ISSN: 2755-9866

Journal of Health Statistics Reports

SCIENTIFIC Research and Community

Review Article

Open Access

A Review on Building Knowledge Graphs from Scanned Textual Data in Medical Literature for Structured Insights

Syed Arham Akheel

Senior Solutions Architect Bellevue, WA

ABSTRACT

The exponential growth of medical literature presents challenges in extracting, organizing, and utilizing valuable information for healthcare applications. This paper presents a comprehensive review of methodologies for building knowledge graphs (KGs) from scanned medical literature, integrating Optical Character Recognition (OCR) and Natural Language Processing (NLP) techniques. The review addresses the challenges of low-quality scans, domain-specific terminology, data heterogeneity, and entity ambiguity. Furthermore, the paper explores various use cases in healthcare, such as clinical decision support and drug discovery, demonstrating the transformative potential of structured knowledge extraction.

***Corresponding author**

Syed Arham Akheel, Senior Solutions Architect Bellevue, WA.

Received: August 02, 2023; **Accepted:** August 09, 2023; **Published:** August 20, 2023

Keywords: Knowledge Graphs, Optical Character Recognition, Natural Language Processing, Medical Literature, Entity Extraction, Healthcare Informatics

Introduction

The exponential increase in medical literature over recent years has presented significant challenges in data extraction, organization, and utilization for healthcare applications. This vast body of information exists across unstructured formats, including scanned journal articles, clinical records, and handwritten physician notes. These diverse document types often result in data that is difficult to search, query, and integrate effectively into healthcare systems [1]. As the volume of medical literature continues to grow, efficient methods for processing and leveraging this knowledge are becoming critical.

Knowledge Graphs (KGs) provide a structured framework to represent relationships between different medical entities, such as diseases, symptoms, and treatments. This structured representation facilitates advanced search capabilities, data interoperability, and the ability to derive new insights through inferencing [2]. In healthcare, KGs have been particularly beneficial for applications like clinical decision support, where they help to integrate Electronic Health Records (EHRs), biomedical literature, and patient data to assist healthcare professionals in making informed decisions [3]. Furthermore, industry-scale KGs developed by technology giants such as Google and Microsoft illustrate the potential of integrating medical and general knowledge to provide comprehensive answers and insights [4].

Graph database models are particularly useful for storing and representing medical knowledge graphs due to their ability to efficiently manage interconnected data, allowing for rapid traversal of relationships among medical entities. As discussed in, graph databases provide an effective model for data whose relationships are as important as the data itself [5]. This is especially relevant for healthcare, where understanding relationships between symptoms, treatments, and patient history can have significant implications for clinical outcomes. The Neo4j graph database, one of the leading graph database technologies, has been used in various applications including health, government, and social network analysis, demonstrating its capability to handle large-scale, highly interconnected data [6].

Despite these advantages, constructing KGs from scanned textual data presents substantial challenges. One primary obstacle is the inherent complexity of medical language, which includes abbreviations, jargon, and polysemous terms that require specialized processing techniques. Furthermore, the quality of scanned medical documents varies significantly, especially when dealing with handwritten notes or poorly scanned images, which hinders Optical Character Recognition (OCR) accuracy [7]. In particular, conventional OCR tools like Tesseract struggle with these challenges, making it necessary to explore advanced machine learning techniques such as Multidimensional Recurrent Neural Networks (MDRNN) and Connectionist Temporal Classification (CTC) to improve recognition rates, especially for handwritten documents [7].

Natural Language Processing (NLP) plays a critical role in converting extracted text into a format suitable for knowledge graph construction. NLP techniques such as Named Entity Recognition (NER), relation extraction, and entity linking are used to identify and classify medical concepts from unstructured text. Domain-specific models like BioBERT and ClinicalBERT have been particularly effective in processing medical literature, outperforming general NLP models due to their focus on biomedical corpora [8]. Tools such as cTAKES and MetaMap have also been successfully employed to extract clinical information, demonstrating the adaptability of NLP methods for different types of medical text [9].

