texts.

The tagging scheme BiTT represents medical relation triples as two binary trees, facilitating efficient extraction of overlapping entities and relations [49]. This innovative approach addresses the challenge of overlapping entity recognition, crucial for accurate joint extraction.

Benchmarking efforts, evaluating models like BERT, CNN, and GRU, provide critical insights into the performance of state-of-the-art models on relation extraction tasks, highlighting strengths and limitations of different methodologies [53]. These benchmarks guide future research and development in joint extraction.

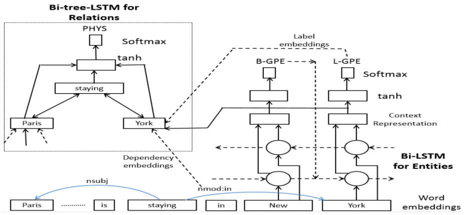
The use of abstractified multi-instance learning addresses the long-tail distribution of fact triples in biomedical relation extraction, where many entity pairs are supported by limited evidence [54]. This approach enhances generalization capabilities of joint extraction models, enabling effective handling of sparse data scenarios.

Overall, methodologies for joint extraction advance through integrating sophisticated neural architectures, external knowledge bases, and ensemble learning techniques. These innovations enhance biomedical text mining by leveraging advanced techniques such as Literature Based Discovery (LBD) and state-of-the-art natural language processing models, including BioBERT and PubTator 3.0. By automating novel associations between medical terms and extracting meaningful insights from scientific literature, these tools improve data analysis accuracy and efficiency, facilitating knowledge discovery in healthcare, supporting applications like precision oncology and biomedical knowledge graph development, accelerating research translation into clinical practice [4, 34, 35].



Category	Count
Total	137
Dataset Papers	56
Research Papers (ACL Conferences)	81
Passed Inclusion Criteria	65
Failed Inclusion Criteria	16

(a) Dataset Papers and Research Papers in ACL Conferences[55]



(b) Bi-tree-LSTM for Relations[56]

Figure 3: Examples of Methodologies for Joint Extraction

As shown in Figure 3, two notable examples illustrate diverse approaches within joint extraction. The first, derived from ACL Conferences, highlights the categorization and distribution of papers, showcasing 137 papers divided into 56 dataset and 81 research papers, with 65 passing inclusion criteria. This structured presentation underscores the breadth of research dedicated to advancing joint extraction techniques. The second introduces a neural network architecture, Bi-tree-LSTM for Relations, emphasizing sentence structure and entity processing integration. This architecture employs Bi-tree-LSTM for sentence dependencies and Bi-LSTM for entity recognition, utilizing dependency, label, and word embeddings to enrich input data representation. Together, these examples provide a comprehensive overview of methodologies employed in joint extraction, highlighting the research landscape and innovative neural network architectures developed to enhance this field [55, 56].

4.3 Applications of Joint Extraction

Joint extraction techniques have emerged as a powerful approach in biomedical text mining, offering significant improvements over traditional methods by simultaneously extracting entities and their relationships. These techniques enhance performance in both entity and relation extraction tasks, leveraging the interconnected nature of these processes to achieve higher accuracy and efficiency [56]. By addressing the limitations of sequential extraction models, joint extraction frameworks reduce error propagation and improve the coherence of extracted information.

One practical application is in temporal ordering of clinical events, where techniques like TEEM provide robust solutions with high precision and recall rates [57]. This capability is crucial for constructing accurate timelines of patient histories, facilitating better clinical decision-making and patient management.

The integration of neural networks with log-linear models exemplifies joint extraction methodologies' potential to combine different models' strengths, leading to improved recall and precision for various relation types [58]. This hybrid approach benefits complex biomedical texts, where relation diversity requires sophisticated modeling techniques.

Joint learning models enhance extraction performance by utilizing the relationship between relation extraction and function detection, as demonstrated in BEL statement extraction [59]. This integration allows for comprehensive analysis and interpretation of biomedical data, supporting advanced research and healthcare applications.

Challenges remain, particularly in cross-domain evaluations where current studies often lack clarity in setups [60]. The use of multiple ontologies, as seen in frameworks like BiOnt, addresses some challenges by enhancing model performance in identifying relations beyond training data limitations [52]. This approach underscores the importance of external knowledge integration in improving joint extraction models' adaptability and accuracy.

Applications of joint extraction in biomedical text mining demonstrate potential to significantly enhance data analysis and knowledge discovery. By addressing traditional information processing methods' limitations, advanced techniques in information extraction enhance healthcare delivery and medical research. These innovations streamline critical data extraction from unstructured medical documents, such as clinical trial reports and electronic medical records, reducing medical errors and improving decision-making. This progress facilitates more accurate treatment and outcome identification, supporting sophisticated information extraction system development adapting to the evolving medical information landscape [36, 33].

In recent years, the field of relation extraction (RE) in biomedical texts has witnessed significant advancements, particularly through the integration of deep learning methodologies and hybrid approaches. To illustrate this progress, Figure 4 presents a comprehensive overview of the hierarchical categorization of various relation extraction techniques. This figure delineates the main categories, including advancements in deep learning, which have fundamentally transformed the landscape of RE, as well as hybrid approaches that combine multiple methodologies for enhanced performance.

Moreover, it underscores the importance of foundational and specialized datasets, such as clinical and social media datasets, which are crucial for the development and evaluation of RE models. The figure further highlights recent innovations that have bolstered RE techniques, specifically focusing on knowledge-driven approaches and scalability improvements. These advancements are pivotal in facilitating more efficient and accurate information extraction within the biomedical domain, ultimately contributing to the enhancement of research and clinical applications.

5 Relation Extraction in Biomedical Texts

5.1 Overview of Relation Extraction Techniques

Relation extraction (RE) in biomedical texts is pivotal for identifying and categorizing relationships among entities such as genes, proteins, diseases, and drugs in unstructured data, thereby facilitating the construction of biomedical knowledge bases and advancing research and healthcare applications. The complexity of biomedical data, characterized by dense relational triples and diverse relationship types, poses unique challenges for RE [52].

AI-generated, for reference only.

