



Accelerating Medical Knowledge Discovery Through Automated Knowledge Graph Generation and Enrichment

Mutahira Khalid¹, Raihana Rahman², Asim Abbas³, Sushama Kumari⁴, Iram Wajahat⁵, and Syed Ahmad Chan Bukhari⁶(✉)

¹ TIB - Leibniz Information Centre for Science and Technology, 30165 Hannover, Lower Saxony, Germany
Mutahira.Khalid@tib.eu

² Department of Statistics, Harvard University, Cambridge, MA, USA
rrahman@college.harvard.edu

³ School of Computer Science, University of Birmingham, Edgbaston, Birmingham, UK
axa2233@student.bham.ac.uk

⁴ Routing and Planning, Amazon, Bellevue, WA, USA

⁵ Institute for Biotechnology, St. John's University, New York, USA
wajahati@stjohns.edu

⁶ Division of Computer Science, Mathematics and Science, St. John's University, New York, USA
bukharis@stjohns.edu

Abstract. Knowledge graphs (KGs) serve as powerful tools for organizing and representing structured knowledge. While their utility is widely recognized, challenges persist in their automation and completeness. Despite efforts in automation and the utilization of expert-created ontologies, gaps in connectivity remain prevalent within KGs. In response to these challenges, we propose an innovative approach termed “Medical Knowledge Graph Automation (M-KGA)”. M-KGA leverages user-provided medical concepts and enriches them semantically using Bio-Portal ontologies, thereby enhancing the completeness of knowledge graphs through the integration of pre-trained embeddings. Our approach introduces two distinct methodologies for uncovering hidden connections within the knowledge graph: a cluster-based approach and a node-based approach. Through rigorous testing involving 300 frequently occurring medical concepts in Electronic Health Records (EHRs), our M-KGA framework demonstrates promising results obtained 50% accuracy, 57% F1-score, 50% recall and 65% precision on a cluster base approach. Similarly, we achieve 85% accuracy, 87% F1-score, 89% recall and 88% precision on a node base approach indicating its potential to address the limitations of existing knowledge graph automation techniques.

Keywords: Knowledge Graphs · Ontologies · Healthcare · Decision Support systems

1 Introduction

While once considered a relic of early Artificial Intelligence (AI) research [1], knowledge graphs (KGs) have experienced a remarkable resurgence in recent years. Knowledge graphs, which serve as the foundation of symbolic AI, consist of interconnected knowledge pertaining to many domains like medical, finance, commerce, and education [2]. Particularly in medicine, KGs have emerged as indispensable tool.

KGs offer numerous advantages over traditional relational databases, primarily stemming from their diverse array of nodes and the ability to establish connections between them. This versatility lends itself to applications ranging from search engine optimization to recommendation systems, knowledge discovery, and research facilitation. However, the process of constructing KGs is inherently labor-intensive, especially in the intricate domain of medicine, despite its profound significance. At its core, a KG comprises facts, often represented as triplets, each consisting of a relationship and two nodes. With KGs consisting of millions to billions of these triplets, their aggregation holds immense potential for information discovery, data integration, and effective management. Yet, crafting KGs, particularly within the medical domain, presents formidable challenges due to the complexity of medical concepts and relationships between them. Compounding these challenges is the prevalence of unstructured medical data, further complicating the KG creation process.

Various methodologies for graph creation have emerged in recent years, ranging from automatic to semi-automated and manual approaches [3]. While these methods addresses some challenges. However they often suffer from significant deficiencies such as they lack standardized platforms or code for graph creation despite offering graphical methodologies. Furthermore, some approaches utilize hospital notes to generate nodes and relations, they overlook the potential benefits of data augmentation, resulting in incomplete graphs. Additionally, there is currently no promising technology capable of generating graphs in real-time, further impeding the process.