Previous studies have also attempted to integrate OCR and NLP to automate the extraction of structured information from medical literature. For example, the 2010 i2b2/VA challenge illustrated the feasibility of combining OCR and NLP techniques for extracting

medical concepts and identifying relationships between these concepts within clinical narratives [10]. However, accurately capturing the relationships between entities in heterogeneous and often noisy content remains a significant challenge.

This paper reviews existing methodologies for constructing knowledge graphs from scanned medical literature, integrating OCR and NLP techniques to overcome the key challenges posed by unstructured medical data. We focus on recent advances in these areas and highlight the challenges that need to be addressed to make the process more efficient and reliable. By addressing these issues, we aim to create more effective knowledge graphs that can significantly impact various areas of healthcare, such as clinical decision support, drug discovery, and medical education.

Literature Review

Knowledge Graphs in Healthcare

Knowledge graphs have become increasingly important in healthcare, providing structured ways to represent biomedical data, facilitate advanced queries, and support decision-making processes. They have been employed in various applications, such as integrating Electronic Health Records (EHRs) and representing adverse drug events (ADEs) in structured formats like ADEpedia [3]. Additionally, KGs developed by technology giants such as Google and Microsoft have shown the potential to integrate general and medical knowledge for broader applications [4]. KGs also serve as valuable tools in enhancing healthcare analytics, enabling researchers to uncover complex interactions between medical entities, such as drugs and diseases, which are otherwise difficult to identify using traditional data management techniques [5].

Graph databases are commonly used for storing and managing the data required for healthcare knowledge graphs. The use of graph databases like Neo4j allows efficient storage and retrieval of interconnected data, which is crucial for representing complex medical knowledge [6]. Neo4j has been widely adopted in healthcare applications due to its ability to efficiently handle large-scale, highly connected datasets, enabling the effective analysis of relationships among various medical entities. This capability has proven beneficial for use cases involving patient records, treatment pathways, and drug interactions, making it a vital component in modern healthcare informatics [6].

Optical Character Recognition (OCR) Technologies

Optical Character Recognition has been widely adopted for digitizing scanned medical documents. While commercial tools like Tesseract, Google Vision, and Amazon Textract have been effective for printed documents, OCR accuracy tends to degrade significantly for handwritten or low-quality scans [1]. Recent advances in Multidimensional Recurrent Neural Networks (MDRNN) and Connectionist Temporal Classification (CTC) have demonstrated potential for enhancing OCR performance, particularly in recognizing complex handwritten documents [7]. Furthermore, the integration of graph-based representations with OCR systems has been explored to improve the accuracy of text recognition from scanned medical documents. By leveraging graph structures, relationships between different textual components can be captured, which can significantly enhance OCR performance in cases involving complex page layouts [5].

Natural Language Processing (NLP) in Medicine

The application of NLP to medical literature includes Named Entity Recognition (NER), relation extraction, and entity linking, which are crucial for transforming unstructured medical text into a structured form suitable for knowledge graphs. BioBERT and ClinicalBERT are examples of domainspecific language models that have outperformed traditional NLP models in extracting entities from clinical narratives [8]. Tools like cTAKES and MetaMap have also been employed for clinical information extraction, showcasing the adaptability of NLP techniques to diverse medical texts [9]. Graph databases have also been used to store the outputs of NLP systems, particularly for managing the complex relationships between extracted entities [6]. The use of graph databases facilitates the representation of extracted medical entities and their interrelationships, enabling more advanced querying and inferencing capabilities that are essential for healthcare applications.

Integration of OCR and NLP

Previous studies have explored the integration of OCR and NLP for automating data extraction from scanned clinical documents. The 2010 i2b2/VA challenge, for instance, demonstrated the feasibility of combining NLP systems with OCR to extract medical concepts and classify relations, leading to improvements in structured knowledge representation [7]. However, challenges remain in effectively handling the heterogeneity of document types and accurately capturing relationships in noisy or handwritten content. The integration of graph databases into the OCR-NLP pipeline has been proposed as a way to enhance the management of extracted data, allowing for better handling of the relationships between entities extracted from scanned medical documents [5].

How Graph Databases Work

Graph databases are a type of NoSQL database that uses graph structures for semantic queries, with nodes, edges (relationships), and properties to represent and store data. This approach makes them ideal for managing highly interconnected data and understanding complex relationships within datasets, particularly in domains such as healthcare, social networks, and recommendation systems [5].