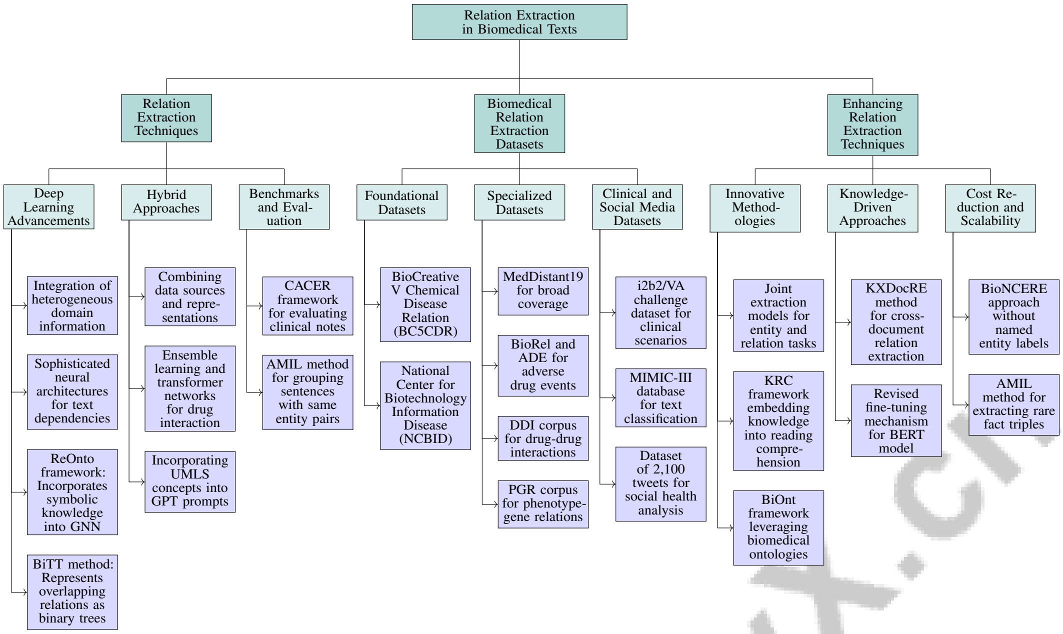

Relation Extraction in Biomedical Texts
Relation Extraction Techniques
Biomedical Relation Extraction Datasets
Enhancing Relation Extraction Techniques
Deep Learning Advancements
Hybrid Approaches
Benchmarks and Evaluation
Foundational Datasets
Specialized Datasets
Clinical and Social Media Datasets
Innovative Methodologies
Knowledge-Driven Approaches
Cost Reduction and Scalability
Integration of heterogeneous domain information
Sophisticated neural architectures for text dependencies
ReOnto framework: Incorporates symbolic knowledge into GNN
BiTT method: Represents overlapping relations as binary trees
Combining data sources and representations
Ensemble learning and transformer networks for drug interaction
Incorporating UMLS concepts into GPT prompts
CACER framework for evaluating clinical notes
AMIL method for grouping sentences with same entity pairs
BioCreative V Chemical Disease Relation (BC5CDR)
National Center for Biotechnology Information Disease (NCBID)
MedDistant19 for broad coverage
BioRel and ADE for adverse drug events
DDI corpus for drug-drug interactions
PGR corpus for phenotype-gene relations
i2b2/VA challenge dataset for clinical scenarios
MIMIC-III database for text classification
Dataset of 2,100 tweets for social health analysis
Joint extraction models for entity and relation tasks
KRC framework embedding knowledge into reading comprehension
BiOnt framework leveraging biomedical ontologies
KXDocRE method for cross-document relation extraction
Revised fine-tuning mechanism for BERT model
BioNCERE approach without named entity labels
AMIL method for extracting rare fact triples

Figure 4: This figure illustrates the hierarchical categorization of relation extraction techniques in biomedical texts. The main categories include advancements in deep learning, hybrid approaches, and benchmarks for evaluation. Furthermore, it highlights the significance of foundational and specialized datasets, including clinical and social media datasets, essential for developing and evaluating RE models. The figure also emphasizes recent advancements enhancing RE techniques, focusing on innovative methodologies, knowledge-driven approaches, and scalability improvements, contributing to more efficient and accurate information extraction in the biomedical domain.

Recent advances in deep learning have significantly enhanced RE by integrating heterogeneous domain information and employing sophisticated neural architectures to capture intricate text dependencies. The ReOnto framework exemplifies this by incorporating symbolic knowledge from ontologies into a Graph Neural Network (GNN), enhancing extraction accuracy through local and global information [61]. The BiTT method excels in extracting overlapping medical relation triples by representing them as binary trees and converting them into word-level tag sequences [49].

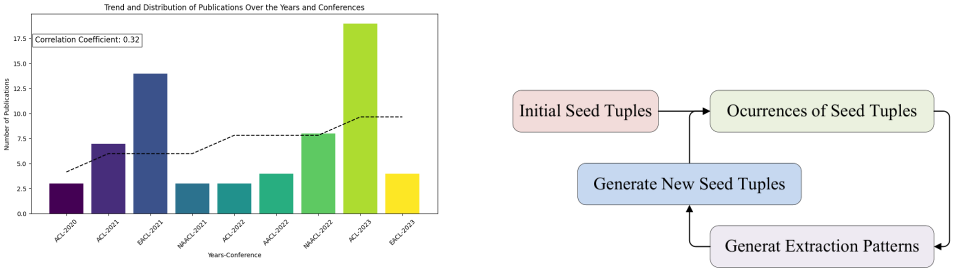
Hybrid approaches, combining various data sources and representations, have shown promise in improving RE techniques. Ensemble learning and transformer networks, particularly for drug interaction identification, enhance model robustness and reliability [13]. Incorporating UMLS concepts into generative prompts for GPT models aims to improve clinical entity and relationship identification [8].

Benchmarks like CACER provide accurate evaluation frameworks for extracting events and relationships from clinical notes, addressing existing shortcomings and improving RE technique assessments [31]. The Abstractified Multi-instance Learning (AMIL) method enhances coherence and accuracy by grouping sentences with the same entity pair into bags based on semantic types [54].

The continuous evolution of RE techniques, driven by deep learning advancements and external knowledge integration, significantly enhances data analysis and knowledge discovery in the biomedical domain. As methodologies like Named Entity Recognition (NER) and Relation Extraction (RE) advance, they are expected to improve the accuracy and efficiency of extracting critical medical information from unstructured text, reducing medical errors and streamlining decision-making processes for healthcare professionals, ultimately leading to better healthcare outcomes and accelerated medical research [21, 36, 33].

As illustrated in Figure 5, relation extraction in biomedical texts is a critical research area focused on identifying and classifying relationships between entities. This process is vital for converting unstructured data into structured knowledge, facilitating advancements in biomedical research and applications. The first image depicts the growth and distribution of publications from 2020 to

AI-generated, for reference only.



(a) Trend and Distribution of Publications Over the Years and Conferences[55]

(b) Algorithmic Process for Seed Tuples Generation and Extraction[62]

Figure 5: Examples of Overview of Relation Extraction Techniques

2023, highlighting contributions from various conferences such as ACL-2020, underscoring the increasing scholarly interest in this field. The second image outlines the algorithmic process for seed tuples generation and extraction, illustrating a systematic approach to identifying and extracting patterns from seed tuples. Together, these examples provide a snapshot of the dynamic landscape and methodological rigor characterizing relation extraction in biomedical texts [55, 62].