In response to these challenges, our proposed approach, Medical Knowledge Graph Automation (M-KGA), effectively addresses these obstacles by seamlessly processing both structured and unstructured data in real-time. The preprocessing steps involve Named Entity Recognition (NER)-based keyword extraction from unstructured data using the SciSpacy library [4], tailored for scientific and biomedical content. Subsequently, a knowledge filtration phase eliminates duplicates and extraneous terms before rapidly generating the knowledge graph using Neo4j's query language, 'Cypher'. Furthermore, we leverage Bioportal [5] for data augmentation, enriching medical terms semantically via incorporating metadata such as definitions, synonyms, and hierarchies. Following data augmentation, a semantic information filtration phase removes duplicates and non-English terms, enhancing the quality of the knowledge graph. To uncover concealed linkages and associations between medical terms, we utilize the pre-trained contextual word embedding model Clinical BERT [6], trained on the MIMIC-III dataset. This facilitates the discovery of valuable insights within the data and contributes

to the creation of a comprehensive knowledge graph. We proposed cluster-based and node-based comparison methods to unveil hidden relationships by exploiting Clinical BERT within the knowledge graph.

Furthermore, our proposed approach enables users to effortlessly navigate complex features and generate autonomous knowledge graphs. Consequently, users can efficiently generate KGs based on input data, eliminating the need for prolonged waiting periods. Additionally, users have access to the generated files for further study and analysis. Ultimately, the discovery of hidden connections through our approach aids clinicians in gaining a deeper understanding of patient symptoms. Also it is benefiting insurance companies in identifying fraudulent claims and examining inaccurate forecasts of medical codes.

In summary, our research makes the following contributions:

- i. Proposed a significant approach for automating the construction of a Medical Knowledge Graph, known as Medical Knowledge Graph Automation (M-KGA).
- ii. Utilization of node-based and cluster-based comparisons for KG completion.
- iii. Conducting rigorous evaluation to demonstrate the efficiency of our technique and the resulting knowledge graph.

Further, the paper is arranged as: Related work on KG automation in the medical and other fields is discussed in Sect. 2. The proposed methodology to create KG is presented in Sect. 3. The result and assessment is describe in Sect. 4, conducted on several medical use cases. Finally, we have comprehensively presented the limitation, future direction and conclusion in Sect. 5.

2 Related Work

Over the past few years, numerous knowledge graph (KG) automation techniques have emerged across sectors like business, healthcare, finance, and education, each designed for specific use cases. KG creation in healthcare, however, faces unique challenges due to the data's diversity, volume, and complexity. While KG completion methods are valuable for uncovering connections using machine learning, they are rarely applied. Additionally, there is no platform that provides a fully automated KG for any specific area within the medical field.

Medical KG are built using a variety of approaches, including human, semi-automatic, and automated methods, as well as modern and conventional procedures. An important example of a semi-automated graph generation procedure that made use of scientific literature and pre-existing datasets is the COVID-19 graph [7]. Evidence mining, hypothesis ranking, and relation extraction were carried out using hierarchical spherical embeddings, ontology-enriched text embeddings, and cross-media semantic-structure representation. A mechanism for producing reports and responding to inquiries was also devised. 7,230 diseases, 9,123 chemicals, and 50,864 genes are included in the final KG. There are 1,725,518 chemical-gene relationships, 5,556,670 chemical-disease ties, and 77,844,574 gene-disease links.

Another method for creating healthcare knowledge graphs was available; it consisted of eight modules and produced a KG semi-automatically utilizing 16,217,270 de-identified clinical visit data from 3,767,198 patients [8]. Entity recognition, entity normalization, relation extraction, property calculation, graph cleaning, related-entity ranking, and graph embedding are among the processes in the process. To store extra context in healthcare graphs, the quadruplet form is used instead of the traditional triplet. A medical KG of 22,508 entities and 579,094 quadruplets with nine distinct entity kinds was the end result.

Knowledge graphs are widely used in fields such as medicine, diagnosis, and fraud detection. A Fraud, Waste, and Abuse (FWA) detection system, utilizing a Chinese medical knowledge graph, was developed to identify fraudulent claims in insurance [9]. Entities were extracted using deep learning techniques, with human validation ensuring accuracy. The graph, built from medical texts and drug labels from the Chinese Food and Drug Administration, comprised 1,616,549 nodes and 5,963,444 relations, achieving a 70% accuracy in detecting fraud claims.