Core Components of Graph Databases

The fundamental components of graph databases include:

- **Nodes:** Nodes represent entities or objects, analogous to rows in a relational database. For example, in healthcare, nodes might represent patients, doctors, diseases, or treatments.
- **Relationships (Edges):** Relationships, also called edges, connect nodes and represent how they interact. Relationships are directional and indicate the nature of the interaction between entities. For instance, a relationship can describe that a doctor "treats" a patient.
- **Properties:** Nodes and relationships have associated properties, similar to attributes or columns in a relational database. For example, a "Patient" node could have properties like name, age, and medical history, while a "treats" relationship might include treatment date.
- **Labels:** Labels are used to group nodes into categories, making querying more efficient. For example, nodes with the label "Doctor" might share attributes specific to medical professionals [6].

Graph Traversal and Querying

The key strength of graph databases is their ability to efficiently traverse the graph to uncover patterns and relationships between nodes. Traversal is the process of visiting nodes and edges in a graph, allowing for complex queries that can explore multiple degrees of relationships without the need for computationally expensive join operations, as in traditional relational databases [5].

Traversal uses a pathfinding mechanism to explore nodes, commonly employing algorithms such as:

- **Breadth-First Search (BFS):** BFS starts at a given node and explores all its neighbors before moving to the next level. This algorithm is useful for finding the shortest path in terms of the number of edges.
- **Depth-First Search (DFS):** DFS starts at a given node and follows a path as deep as possible before backtracking. This algorithm is suitable for discovering long chains of relationships.

The graph traversal can be expressed mathematically through matrix operations. Consider an adjacency matrix representation of a graph A, where Aij represents an edge between nodes i and j. The traversal through k steps can be calculated using the matrix power A^k, where each element (i,j) of A^k indicates the number of possible paths of length k between nodes i and j [5].

Graph Query Languages

To interact with graph databases, specialized query languages are used, such as:

- **Cypher:** Cypher is the query language used by Neo4j and is designed to be easy to read and write, making it intuitive for describing patterns of nodes and relationships [6]. An example query to find all patients treated by a particular doctor might look like:
- **Gremlin:** Gremlin is used by Apache TinkerPop and allows for graph traversal, enabling developers to navigate graph paths, filter nodes, and project query results. It provides a functional approach to describe how the graph should be traversed [5].

Mathematical Representation

Graph databases can be described using graph theory concepts. A graph G can be defined as:

$$G = (V,E) \qquad (1)$$

where V represents the set of nodes (vertices), and E represents the set of edges (relationships) between nodes. Each edge e ∈ E is a pair (v_i, v_j), where $v_i, v_j \in V$.

To model relationships mathematically, graph databases often utilize adjacency matrices or incidence matrices:

- **Adjacency Matrix (A):** For a graph with n nodes, an adjacency matrix A is an n × n matrix where:

(
1 if there is an edge from node i to node j A_{ij} =
0 otherwise
The adjacency matrix helps in determining direct connections and is fundamental in graph traversal algorithms like
BFS and DFS.

- **Graph Traversal Using Adjacency Matrix:** Traversing the graph over multiple steps can be computed using powers of the adjacency matrix. For example, A2 represents all possible two-step paths between nodes. This approach is used in various graph traversal algorithms to calculate path reachability and discover relationships across multiple hops.

Benefits of Graph Databases

Graph databases are particularly advantageous when dealing with highly interconnected data, due to their ability to:

- **Efficient Relationship Management:** Unlike relational databases that require expensive join operations, graph databases can traverse relationships directly, making queries across connected data much faster [5].
- **Schema Flexibility:** Graph databases are schema optional, which means that they can adapt to changes more easily compared to relational databases, where altering the schema can be complex and time-consuming [6].
- **Natural Representation of Real-World Data:** Many real-world datasets, such as social networks and healthcare systems, naturally fit into graph models where relationships between entities are critical. This makes graph data structures make them a powerful tool for applications that require an understanding of relationships among multiple entities [6].