5.2 Biomedical Relation Extraction Datasets

Benchmark	Size	Domain	Task Format	Metric
MD19[53]	450,071	Biomedical Relation Extraction	Relation Extraction	AUC, F1-micro
BEAR[48]	6,324	Biomedical	Entity Recognition And Relation Extraction	F1-score
BioRE[63]	25,000	Biomedical Relation Extraction	Relation Extraction	micro F1
Kiwi[64]	1,588	Clinical Information Extraction	Named Entity Recognition	F1
RadGraph[65]	500	Radiology	Entity And Relation Extraction	Micro F1
DTR[66]	4075	Clinical Text Analysis	Temporal Relation Extraction	F-measure, Precision
MedMentions[67]	4,392	Biomedical	Entity Recognition	F1
K-CNER[68]	536	Clinical Entity Recognition	Named Entity Recognition	F1-score, Precision

Table 2: This table presents a comprehensive overview of key benchmarks utilized in the field of biomedical relation extraction. It includes details on dataset size, domain, task format, and evaluation metrics, thereby highlighting the diversity and scope of resources available for advancing research in this domain.

Developing and evaluating relation extraction (RE) models in the biomedical domain heavily rely on diverse datasets that cover a broad spectrum of biomedical relationships. These datasets are crucial for training models to accurately identify and interpret complex interactions among entities such as genes, proteins, diseases, and drugs. The BioCreative V Chemical Disease Relation (BC5CDR) task corpus and the National Center for Biotechnology Information Disease (NCBID) corpus are foundational datasets for evaluating RE models, providing a robust framework for understanding chemical-disease interactions [69].

The MedDistant19 dataset, with over 450,000 instances across training, validation, and test sets, serves as a benchmark for evaluating RE models [53]. Similarly, the BioRel and ADE datasets assess RE methods' effectiveness regarding adverse drug events and other biomedical interactions [61].

Datasets like the DDI corpus for drug-drug interactions, the PGR corpus for phenotype-gene relations, and the BC5CDR corpus for chemical-induced disease relations are essential resources for testing and refining RE systems [52]. Additionally, the i2b2/VA challenge dataset and collections of de-identified clinical notes provide diverse representations of clinical scenarios, crucial for training models in clinical relation extraction [30].

The MIMIC-III database, which includes electronic health records for text classification and temporal relation extraction tasks, has been extensively used to explore the application of artificial data in enhancing biomedical NLP models [17]. Furthermore, a dataset containing 2,100 tweets with approximately 6,000 entity annotations and 3,000 relation annotations offers a rich resource for analyzing patient experiences and understanding social determinants of health [48].

The diverse range of datasets available for biomedical relation extraction underscores the intricate and multifaceted nature of the field. These datasets not only enhance the training and evaluation of RE models, such as the newly proposed accuracy-optimized BioBERT and speed-optimized Fully Connected Neural Network (FCNN) models, but also facilitate significant advancements in understanding complex relationships within biomedical texts. Recent benchmarks have demonstrated state-of-the-art performance across various datasets, including the i2b2 Clinical Relations challenges and the Phenotype-Gene Relations dataset, which is critical for developing robust applications like biomedical knowledge graphs and improving clinical coding accuracy, ultimately driving innovations in text mining techniques capable of processing the growing volume of unstructured biomedical data [70, 35, 71, 27, 63]. Table 2 provides a detailed examination of the representative benchmarks used for evaluating biomedical relation extraction models, underscoring their significance in advancing the field.

5.3 Enhancing Relation Extraction Techniques

Recent advancements in relation extraction (RE) techniques emphasize improving accuracy and efficiency through innovative methodologies and the integration of domain-specific knowledge. Joint extraction models, which simultaneously address entity and relation extraction tasks, demonstrate substantial performance improvements by effectively reducing redundant information [72].

Incorporating domain knowledge has been pivotal in enhancing RE techniques. The KRC framework exemplifies this by embedding knowledge into a reading comprehension model, significantly improving relation extraction capabilities [27]. Similarly, the BiOnt framework shows that leveraging knowledge from biomedical ontologies can lead to notable performance enhancements, achieving substantial F-score improvements across multiple datasets [52].

Cross-document relation extraction has also progressed through knowledge-driven approaches. The KXDocRE method incorporates domain knowledge to enhance predictive performance and interpretability, providing explanatory text for predicted relations [73]. This highlights the importance of domain knowledge in improving the interpretability and accuracy of RE models.

In terms of model architecture, a revised fine-tuning mechanism for the BERT model has achieved state-of-the-art results on relation extraction tasks, underscoring the potential of optimized neural architectures in enhancing RE performance [74]. The ReOnto framework, which combines symbolic knowledge with deep learning, has demonstrated improved accuracy in capturing complex relationships within biomedical texts [61].

The BioNCERE approach stands out by performing relation extraction without the need for named entity labels, simplifying the training process and reducing annotation costs [75]. This innovation addresses data annotation challenges, making RE more scalable and adaptable.

Additionally, the Abstractified Multi-instance Learning (AMIL) method significantly improves performance in extracting rare fact triples, showcasing its effectiveness in sparse data scenarios [54].

These recent advancements in RE techniques signify a transformation in the field, driven by innovative methodologies such as cross-document relation extraction (CrossDocRE) and the integration of domain-specific knowledge through frameworks like KXDocRE. Collectively, these developments enhance the interpretability and performance of RE systems, addressing critical challenges by leveraging deep learning approaches and large pre-trained language models. This shift towards more sophisticated and context-aware RE applications is essential for knowledge graph completion and information retrieval across various domains, including biomedicine [73, 12]. These improvements contribute to more accurate and efficient information extraction in the biomedical domain, ultimately enhancing data analysis and knowledge discovery.

6 Deep Learning and NLP in Biomedical Text Mining

The integration of deep learning methodologies in biomedical text mining, particularly through transformer architectures and large language models (LLMs), has been transformative. These innovations have enhanced the extraction and interpretation of complex medical information, facilitating sophisticated applications in clinical settings. This section examines the impact of these architectures on natural language processing (NLP) within the biomedical domain, paving the way for deeper exploration of their capabilities and applications.

6.1 Transformer Architectures and Language Models

Transformer architectures and LLMs have revolutionized biomedical text mining by significantly improving information extraction from complex medical texts. Models such as BERT, RoBERTa, and XLNet have advanced clinical relation extraction, especially when pretrained on clinical texts, capturing domain-specific nuances [76]. GatorTron, tailored for processing clinical narratives in electronic health records (EHRs), exemplifies the application of these models in healthcare [77].

Adapting transformer architectures to medical entity recognition has addressed prior limitations, enhancing extraction accuracy [78]. BioBERT has notably improved relation extraction (RE) frameworks, demonstrating the influence of transformer architectures on biomedical text mining [35]. This is further evidenced by state-of-the-art performance in automated adverse drug event (ADE) extraction using contextualized language models like BioBERT [79].