Knowledge Graphs are often constructed for text data, which makes up 80% of all available data. The Semi-automated KG Construction and Application (SAKA) framework [10] is an interesting use case that utilizes both auditory and structured data for KG generation. The audio-based KG Information Extraction (AGIE) technique employs Voice Activity Detection (VAD), Speaker Diarization (SD), and the Medical Information Extractor (MIE) model to extract entities. Additionally, a system was developed to handle user inquiries and ensure the data remains relevant and updated. Testing on datasets like LibriSpeech, Vox-Celeb, and doctor-patient dialogues showed promising results. An automated medical knowledge graph for “Subarachnoid hemorrhage” used over a thousand case records, enhanced by Bioportal ontologies across several layers, including semantic, statistical, and predictive knowledge layers [11]. Although the source code is publicly available, adapting it for other medical fields is challenging, as switching ontologies alone isn’t enough, and word embeddings need retraining. The authors also overlooked knowledge graph (KG) completion. Another automation method for evidence-based medicine focused on cerebral aneurysm and COVID-19, using peer-reviewed ontologies and clustering models with deep learning techniques like RNN and BioBERT [12]. Accuracy reached 93% for COVID and 82% for aneurysm datasets.

The majority of approaches depict the KG creation process in a manual or semi-automated manner, as the literature demonstrates. The majority of them are tailored to specific use cases and the medical field, and they rarely incorporate the KG completion technique, which can harness the potential of BIG data to uncover facts. Although the generated KGs are indeed helpful in many ways, they cannot be applied generally. Furthermore, there isn’t a platform that can handle user requests to validate specific use cases and create the appropriate KG in a matter of minutes or seconds. Our method, which offers automation, is distinct and creative in that it uses expert-created ontologies to produce a full and comprehensive KG while satisfying the user’s request for a specific KG

generation. People can benefit from the approach in a variety of fields, and the KGs produced by our method can be used to enhance the research of others.

3 Methods

Our proposed approach is designed in multiple steps. Figure 1 illustrates the entire workflow of the Medical-Knowledge Graph Automation (M-KGA). This approach acquires data in two formats: structured and unstructured. It then applies various natural language processing (NLP) techniques to process the data. Initially, Bioportal is utilized to identify and enhance medical concepts with semantic information. The fetched data is filtered and used to create nodes in a knowledge graph (KG) along with their relationships. A pre-trained contextual word embedding model Clinical BERT is leveraged to discover hidden connections for KG completion. Finally, a Cypher query file is generated to facilitate the creation of the semantically enriched KG in Neo4j. The details of each individual stage are outlined below.

3.1 User Input

The M-KGA technique allows users to input medical data in two distinct formats: structured and unstructured. When we say “structured,” we mean that the user defines the medical terms with precision. The data does not contain any interconnected notions. Here is an example of a text that is organized in a structured manner:

Structured Input Example:

[‘fever’, ‘diarrhea’, ‘insomnia’, ‘severe acute respiratory syndrome’, ‘diabetes’]

Unstructured text, on the other hand, is free natural language text that is understandable to people but not to computers. It is the text written by a medical professional for instance diagnosis of a patient. An illustration is:

Unstructured Input Example:

[“If you have a condition called polyuria, it’s because your body makes more pee than normal. Adults usually make about 3 liters of urine per day. But with polyuria, you could make up to 15 liters per day. It’s a classic sign of diabetes.”]

The developed code can take data in both formats; if structured text is needed, it will ask for the data numerous times. If unstructured formatted text is needed, it will accept it all at once and find out the concepts on its own.