MATCH (d:Doctor {name: “Dr. Smith”})- databases an intuitive choice for applications that require [:TREATS]->(p:Patient) complex relationship analysis.

RETURN p.name Graph databases like Neo4j have been widely adopted in various fields, including healthcare, social networks, and This query matches all nodes labeled as ”Doctor” with fraud detection, due to their ability to efficiently manage the name ”Dr. Smith” and traverses the ”TREATS” rela- and query highly connected data. Their flexibility, speed in tionship to return the names of patients. traversing relationships, and suitability for storing complex

Key challenges

Low-Quality Scans and Handwriting

Medical literature often includes low-resolution scans and handwritten notes, which hinder OCR accuracy. Techniques such as image enhancement, noise reduction, and specialized neural networks like Multidimensional Recurrent Neural Networks (MDRNN) have been developed to address these challenges [7]. Image preprocessing methods, such as binarization, contrast adjustment, and skew correction, have also shown promise in improving the quality of scanned images before OCR processing [1]. Nonetheless, further advancements are needed to improve robustness against variations in handwriting styles and poor image quality, which remain significant barriers to high-accuracy OCR.

Recent research has also focused on incorporating machine learning models, such as Convolutional Neural Networks (CNNs), to enhance OCR capabilities for complex documents with handwritten annotations. However, the variability in handwriting, including differences in language, style, and character spacing, presents an ongoing challenge [7]. To handle these issues, hybrid models combining CNNs with Recurrent Neural Networks (RNNs) have been explored, providing some improvement but still facing challenges in generalizing across diverse handwriting samples.

Domain-Specific Terminology

Medical terminology is highly complex, involving numerous abbreviations, jargon, and rarely used terms. Generic OCR and NLP models struggle with this specificity, necessitating the use of domain-specific resources like UMLS (Unified Medical Language System) to improve entity recognition and relationship extraction accuracy [3]. Domain-specific NLP models, such as BioBERT and ClinicalBERT, have been trained on biomedical corpora and have demonstrated superior performance in understanding complex medical terms [8]. However, these models are often limited by the diversity of medical sub-domains, making it essential to further train or fine-tune models on specific datasets.

Additionally, medical documents may contain multiple languages or even mixed-language content, which complicates entity recognition further. Tools like cTAKES and MetaMap have been employed to

address some of these challenges, but their performance may be inconsistent when dealing with multilingual documents or specialized sub-domains of medicine [9].

Data Heterogeneity

The diversity of medical documents, which include tables, charts, handwritten annotations, and text in various languages, complicates data extraction. Each type of content requires a different processing strategy, making it challenging to develop a unified pipeline for knowledge graph construction. For

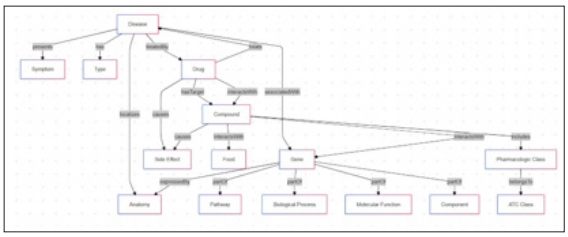

Figure 1: A Sample Knowledge Graph for Healthcare

instance, tables and charts often need specialized extraction methods, such as table detection algorithms and cell parsing techniques, to accurately extract structured data from scanned documents [3].

Furthermore, handwritten notes and annotations require different OCR approaches compared to printed text, and mixedmedia documents complicate the segmentation process. To tackle data heterogeneity, several hybrid approaches have been developed, which integrate computer vision techniques for layout detection along with NLP tools for text extraction. Graph-based representations have also been used to model document structures, thereby enabling better segmentation and extraction of heterogeneous content [9].

Entity and Relationship Extraction

Accurately extracting entities and relationships is a complex process that is often affected by the nuances of medical language, including polysemy (the existence of multiple meanings for a word) and synonymy (different words having the same meaning). These nuances can lead to ambiguity in identifying entities and understanding relationships within medical texts. Dependency parsing and rule-based extraction methods have shown promise in addressing some of these issues, but they often require substantial adaptation for medical content [6].