Transformers also play a crucial role in multi-task learning models, with MT-BERT-Refinement and MT-BERT-Fine-Tune enhancing performance across tasks [80]. The integration of LLMs in frameworks such as RAMIE has improved contextual understanding during extraction processes, showcasing versatility in managing complex biomedical data [38].

The LLM-IE package exemplifies how LLMs enhance deep learning capabilities in biomedical text mining, offering a user-friendly interface for practitioners [81]. Additionally, transformer networks' ability to capture inter-sentence relations boosts relation extraction accuracy, as demonstrated in methods aimed at improving relationship identification within biomedical texts [29].

Recent advancements in transformer architectures and LLMs have improved NLP in biomedical text mining, facilitating more accurate and efficient information extraction tasks, including Named Entity Recognition (NER) and Relation Extraction (RE). Models like BioBERT and PubTator 3.0 have achieved F1-scores of 0.93 and 0.79, respectively, showcasing LLMs' capacity to leverage extensive biomedical knowledge for high-throughput information extraction, supporting critical applications in precision oncology and enhancing knowledge discovery [34, 82, 83, 21, 6]. These developments continue to drive innovation, presenting promising avenues for future research and application in healthcare.

6.2 Integration of Domain Knowledge

Integrating domain knowledge into deep learning models is crucial for enhancing the accuracy and interpretability of biomedical text mining. This approach refines model predictions and improves the extraction of relevant biomedical entities and relationships. The Hybrid Processing System (HPS), which combines machine learning and rule-based techniques, exemplifies the importance of domain knowledge in deep learning applications [84].

The KECI framework demonstrates the effectiveness of incorporating global relational information into local representations, enhancing entity mention linking through domain-specific knowledge [85]. This integration fosters precise and context-aware entity recognition and relation extraction in complex biomedical texts.

DeepReader illustrates the application of domain knowledge by allowing users to query extracted data via SQL or natural language interfaces, enhancing the comprehension and utility of extracted information [15]. This method showcases the potential of combining domain knowledge with user-friendly query systems for effective information retrieval.

Incorporating ontological embeddings with traditional word embeddings, as seen in the BiOnt framework, significantly improves the identification of relations between biomedical entities [52]. This strategy leverages structured domain knowledge to enhance semantic understanding, leading to more accurate extraction outcomes.

The use of guided prompt design in LLMs further illustrates the integration of domain knowledge to produce structured outputs, improving the precision and relevance of extracted information [86]. This technique highlights how domain-specific prompts can guide model outputs toward more accurate representations of biomedical data.

AI-generated, for reference only.

Moreover, the joint learning framework for causal relation extraction utilizes shared information between tasks to enhance detection precision and overall extraction performance [59]. This method emphasizes the significance of task interdependency and shared domain knowledge in improving extraction accuracy.

Lightweight transformers, which capitalize on the knowledge of existing large models while minimizing computational demands, exemplify the application of domain knowledge in resource-constrained clinical environments [40]. This approach ensures advanced deep learning capabilities remain accessible and practical for clinical applications, facilitating improved healthcare outcomes through efficient information extraction.

6.3 Advancements in Learning Techniques

Advancements in learning techniques for biomedical text mining have been significantly influenced by innovative models and approaches that enhance the extraction and classification of complex medical information. Large language models (LLMs) such as GPT-4 have shown promise in reasoning tasks, underscoring their potential to improve accuracy in biomedical contexts [37]. These models leverage extensive knowledge and reasoning capabilities, supported by contextual information, to enhance relation extraction [82].

BioBERT, recognized for capturing domain-specific nuances, has demonstrated substantial improvements in clinical document understanding and relation extraction when applied to benchmark datasets [71]. Its performance is further enhanced by integrating attention mechanisms and domain-specific contextual embeddings, improving feature extraction and relation identification [87]. The focused attention model exemplifies these advancements by leveraging pre-trained knowledge from BERT while allowing for task-specific adaptations, enhancing information extraction precision [7].

Transfer learning has emerged as a pivotal technique, enabling models to benefit from previously acquired knowledge, improving their ability to interpret complex clinical narratives [88]. Additionally, extending BERT models to incorporate multiple instance learning and novel data encoding schemes has been proposed to reduce noise and enhance relation extraction [89]. Emphasizing semantic relations as a comprehensive method has also been highlighted for improving retrieval outcomes in clinical contexts [45].

In federated learning, the FedCMC approach leverages major classifier vectors to enhance performance and convergence speed, showcasing advancements in collaborative learning techniques [90]. This method exemplifies the potential of federated learning to improve model performance by effectively utilizing distributed data sources.

The DRAA framework's dynamic resource allocation principle maximizes resource utilization, crucial for enhancing learning techniques in biomedical text mining [91]. AI-assisted knowledge discovery frameworks have demonstrated the effectiveness of NLP solutions in extracting and synthesizing knowledge from biomedical literature, with BioBERT showing promise in identifying key entities and relations relevant to precision oncology [34].

Instruction tuning has significantly enhanced the performance of Llama2 models in biomedical NLP tasks, indicating further advancements in learning techniques [10]. This approach highlights the potential of tailored instruction sets to guide model outputs toward more accurate representations of biomedical data.

These advancements underscore the continuous evolution of learning techniques in biomedical text mining, driven by innovative models and methodologies that enhance the extraction, classification, and interpretation of complex medical data. Future research may focus on expanding datasets, improving model performance, and exploring new techniques for handling complex entity relationships [92].

7 Applications and Case Studies

7.1 Clinical Applications in Oncology

The integration of natural language processing (NLP) and deep learning in oncology has significantly improved the extraction and analysis of clinical information, enhancing patient outcomes through personalized treatment approaches. By leveraging NLP with electronic health records (EHRs), critical data such as tumor characteristics and treatment responses can be extracted from unstructured clinical narratives, refining oncological care and individualizing treatment plans [11]. Large language models (LLMs) have advanced clinical language understanding, enabling the identification of intricate relationships among cancer-related entities and facilitating structured semantic representations from clinical documents [31]. These advancements are crucial for developing comprehensive cancer registries and identifying novel therapeutic targets.

Deep learning models have enhanced the accuracy of named entity recognition (NER) tasks, allowing precise extraction of oncological entities such as tumor markers and genetic mutations [9]. This precision is vital for advancing personalized medicine, where treatment regimens are tailored based on genetic profiles. AI-driven solutions have also improved the identification and management of adverse drug events (ADEs) in cancer treatment, where polypharmacy complexity poses significant risks [29]. By employing NLP techniques, healthcare providers can better monitor and mitigate these risks, thus enhancing patient safety and treatment efficacy.

In addition to optimizing clinical workflows, NLP in oncology supports research by facilitating the extraction of insights from extensive biomedical literature. Advanced NLP techniques, such as Bidirectional Encoder Representations from Transformers (BERT), enable systematic analysis of vast literature, identifying emerging trends and novel associations in cancer research [34, 4, 22, 93, 36]. This capability accelerates innovation and deepens understanding of cancer biology and treatment efficacy.