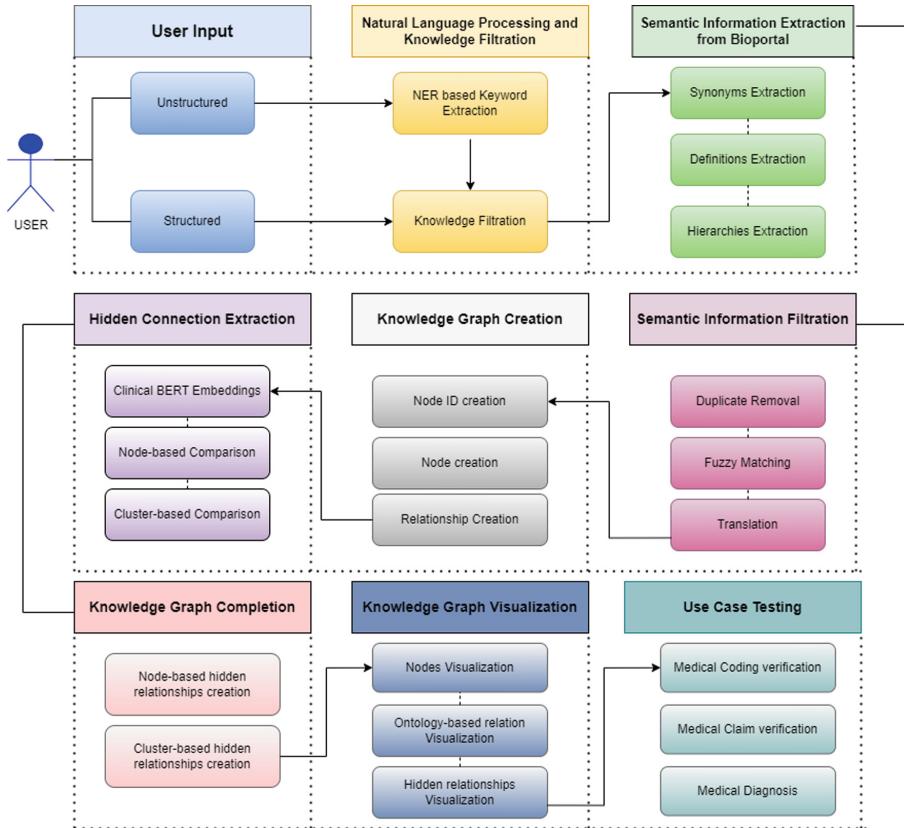


Fig. 1. Medical-Knowledge Graph Automation Approach (M-KGA).

3.2 NLP-Based Knowledge Filtration

We propose two NLP-based approaches for knowledge filtration: i) NER-based keyword extraction, and ii) knowledge filtration. The NER-based extraction identifies medical concepts from unstructured text and converts them into a structured format. Then, the knowledge filtration process selects the most relevant information from the extracted data. Each process is detailed below:

3.2.1 NER-Based Keyword Extraction and Knowledge Filtration

An unstructured document is input to extract clinical entities using an NER-based keyword function. This process populates nodes in the medical knowledge graph, integrating entities like diseases, treatments, and clinical concepts. We used the SciSpacy library [6], an extension of spaCy tailored for scientific and biomedical text, specifically leveraging the en_core_sci_sm model optimized for this domain. The NER-based function processes the unstructured text to extract

clinical concepts, producing a list of entities that can be seamlessly integrated into the medical knowledge graph.

Subsequently, acquiring a list of clinical or medical concepts, knowledge filtration function would be applied to choose only relevant and prominent concepts. The concepts that are previously extracted may contain some redundancies. Knowledge Filtration looks for duplicates in the data by Fuzzy-matching and filters it further. Additionally, the knowledge filtration assists the medical practice in decision making towards building a comprehensive KG by including and excluding certain concepts. The concept that is extracted during the NER-based keyword extraction process may be not relevant to current problem, disease diagnosis or treatment. By mapping these words to Bioportal Ontologies, more filtering is applied.

3.3 Semantic Information Extraction

The organized, enhanced, and sifted material from the preceding stage is used in the semantic information extraction step. Mapping these terms to expert-created Bioportal ontologies [5] allows you to retrieve the semantically enriched information using the Bioportal REST API. Our approach's strength is rooted in the notion that M-KGA is not exclusive to any particular medical condition or illness. Instead, it attempts to enrich data without being limited to particular ontologies.

Different sorts of semantically enriched information are retrieved from ontologies Such as Synonyms words, phrases, or morphemes that share the same meaning as the term being mapped, Definition which is a group of terms or phrases that provide a longer description of the term that is mapped. In this process, the two types of semantic information extraction took place.