Pre-trained models such as SciBERT and BioGPT have been employed to perform entity extraction and relationship identification in the medical domain with considerable success. These models leverage contextual information to disambiguate entities and derive relationships from unstructured text. However, a key challenge lies in the need for large amounts of annotated medical data for effective model training, which is often difficult to obtain due to privacy concerns and the specialized nature of medical information [8].

In addition to extraction, linking entities to canonical knowledge bases such as UMLS or SNOMED CT is challenging due to inconsistencies in terminology and overlapping meanings. Effective entity linking requires sophisticated disambiguation techniques and integration of external ontologies to ensure the accuracy and completeness of the resulting knowledge graphs [3]. Incorporating graph databases to store these entities and relationships has helped manage the complexity of linking and querying data, enabling more efficient traversal of relationships among medical concepts [6].

Use Cases in Healthcare

Clinical Decision Support Systems

Knowledge graphs can enhance clinical decision-making by linking symptoms, diagnoses, treatments, and patient history in a structured format. By representing this data in a way that is easy to query and traverse, healthcare professionals can quickly access relevant information and derive insights to guide patient care. Such applications enable clinicians to make evidence-based decisions more efficiently and effectively, reducing the risk of medical errors and improving patient outcomes [2]. Additionally, KGs have been used to combine EHR data with guidelines from medical literature, thereby supporting personalized treatment plans and facilitating decision-making in complex cases [9]. For example, integrating patient-specific data with public health databases has helped identify potential risks for patients based on the presence of co-morbidities and recommended treatments [5].

Drug Discovery and Research

The use of knowledge graphs in drug discovery allows researchers to uncover non-obvious relationships between genes, proteins, and diseases, thereby identifying novel therapeutic targets. KGs facilitate the integration of diverse datasets, including genetic, proteomic, and clinical data, enabling researchers to form new hypotheses about disease mechanisms and potential drug candidates [3]. ADEpedia, for instance, has demonstrated the utility of KGs in managing adverse drug events through structured data integration [3]. By connecting information about drugs, their known side effects, and patient specific data, researchers can better understand drug safety profiles and identify potential adverse effects that may not have been detected in clinical trials. The integration of graph based representations of molecular interactions also allows for enhanced visualization of complex biological processes, supporting hypothesis generation in early-stage drug discovery [2].

Medical Education

Interactive knowledge graphs can be used as educational tools, providing medical students with dynamic ways to explore disease interactions, drug mechanisms, and treatment protocols. KGs enable visualization of intricate relationships between different medical entities, which is particularly useful for understanding multifaceted concepts in pathology and pharmacology [2]. Such visualizations enhance understanding and retention of complex medical concepts by presenting information in a contextually rich format that aids in pattern recognition and the comprehension of causal relationships. Additionally, KGs can serve as an interactive resource that medical educators use to demonstrate real-world applications, such as how different treatments interact with various biological pathways [5]. For example, the Neo4j graph database has been used to create interactive educational tools that allow users to explore relationships between diseases, symptoms, and treatments in a visual and intuitive manner [6].

Pandemic Management

During pandemics, knowledge graphs can map disease transmission pathways, identify high-risk areas, and inform intervention strategies. The ability of KGs to integrate and analyze diverse data sources, including epidemiological data, healthcare resources, and demographic information, enables public health officials to make informed decisions regarding containment and mitigation efforts [2]. For instance, during the COVID-19 pandemic, KGs were utilized to understand the relationships between patient demographics, pre-existing conditions, and disease severity, helping in risk stratification and resource allocation. By integrating data from public health databases, KGs can assist in predicting outbreak hotspots and

optimizing the distribution of healthcare resources, such as vaccines and medical supplies [5]. Furthermore, the use of KGs in pandemic management helps identify knowledge gaps in existing data, guiding research efforts to better understand the disease and its impact on various populations [4].

Conclusion

Integrating OCR and NLP technologies for constructing knowledge graphs from scanned medical literature offers significant potential for enhancing healthcare research, education, and practice. Addressing challenges such as low-quality scans, domain-specific terminology, and data heterogeneity is crucial for developing robust, actionable knowledge graphs. Improvements in OCR accuracy, particularly for low-quality and handwritten documents, are essential for ensuring that medical data is digitized accurately, allowing for reliable downstream processing [7].