The clinical applications of NLP and deep learning in oncology highlight their transformative potential in enhancing cancer care and research. By extracting and structuring critical data from unstructured texts, these technologies support decision-making through advanced knowledge discovery and improve clinical trial processes. They facilitate diagnostic and prognostic model development, optimize trial designs through automated analytics, and provide insights into treatment effectiveness. Successful implementation requires careful evaluation of performance, biases, and ethical implications to ensure reliability and safety in clinical settings [94, 95, 36, 34]. These advancements ultimately contribute to better patient outcomes and the ongoing evolution of precision oncology.

7.2 Integration of AI in Clinical Practice

The integration of artificial intelligence (AI) into routine clinical practice has fundamentally transformed healthcare delivery, enhancing diagnostic accuracy, treatment personalization, and operational efficiency. AI technologies, particularly those utilizing NLP and deep learning, play a crucial role in processing vast amounts of clinical data, thereby improving decision-making and patient outcomes [11]. AI facilitates the extraction and analysis of unstructured data from EHRs, enabling healthcare providers to derive actionable insights from complex medical narratives [26].

A significant area of AI impact is the automation of clinical documentation and coding. By employing machine learning algorithms and NLP techniques, AI systems can automatically extract pertinent medical information from clinical notes, alleviating the burden on healthcare professionals and enhancing the accuracy of medical records [84]. This automation improves healthcare delivery efficiency and ensures compliance with regulatory requirements by maintaining accurate patient records.

AI's role in personalized medicine is evident as AI-driven models analyze patient-specific data to tailor treatment plans. This approach is particularly beneficial in oncology, where AI systems can identify genetic mutations and tumor markers, facilitating the development of individualized regimens [9]. By leveraging AI's predictive capabilities, clinicians can make more informed decisions, ultimately improving patient outcomes and minimizing adverse drug events [29].

Furthermore, AI technologies enhance clinical decision support systems (CDSS), providing real-time insights and recommendations to healthcare providers. These systems utilize AI algorithms to analyze patient data and evidence-based guidelines, offering valuable support in diagnosing and treating complex conditions [11]. The integration of AI into CDSS has been shown to improve diagnostic accuracy and reduce variability in clinical practice, leading to more consistent and effective patient care.

AI is also integrated into healthcare operations, optimizing resource allocation and streamlining workflows. AI-driven analytics provide healthcare administrators with insights into patient flow, resource utilization, and operational efficiency, enabling strategic planning and management of healthcare facilities [40]. This operational integration contributes to the overall sustainability and effectiveness of healthcare systems.

The integration of AI into routine clinical practice signifies a paradigm shift in healthcare delivery, offering numerous advantages in efficiency, accuracy, and patient-centered care. As AI technologies continue to evolve, their integration into clinical practice is expected to significantly enhance decision-making processes, particularly through sophisticated NLP techniques that facilitate knowledge discovery from extensive biomedical literature. These advancements are anticipated to streamline the extraction of relevant information from clinical trial results and EHRs, thereby improving the efficiency of evidence-based medicine and fostering further innovations in healthcare and medical research [34, 96, 35, 45, 36].

8 Challenges and Future Directions

8.1 Data Availability and Quality

In biomedical text mining, data availability and quality are pivotal challenges affecting the performance and generalizability of relation extraction models. A significant issue is the scarcity of large, high-quality annotated datasets necessary for training models that can generalize across various biomedical contexts [17]. The inefficiency in capturing joint features between named entity recognition (NER) and relation classification (RC) tasks further hampers the models' ability to leverage shared information effectively [7]. The reliability of current biomedical relation extraction methods is often compromised by noisy training signals from distant supervision, leading to inaccurate training and poor generalization [54]. Additionally, the high computational costs of accurately representing overlapping triples pose challenges, particularly given the frequent overlap of entities and relationships in biomedical texts [49].

Benchmarks for evaluating relation extraction models often suffer from high train-test overlap and inconsistencies in data construction, which can result in misleading performance evaluations [97, 60]. The informal and unstructured nature of electronic health records (EHRs) further complicates the extraction of relevant interactions and relationships, leading to ambiguities and potential inaccuracies [31]. Over-parameterization of existing models also limits their deployment in resource-constrained environments [40]. Addressing these challenges requires the development of diverse datasets, improved annotation practices, and advanced computational techniques to enhance data quality and availability in biomedical text mining.

8.2 Model Interpretability and Ethical Concerns

In the biomedical domain, model interpretability and ethical concerns are critical when applying deep learning and natural language processing (NLP). The complexity of neural architectures presents significant interpretability challenges, as demonstrated by misclassifications in CoEx-Bert model experiments, which can undermine trust in AI-driven healthcare solutions [39]. Ethical concerns also arise from potential hallucinations in outputs generated by large language models, which can adversely affect patient outcomes [13]. The reliance on high-quality annotated datasets, as illustrated by the ClinLinker model, highlights ethical implications, especially in less-resourced languages or domains [28].

Error propagation remains a significant ethical concern, particularly when NER and relation extraction are performed separately [98]. This issue is exacerbated by the computational demands of fine-tuning deep architectures, which limit model accessibility and scalability in resource-constrained settings [41]. The complexity of social determinants of health (SDOH) annotations necessitates sophisticated models to capture intricate biomedical relationships accurately [25]. The limitations of current studies, such as small datasets, restrict generalizability and model performance [16]. Addressing these challenges requires comprehensive frameworks prioritizing transparency, fairness, and privacy in AI-driven healthcare systems, including effective de-identification of clinical data and leveraging NLP to extract critical health risk factors [84, 99].

8.3 Future Directions and Research Opportunities

Future research in biomedical text mining aims to explore several promising directions to enhance relation extraction methodologies. Integrating multimodal data could significantly improve model performance by providing a comprehensive understanding of biomedical contexts [10]. Developing joint learning techniques that emphasize attention mechanisms and consider relationships among all relatives is a promising avenue [98]. Enhancing domain-specific models and improving few-shot learning techniques are critical areas for advancing relation extraction in specialized biomedical domains [12]. Incorporating additional background knowledge from unsupervised data and utilizing ontology reasoners could enhance frameworks like ReOnto, providing robust relation extraction capabilities [61].

Future research should also focus on developing multilingual biomedical Named Entity Recognition (bNER) systems, improving annotation methods for nested entities, and enhancing model robustness to handle clinical data complexities [9]. Enhancements to the BiTT scheme, including exploring binary forests for result extraction and integrating advanced pre-trained encoders, are anticipated to improve the extraction of complex biomedical relationships [49]. Exploring advanced bag aggregation methods and applying relationship embedding architectures to general-domain datasets present additional research opportunities [54]. Future efforts will focus on validating generated data with clinical experts and exploring additional NLP tasks to address current methodologies' limitations [17].