3.4 Semantic Knowledge Filtration

Since our method is not limited to a specific medical condition or issue, M-KGA anticipates a high degree of data diversity, variation, and redundancy. In this step, data that has been semantically enhanced is filtered using a variety of techniques. The method attempts to translate data from many languages to English, eliminate duplicates from the retrieved data, and then use fuzzy-matching to further filter the findings.

3.4.1 Translation, Duplicate Removal and Fuzzy Matching

The data used in the enrichment stage is multilingual and comes from many ontologies. To translate this material, we exploited the 'translate' library in python. We need to take this action since, in the absence of translation, we will lose some important information. Text that we identified as non-English was translated into English; if the language cannot be identified or there are any exceptions, the text is eliminated. The subsequent stage does not include this deleted portion. Since the non-English text cannot be used in any further steps,

it has been removed. Pre-trained model employed to find hidden connections are unable to comprehend the data, which will result in problems. Aside from that, this step will address the limitations imposed by Neo4j on the creation of a Cypher Node ID.

This duplicate removal step takes the enriched data from the previous stage and tries to remove the duplicates. As we are fetching data from 1000+ Bioportal ontologies, we will likely have massive duplications. In this step we used semantic information, changed all synonyms and definitions to lower case and then used simple set operation on python to remove redundant entries.

An expansion of the duplicate removal process is fuzzy matching. Certain enriched data may contain semantically comparable text that cannot be removed with set procedures. In an effort to maintain the content's uniqueness, we employed this strategy. This also fixes problems in the ID creation stage and greatly aids in the removal of semantically duplicate items.

3.5 Knowledge Graph Creation

Creating a KG is a challenging endeavor in and of itself because it requires extreme caution while creating nodes and interactions. The KG was developed in Neo4j's Cypher query language. The format has unique limitations. The node ID in Cypher ought to begin with a character rather than a number, special character, non-English phrase, etc. Taking these factors into account, we produced graphs. This will be further explained in the steps following.

3.5.1 Node ID Creation

Nodes ID was developed with the understanding that hundreds of connections—both hidden and provided by the ontology—must be made between nodes. We translated the Node content or enriched data into ID by adhering to the ID requirements for different KG formats in order to reduce the amount of computing resources required for ID retrieval for comparison and connection formation. As previously stated, the Node ID in Cypher only accepts data in English format; special characters are not permitted, etc. Using this method, the enriched data on *polyuria* is transformed into IDs such as *excessive secretion of urine* from the definition of 'excessive secretion of urine'. Therefore, we don't have to go look for the ID connected to that Node every time we need to establish a connection. All we had to do was apply our function and turn the content into ID.

3.5.2 Nodes and Relationships Creation

This step builds the nodes for the structured and unstructured (converted to structured) data, as well as the semantically enriched data, according to the ID creation technique previously outlined. Different kinds of nodes have been created. Synonyms, medical concepts, definitions and so on are among the categories. Depending on its kind, every node in the graph is represented by a distinct color. The node displays the content. All KG nodes are constructed in this step.

After node creation, relationships among different nodes based on the expert crafted ontology-provided connections are created. With each iteration, a semantically enriched node is created with ID, the connection creation step, uses the ID and connects the node with the main medical concept. Same is the case for all semantically enriched data. Here, relationships are also of different types such as synonyms, definitions, and so on. Relationships are labeled and directed.

3.6 Hidden Connection Extraction

The earlier processes collect user data, filter it, obtain enriched data from ontologies produced by experts, and produce a knowledge graph. In addition to the links supplied by experts, our method looks for hidden connections absent from ontologies. The ontologies offer richer medical terminology, but it might be challenging to determine whether or not these concepts are related to one another. Exist any connections that could be omitted to improve the analysis of the medical data? We made knowledge graphs, but how do we complete them?

To address these problems, we attempted using KG embeddings for our method, which can predict links given KG triplets. Sadly, these methods are ineffective for tiny graphs. Our method can create both large and tiny KG in response to user requests; nevertheless, KG embeddings are unable to function on small networks because these models need thousands of triplets. Therefore, in order to create connections, we took advantage of word embeddings to determine a word's meaning and relationships with other words. It should be highlighted that our method looks for connections with other medical concepts and their enriched content rather than trying to establish links with its semantic enriched nodes, which are all already connected.