The integration of domain-specific NLP models, such as BioBERT and ClinicalBERT, plays a critical role in enabling more effective extraction of entities and relationships from complex medical texts [8]. These models outperform traditional NLP methods when applied to healthcare settings due to their focus on biomedical terminology. However, there remains a significant need for further fine-tuning of these models across various sub-domains of medicine to ensure comprehensive coverage and accuracy.

Another significant challenge lies in handling the heterogeneity of medical data, which comes in multiple formats such as tables, charts, and free text, often mixed within the same document. Developing comprehensive extraction pipelines capable of handling this heterogeneity will be crucial to advancing the quality and usefulness of the resulting knowledge graphs [1]. Tailoring NLP and OCR technologies to process and unify these diverse data types will facilitate more coherent and reliable knowledge representation.

Knowledge graphs themselves offer immense potential in several healthcare use cases. They have shown their ability to enhance clinical decision-making by linking and visualizing relationships among patient data, symptoms, and treatments, thereby improving clinical workflow and patient outcomes [2]. KGs are also increasingly used in drug discovery to uncover non-obvious relationships between biological entities, paving the way for novel therapeutic approaches [3]. During pandemics, KGs have proven to be valuable tools in understanding disease transmission patterns and aiding public health officials in their decision-making processes [4].

Future research should focus on enhancing the capabilities of OCR and NLP systems to deal with handwritten and multilingual content, which is common in medical records globally. Additionally, implementing real-time knowledge graphs, which update automatically as new data becomes available, would ensure that healthcare professionals have access to the most current information, thereby supporting timely and informed decision-making [2]. Graph databases such as Neo4j provide a robust foundation for storing and managing this dynamic and interconnected medical data, enabling more advanced querying and inferencing capabilities [6].

Furthermore, addressing the explainability of AI-based systems remains critical to fostering trust among healthcare practitioners. Enhanced transparency in entity extraction and relationship modeling will enable healthcare professionals to better understand the processes behind the construction of knowledge graphs and validate their correctness [5]. As these technologies evolve, their application in medical informatics is expected to drive significant improvements in clinical decision making, education, and healthcare delivery overall. Future research should focus on developing OCR tools that can accurately recognize handwritten and multilingual content, expanding the applicability of these systems in diverse medical settings [7]. Implementing automated, real-time updates to knowledge graphs as new data becomes available would ensure that healthcare professionals always have access to the latest insights [2]. To foster trust in AI-based systems, it is essential to enhance the explainability of entity extraction and relationship modeling processes. This will allow clinicians to better understand and validate the knowledge represented in the graphs [2].

References

1. Kostrinsky-Thomas AL (2021) Searching the PDF Haystack: Automated Knowledge Discovery in Scanned EHR Documents. Appl Clin Inform 12: 245-250.
2. Ji S, Pan S, Cambria E, Marttinen P, Yu PS (2021) A Survey on Knowledge Graphs: Representation, Acquisition, and Applications. IEEE Trans Neural Netw Learn Syst.
3. Jiang G, Solbrig H, Chute C, Liu H (2019) ADEpedia: A Scalable and Standardized Knowledge Base of Adverse Drug Events. Department of Health Sciences Research, Mayo Clinic.
4. Noy N, Gao Y, Jain A, Narayanan A, Patterson A, et al. (2019) Industry-scale Knowledge Graphs: Lessons and Challenges. ACM Queue 1-28.
5. Angles R, Gutierrez C (2017) Survey of Graph Database Models. ACM Computing Surveys.
6. Guia J, Soares VG, Bernardino J (2017) Graph Databases: Neo4j Analysis. In Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) 351-356.
7. Graves A (2018) Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. TU Munich.
8. Alsentzer E (2019) Publicly Available Clinical BERT Embeddings. MIT CSAIL.
9. Wang Y, Wang L, Rastegar-Mojarad M, Shen F, Liu S, et al. (2018) Clinical Information Extraction Applications: A Literature Review. J Biomed Inform 77: 34-49.
10. (2011) 2010 i2b2/VA Challenge on Concepts, Assertions, and Relations in Clinical Text. J Am Med Inform Assoc 18: 552-556.