Ongoing research directions underscore the necessity for continuous innovation, particularly in Literature Based Discovery (LBD), which automates the identification of novel associations among medical terms across disparate literature [4]. This dynamic field is driven by the increasing volume of scientific publications and the need for efficient information extraction techniques to mitigate medical errors and enhance clinical decision-making. Community challenges in biomedical text mining have emerged as vital platforms fostering interdisciplinary collaboration and technological advancements, highlighting the importance of leveraging large language models and deep learning techniques to address current limitations and explore future opportunities [100, 33]. By focusing on enhancing data analysis and knowledge discovery, future research will continue to drive advancements in this critical field.

9 Conclusion

The survey highlights the profound impact of advanced computational approaches, particularly deep learning and natural language processing, on the field of biomedical text mining. These technologies have significantly enhanced data analysis and knowledge discovery capabilities, with joint-learning models achieving notable improvements in relation extraction while maintaining accuracy in named entity recognition. The effectiveness of deep learning techniques is demonstrated by their superior generalization and performance compared to traditional methods. Furthermore, the integration of structured knowledge has proven to be highly beneficial, as evidenced by models that have enhanced biomedical named entity recognition across various contexts.

Frameworks have achieved strong performance metrics on key datasets, showcasing their effectiveness in biomedical relation extraction. Additionally, models with advanced summarization capabilities have successfully addressed the issue of information overload in clinical and biomedical environments. The potential of generative large language models to transform healthcare practices is evident, though further research is needed to refine techniques and explore additional applications that could benefit from a cohesive approach.

The survey underscores the importance of interdisciplinary collaboration and thorough validation to ensure the safety and efficacy of AI tools in clinical settings. As deep learning and natural language processing techniques continue to evolve, ongoing research is crucial to advancing data analysis, enhancing knowledge discovery, and ultimately improving healthcare outcomes. Future research should focus on incorporating domain-specific knowledge, developing sophisticated models, and exploring innovative methodologies to drive progress in this vital field.

References

[1] Usama Yaseen, Pankaj Gupta, and Hinrich Schütze. Linguistically informed relation extraction and neural architectures for nested named entity recognition in bionlp-ost 2019, 2019.

[2] Marco Basaldella, Lenz Furrer, Carlo Tasso, and Fabio Rinaldi. Entity recognition in the biomedical domain using a hybrid approach. *Journal of biomedical semantics*, 8:1–14, 2017.

[3] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models, 2019.

[4] Balu Bhasuran, Gurusamy Murugesan, and Jeyakumar Natarajan. Literature based discovery (lbd): Towards hypothesis generation and knowledge discovery in biomedical text mining, 2023.

[5] Mehmet Efruz Karabulut and K. Vijay-Shanker. Sectioning of biomedical abstracts: A sequence of sequence classification task, 2022.

[6] Surag Nair. A biomedical information extraction primer for nlp researchers, 2017.

[7] Kui Xue, Yangming Zhou, Zhiyuan Ma, Tong Ruan, Huanhuan Zhang, and Ping He. Fine-tuning bert for joint entity and relation extraction in chinese medical text, 2019.

[8] Kriti Bhattarai, Inez Y. Oh, Zachary B. Abrams, and Albert M. Lai. Document-level clinical entity and relation extraction via knowledge base-guided generation, 2024.

[9] Pir Noman Ahmad, Adnan Muhammad Shah, and KangYoon Lee. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare*, volume 11, page 1268. MDPI, 2023.

[10] Omid Rohanian, Mohammadmahdi Nouriborji, and David A. Clifton. Exploring the effectiveness of instruction tuning in biomedical language processing, 2023.

[11] Yuqing Wang, Yun Zhao, and Linda Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding, 2023.

[12] Xiaoyan Zhao, Yang Deng, Min Yang, Lingzhi Wang, Rui Zhang, Hong Cheng, Wai Lam, Ying Shen, and Ruifeng Xu. A comprehensive survey on relation extraction: Recent advances and new frontiers, 2024.

[13] Cheng Peng, Xi Yang, Aokun Chen, Zehao Yu, Kaleb E Smith, Anthony B Costa, Mona G Flores, Jiang Bian, and Yonghui Wu. Generative large language models are all-purpose text analytics engines: Text-to-text learning is all your need, 2023.

[14] Shantanu Kumar. A survey of deep learning methods for relation extraction, 2017.

[15] Vishwanath D, Rohit Rahul, Gunjan Sehgal, Swati, Arindam Chowdhury, Monika Sharma, Lovekesh Vig, Gautam Shroff, and Ashwin Srinivasan. Deep reader: Information extraction from document images via relation extraction and natural language, 2018.

[16] Sijia Zhou and Xin Li. Feature engineering vs. deep learning for paper section identification: Toward applications in chinese medical literature, 2024.

[17] Zixu Wang, Julia Ive, Sumithra Velupillai, and Lucia Specia. Is artificial data useful for biomedical natural language processing algorithms?, 2019.

[18] Hyejin Cho and Hyunju Lee. Biomedical named entity recognition using deep neural networks with contextual information. *BMC bioinformatics*, 20:1–11, 2019.

[19] Huiwei Zhou, Shixian Ning, Yunlong Yang, Zhuang Liu, Chengkun Lang, and Yingyu Lin. Chemical-induced disease relation extraction with dependency information and prior knowledge, 2020.

[20] Junyi Bian, Rongze Jiang, Weiqi Zhai, Tianyang Huang, Hong Zhou, and Shanfeng Zhu. Dmner: Biomedical entity recognition by detection and matching, 2023.

[21] Parisa Naderi Golshan, HosseinAli Rahmani Dashti, Shahrzad Azizi, and Leila Safari. A study of recent contributions on information extraction, 2018.

[22] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in cell and developmental biology*, 8:673, 2020.

[23] Yiming Li, Xueqing Peng, Jianfu Li, Xu Zuo, Suyuan Peng, Donghong Pei, Cui Tao, Hua Xu, and Na Hong. Relation extraction using large language models: A case study on acupuncture point locations, 2024.

[24] Priyankar Bose, Sriram Srinivasan, William C Sleeman IV, Jatinder Palta, Rishabh Kapoor, and Preetam Ghosh. A survey on recent named entity recognition and relationship extraction techniques on clinical texts. *Applied Sciences*, 11(18):8319, 2021.

[25] Manabu Torii, Ian M. Finn, Son Doan, Paul Wang, Elly W. Yang, and Daniel S. Zisook. Task formulation for extracting social determinants of health from clinical narratives, 2023.

[26] Jun-En Ding, Phan Nguyen Minh Thao, Wen-Chih Peng, Jian-Zhe Wang, Chun-Cheng Chug, Min-Chen Hsieh, Yun-Chien Tseng, Ling Chen, Dongsheng Luo, Chi-Te Wang, Pei fu Chen, Feng Liu, and Fang-Ming Hung. Large language multimodal models for 5-year chronic disease cohort prediction using ehr data, 2024.