3.6.1 Clinical BERT Embeddings

We took advantage of Clinical BERT embeddings to extract vector representation of medical concepts and their contextual meta-data. We utilized the Clinical BERT embeddings, which are trained on a sizable medical corpus, in place of creating our own model. Medical Information Mart for Intensive Care III (MIMIC-III) is used to train the model. We took use of their pretrained nature and open-source nature to comprehend medical concepts and their interrelationships. We calculated the degree of similarity between various terms and built relations that led to knowledge graph completion based on the distance and user-defined threshold.

3.6.2 Cluster-Based Comparison

We offered two methods for locating the links that are buried in knowledge graphs. We treated every medical concept and its semantically enriched data as a cluster in a cluster-based method. Using all available semantic information, we composed a paragraph and then used the Clinical BERT model to look for embeddings. Clinical BERT implementation is not scalable and introduces mistakes on big clusters. In order to address this method, we segmented the

paragraph into chunks, handled exceptions, fetched embeddings for each chunk and then divided by the total number of pieces. A cluster is mapped with other clusters according to a user-defined threshold. The threshold and the degree to which users require specific or general connections are key factors here. In actuality, the threshold is the separation between the clusters. To determine whether clusters have strong relationships or not, users can choose the lower threshold. Cluster-based comparisons or connections lead to KG completion quickly and at low computing cost. This step introduces further relationships named 'embedding_match_cluster' in the KG. Figure 2 is an example of Cluster-based comparison method.

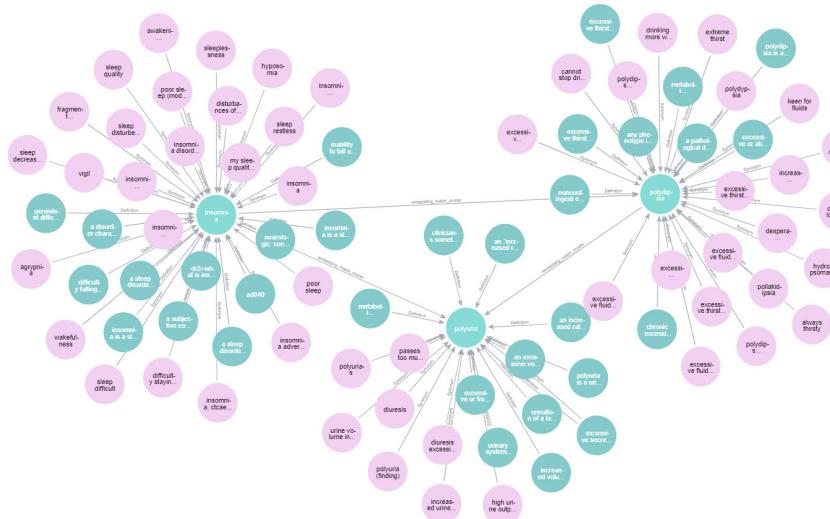


Fig. 2. Cluster-based comparison on Medical diagnosis use-case.

3.6.3 Node-Based Comparison

In contrast to the cluster-based approach, the node-based technique looks for connections with other nodes in the cluster. Using word embeddings, this compares a single node to every other cluster's node based on nodes. The lack of a large amount of text on the nodes means that scalability is not an issue. Furthermore, this method requires around n^2 time and is computationally costly, as opposed to the cluster-based method. Because the node-based technique allows us to determine the exact match of the link, it is much easier to understand. Here, connections are also established according to user-specified thresholds. Depending on the size of the graph, connecting nodes takes minutes. Figure 3 shows the result depiction of this method. This adds relationship named 'embedding_match_node' in the KG.

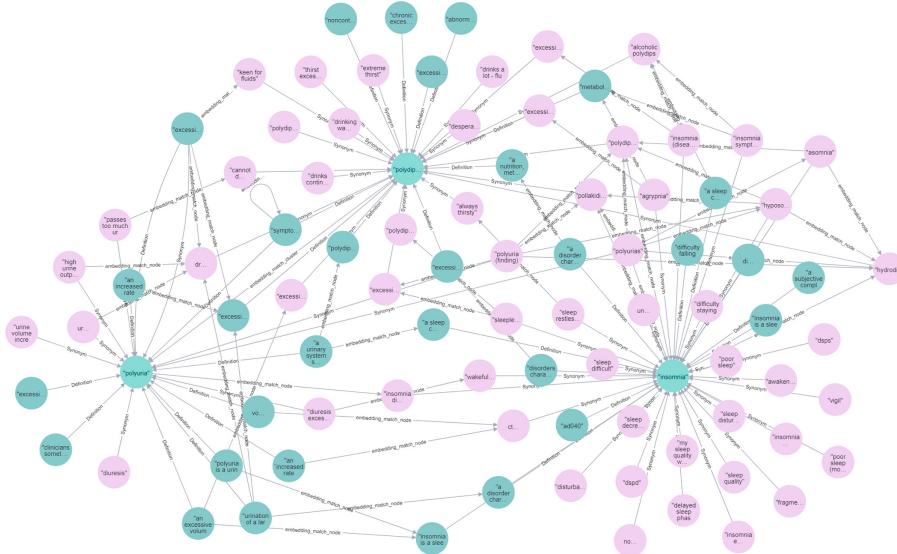


Fig. 3. Node-based comparison on Medical diagnosis use-case.

3.7 Use Case Testing

To determine the efficacy of our method, we ran three separate use cases through it. Three use cases were selected: medical claims, medical diagnostics, and medical coding. We used the dataset “CodiEsp” [13] to test each of these use cases. Experts annotate CodiEsp data with ICD-10 codes. To test our strategy, we used the discharge summaries along with their annotations.

3.7.1 Medical Coding, Claim Verification and Medical Diagnosis

We used several synopses, medical terms, and related medical codes. We extracted the medical terms contained in the descriptions of the medical codes after converting them into descriptions. We then used our Medical-Knowledge Graph Automation (M-KGA) technique to see the outcomes after passing all the medical concepts—such as discharge summaries or descriptions of diagnostic codes—through it. True positive and true negative cases were used in our tests. We took the summaries, applied our method to their annotations, annotated the summaries once again with fictitious examples, and retested the method. The approach demonstrated its usefulness through visualization and proved satisfactory in all cases. We also experimented with various thresholds.

We also tested the method for medical diagnosis using CodiEsp data. Each medical summary's knowledge graph was made using the concepts that were extracted from the summaries using the NER-based keyword extraction stage. We applied both node-based and cluster-based comparisons, and we generated a complete KG. We presume that any medical ideas included in a summary

must be related to one another; this relationship will confirm the usefulness of our method for completing graphs and offer a far more profound comprehension of the relationships taken from ontologies. Using our node-based and cluster-based comparisons technique, the strategy demonstrated significance in all experiments and most medical terms within the same summary generated links. To better understand the approach's operation, we also put it to the test with negative cases as well.

4 Results, Evaluation and Discussion

In this section, we present the results obtained from the implementation and testing of our proposed Medical Knowledge Graph Automation (M-KGA) approach. The evaluation aims to assess the effectiveness and efficiency of M-KGA in constructing comprehensive knowledge graphs from medical concepts provided by users. We conducted experiments using a diverse set of 300 medical concepts to evaluate the performance of our approach across various domains within healthcare.

Table 1. Inter Annotator Agreement (IAA) Among Domain Experts for Node-based Annotation

% of Agreement				Cohen's Kappa Value	
	Expert1	Expert2	Expert3	IAA between two Expert	
Expert1	1	94.25%	93.79%	Expert1, Expert2	0.86
Expert2	94.25%	1	94.55%	Expert1, Expert3	0.87
Expert3	93.79%	94.55%	1	Expert2, Expert3	0.86

Table 2. Inter Annotator Agreement (IAA) Among Domain Experts for Cluster-based Annotation

% of Agreement				Cohen's Kappa Value	
	Expert1	Expert2	Expert3	IAA between two Expert	
Expert1	1	92.56%	91.27%	Expert1, Expert2	0.85
Expert2	92.56%	1	91.85%	Expert1, Expert3	0.84
Expert3	91.27%	91.85%	1	Expert2, Expert3	0.84