[27] Jing Chen, Baotian Hu, Weihua Peng, Qingcai Chen, and Buzhou Tang. Biomedical relation extraction via knowledge-enhanced reading comprehension. *BMC bioinformatics*, 23(1):20, 2022.

[28] Fernando Gallego, Guillermo López-García, Luis Gasco-Sánchez, Martin Krallinger, and Francisco J. Veredas. Clinlinker: Medical entity linking of clinical concept mentions in spanish, 2024.

[29] Fenia Christopoulou, Thy Thy Tran, Sunil Kumar Sahu, Makoto Miwa, and Sophia Ananiadou. Adverse drug events and medication relation extraction in electronic health records with ensemble deep learning methods. *Journal of the American Medical Informatics Association*, 27(1):39–46, 2020.

[30] Shweta Yadav, Srivatsa Ramesh, Sriparna Saha, and Asif Ekbal. Relation extraction from biomedical and clinical text: Unified multitask learning framework. *IEEE/ACM transactions on computational biology and bioinformatics*, 19(2):1105–1116, 2020.

[31] Yujuan Fu, Giridhar Kaushik Ramachandran, Ahmad Halwani, Bridget T. McInnes, Fei Xia, Kevin Lybarger, Meliha Yetisgen, and Özlem Uzuner. Cacer: Clinical concept annotations for cancer events and relations, 2024.

[32] Shyni Sharaf and V. S. Anoop. An analysis on large language models in healthcare: A case study of biobert, 2023.

[33] Mohamed Yassine Landolsi, Lobna Hlaoua, and Lotfi Ben Romdhane. Information extraction from electronic medical documents: state of the art and future research directions. *Knowledge and Information Systems*, 65(2):463–516, 2023.

[34] Ting He, Kory Kreimeyer, Mimi Najjar, Jonathan Spiker, Maria Fatteh, Valsamo Anagnostou, and Taxiarchis Botsis. Ai-assisted knowledge discovery in biomedical literature to support decision-making in precision oncology, 2024.

[35] Hasham Ul Haq, Veysel Kocaman, and David Talby. Deeper clinical document understanding using relation extraction. *arXiv preprint arXiv:2112.13259*, 2021.

[36] Benjamin E. Nye, Jay DeYoung, Eric Lehman, Ani Nenkova, Iain J. Marshall, and Byron C. Wallace. Understanding clinical trial reports: Extracting medical entities and their relations, 2022.

[37] Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina Kuttichi Keloth, Kalpana Raja, Jiming Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, Zhiyong Lu, and Hua Xu. A systematic evaluation of large language models for biomedical natural language processing: benchmarks, baselines, and recommendations, 2024.

[38] Zaifu Zhan, Shuang Zhou, Mingchen Li, and Rui Zhang. Ramie: Retrieval-augmented multi-task information extraction with large language models on dietary supplements, 2024.

[39] Fan Lu, Quan Qi, and Huaibin Qin. Joint extraction of uyghur medicine knowledge with edge computing, 2024.

[40] Omid Rohanian, Mohammadmahdi Nouriborji, Hannah Jauncey, Samaneh Kouchaki, ISARIC Clinical Characterisation Group, Lei Clifton, Laura Merson, and David A. Clifton. Lightweight transformers for clinical natural language processing, 2023.

[41] Walid Hafiane, Joel Legrand, Yannick Toussaint, and Adrien Coulet. Experiments on transfer learning architectures for biomedical relation extraction, 2020.

[42] Yucong Lin, Keming Lu, Sheng Yu, Tianxi Cai, and Marinka Zitnik. Multimodal learning on graphs for disease relation extraction, 2022.

[43] Hang Dong, Matúš Falis, William Whiteley, Beatrice Alex, Joshua Matterson, Shaoxiong Ji, Jiaoyan Chen, and Honghan Wu. Automated clinical coding: What, why, and where we are?, 2022.

[44] Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. *arXiv preprint arXiv:2105.13456*, 2021.

[45] Maristella Agosti, Giorgio Maria Di Nunzio, Stefano Marchesin, and Gianmaria Silvello. A relation extraction approach for clinical decision support, 2019.

[46] Anfu Tang, Claire Nédellec, Pierre Zweigenbaum, Louise Deléger, and Robert Bossy. Global alignment for relation extraction in microbiology, 2021.

[47] Elham Shahab. A short survey of biomedical relation extraction techniques, 2017.

[48] Amelie Wührl and Roman Klinger. Recovering patient journeys: A corpus of biomedical entities and relations on twitter (bear), 2022.

[49] Xukun Luo, Weijie Liu, Meng Ma, and Ping Wang. A bidirectional tree tagging scheme for joint medical relation extraction, 2022.

[50] Trapit Bansal, Pat Verga, Neha Choudhary, and Andrew McCallum. Simultaneously linking entities and extracting relations from biomedical text without mention-level supervision, 2019.

[51] Giannis Bekoulis, Johannes Deleu, Thomas Demeester, and Chris Develder. Joint entity recognition and relation extraction as a multi-head selection problem, 2018.

[52] Diana Sousa and Francisco M. Couto. Biont: Deep learning using multiple biomedical ontologies for relation extraction, 2020.

[53] Saadullah Amin, Pasquale Minervini, David Chang, Pontus Stenetorp, and Günter Neumann. Meddistant19: Towards an accurate benchmark for broad-coverage biomedical relation extraction, 2022.

[54] William Hogan, Molly Huang, Yannis Katsis, Tyler Baldwin, Ho-Cheol Kim, Yoshiki Vazquez Baeza, Andrew Bartko, and Chun-Nan Hsu. Abstractified multi-instance learning (amil) for biomedical relation extraction, 2021.

[55] Jose A. Diaz-Garcia and Julio Amador Diaz Lopez. A survey on cutting-edge relation extraction techniques based on language models, 2024.

[56] Sachin Pawar, Pushpak Bhattacharyya, and Girish K. Palshikar. Techniques for jointly extracting entities and relations: A survey, 2021.

[57] Azad Dehghan. Temporal ordering of clinical events, 2015.

[58] Thien Huu Nguyen and Ralph Grishman. Combining neural networks and log-linear models to improve relation extraction, 2015.

[59] Dongling Li, Pengchao Wu, Yuehu Dong, Jinghang Gu, Longhua Qian, and Guodong Zhou. Joint learning-based causal relation extraction from biomedical literature, 2022.

[60] Elisa Bassignana and Barbara Plank. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification, 2022.

[61] Monika Jain, Kuldeep Singh, and Raghava Mutharaju. Reonto: A neuro-symbolic approach for biomedical relation extraction, 2023.

[62] Hailin Wang, Ke Qin, Rufai Yusuf Zakari, Guoming Lu, and Jin Yin. Deep neural network based relation extraction: An overview, 2021.

[63] Yongkang Li. An empirical study on relation extraction in the biomedical domain, 2021.

[64] Yan Hu, Xu Zuo, Yujia Zhou, Xueqing Peng, Jimin Huang, Vipina K. Keloth, Vincent J. Zhang, Ruey-Ling Weng, Qingyu Chen, Xiaoqian Jiang, Kirk E. Roberts, and Hua Xu. Information extraction from clinical notes: Are we ready to switch to large language models?, 2025.

[65] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021.

[66] Hong Guan, Jianfu Li, Hua Xu, and Murthy Devarakonda. Robustly pre-trained neural model for direct temporal relation extraction, 2020.

[67] Kathleen C. Fraser, Isar Nejadgholi, Berry De Bruijn, Muqun Li, Astha LaPlante, and Khaldoun Zine El Abidine. Extracting umls concepts from medical text using general and domain-specific deep learning models, 2019.

[68] Young-Min Kim and Tae-Hoon Lee. Korean clinical entity recognition from diagnosis text using bert. *BMC Medical Informatics and Decision Making*, 20:1–9, 2020.

[69] Shogo Ujiie, Hayate Iso, Shuntaro Yada, Shoko Wakamiya, and Eiji Aramaki. End-to-end biomedical entity linking with span-based dictionary matching, 2021.

[70] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. Biored: A rich biomedical relation extraction dataset, 2022.

[71] Hasham Ul Haq, Veysel Kocaman, and David Talby. Deeper clinical document understanding using relation extraction, 2021.

[72] Yuanhao Shen and Jungang Han. Joint extraction of entity and relation with information redundancy elimination, 2020.

[73] Monika Jain, Raghava Mutharaju, Kuldeep Singh, and Ramakanth Kavuluru. Knowledge-driven cross-document relation extraction, 2024.

[74] Peng Su and K. Vijay-Shanker. Investigation of bert model on biomedical relation extraction based on revised fine-tuning mechanism, 2020.

[75] Farshad Noravesh. Bioncere: Non-contrastive enhancement for relation extraction in biomedical texts, 2024.

[76] Xi Yang, Zehao Yu, Yi Guo, Jiang Bian, and Yonghui Wu. Clinical relation extraction using transformer-based models, 2021.

AI-generated, for reference only.

[77] Xi Yang, Aokun Chen, Nima PourNejatian, Hoo Chang Shin, Kaleb E Smith, Christopher Parisien, Colin Compas, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, Christopher A Harle, Gloria Lipori, Duane A Mitchell, William R Hogan, Elizabeth A Shenkman, Jiang Bian, and Yonghui Wu. Gatortron: A large clinical language model to unlock patient information from unstructured electronic health records, 2022.

[78] Louis Falissard, Claire Morgand, Sylvie Roussel, Claire Imbaud, Walid Ghosn, Karim Bounebache, and Grégoire Rey. Neural translation and automated recognition of icd10 medical entities from natural language, 2020.

[79] Darshini Mahendran and Bridget T. McInnes. Extracting adverse drug events from clinical notes, 2021.

[80] Yifan Peng, Qingyu Chen, and Zhiyong Lu. An empirical study of multi-task learning on bert for biomedical text mining, 2020.

[81] Enshuo Hsu and Kirk Roberts. Llm-ie: A python package for generative information extraction with large language models, 2024.

[82] Songchi Zhou and Sheng Yu. High-throughput biomedical relation extraction for semi-structured web articles empowered by large language models, 2024.

[83] Sonit Singh. Natural language processing for information extraction, 2018.

[84] Xavier Tannier, Perceval Wajsbürt, Alice Calliger, Basile Dura, Alexandre Mouchet, Martin Hilka, and Romain Bey. Development and validation of a natural language processing algorithm to pseudonymize documents in the context of a clinical data warehouse, 2023.

[85] Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference, 2021.

[86] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors, 2022.

[87] Amarin Jettakul, Duangdao Wichadakul, and Peerapon Vateekul. Relation extraction between bacteria and biotopes from biomedical texts with attention mechanisms and domain-specific contextual representations. *BMC bioinformatics*, 20:1–17, 2019.

[88] Shaina Raza and Brian Schwartz. Entity and relation extraction from clinical case reports of covid-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, 23(1):20, 2023.

[89] Saadullah Amin, Katherine Ann Dunfield, Anna Vechkaeva, and Günter Neumann. A data-driven approach for noise reduction in distantly supervised biomedical relation extraction, 2020.

[90] Chunhui Du, Hao He, and Yaohui Jin. Contrast with major classifier vectors for federated medical relation extraction with heterogeneous label distribution, 2023.

[91] Paul Barry, Sam Henry, Meliha Yetisgen, Bridget McInnes, and Ozlem Uzuner. Jointly learning clinical entities and relations with contextual language models and explicit context, 2021.

[92] Sarah Schulz, Jurica Ševa, Samuel Rodriguez, Malte Ostendorff, and Georg Rehm. Named entities in medical case reports: Corpus and experiments, 2020.

[93] Eva K Lee and Karan Uppal. Cerc: an interactive content extraction, recognition, and construction tool for clinical and biomedical text. *BMC medical informatics and decision making*, 20:1–14, 2020.

[94] Reza Khanmohammadi, Mohammad M. Ghassemi, Kyle Verdecchia, Ahmed I. Ghanem, Luo Bing, Indrin J. Chetty, Hassan Bagher-Ebadian, Farzan Siddiqui, Mohamed Elshaikh, Benjamin Movsas, and Kundan Thind. An introduction to natural language processing techniques and framework for clinical implementation in radiation oncology, 2023.

[95] Miao Chen, Ganhui Lan, Fang Du, and Victor Lobanov. Joint learning with pre-trained transformer on named entity recognition and relation extraction tasks for clinical analytics. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 234–242, 2020.

[96] Jetsun Whitton and Anthony Hunter. Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations, 2021.

[97] Haiyun Jiang, Qiaoben Bao, Qiao Cheng, Deqing Yang, Li Wang, and Yanghua Xiao. Complex relation extraction: Challenges and opportunities, 2020.

[98] Xue Shi, Dehuan Jiang, Yuanhang Huang, Xiaolong Wang, Qingcai Chen, Jun Yan, and Buzhou Tang. Family history information extraction via deep joint learning. *BMC medical informatics and decision making*, 19:1–6, 2019.

[99] Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Ozlem Ozuner, and Meliha Yetisgen. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record, 2023.

[100] Hui Zong, Rongrong Wu, Jiaxue Cha, Weizhe Feng, Erman Wu, Jiakun Li, Aibin Shao, Liang Tao, Zuofeng Li, Buzhou Tang, and Bairong Shen. Advancing chinese biomedical text mining with community challenges, 2024.

AI-generated, for reference only.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.