In our evaluation, we partitioned the 300 medical concepts into two sets: 150 for assessing the cluster-based comparison method and another 150 for evaluating the node-based comparison approach. Each set underwent pairing, facilitated by the GPT-3.5 model, to create pairs of medical concepts. These pairs were

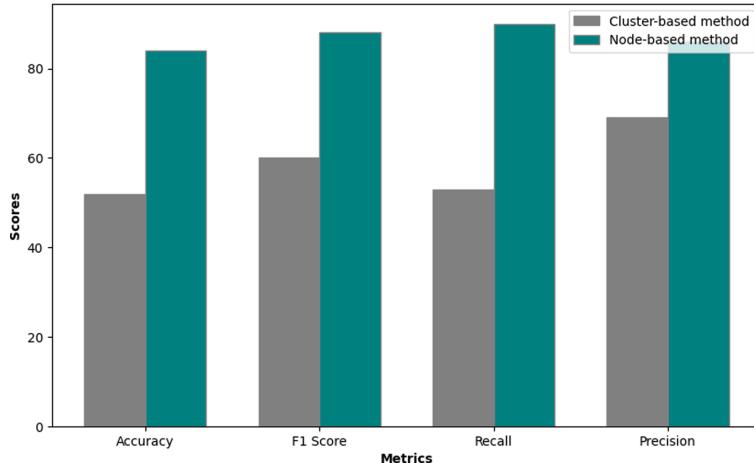


Fig. 4. Node-based verses Cluster-based comparison on Medical diagnosis use-case.

compiled into an Excel file for annotation by human medical experts. Owing to this, we calculated Inter-Annotator Agreement (IAA) assesses how consistently multiple annotators (domain experts) can make identical annotation decisions for a specific category, playing a vital role in validating and reproducing annotation results. Similarly, We utilized a well-known statistical measurements for the evaluation purposes Cohen's Kappa which is a statistical coefficient that quantifies the degree of agreement and reliability in statistical classification. It measures the concordance between two annotators (Domain Experts), each categorizing items into exclusive categories. Kappa value greater than 0.81% presented almost perfect agreements, where as the proposed framework achieve almost perfect annotation agreements of over 90% and Cohen's Kappa over 81% among the three domain experts both at Node-based and Cluster-based annotation as shown in Table 1 and Table 2. This indicates the robustness and reliability of the “M-KGA” framework in facilitating agreement among domain experts in knowledge graph recommendations.

The experts were tasked with annotating each pair based on measures of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN), providing valuable insights into the accuracy and performance of our approach. Following the annotation process, we applied both the cluster-based and node-based comparison methods to the pairs of medical concepts with threshold=4. Utilizing these methods, we constructed knowledge graphs for each pair and analyzed whether they successfully identified connections as annotated by the experts. Given that traditional ontologies often struggle to find connections among certain medical concepts, our objective was to determine if our approach could uncover hidden connections that might otherwise remain undiscovered. This analysis aimed to ascertain the efficacy of our proposed methods in aug-

menting existing knowledge and revealing previously unrecognized relationships within the medical domain.

In Fig. 4, the metrics constructed using True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) are presented. Our analysis focused on evaluating accuracy, F1 score, recall, and precision based on these metrics. The comparison depicted in the Fig. 4 highlights the performance disparity between the node-based and cluster-based methods. Notably, the node-based method emerges as the clear frontrunner, exhibiting significantly higher accuracy, F1 score, recall, and precision compared to the cluster-based approach. This observation underscores the effectiveness of the node-based method in accurately capturing connections within the knowledge graph, ultimately leading to superior performance across all evaluated metrics. The observed limitations in accuracy, F1 score, recall, and precision of the cluster-based method can be attributed to the utilization of Clinical BERT. While Clinical BERT is a powerful pre-trained model, its efficacy is constrained by practical considerations such as computational resources and sample size limitations. Due to the vast scale of clusters within the knowledge graph, it becomes necessary to divide them into smaller, manageable chunks for processing. However, this segmentation introduces a challenge: the loss of contextual coherence across multiple chunks. As a consequence, the embeddings derived from fragmented clusters may lack the holistic context necessary for accurate representation and inference, resulting in diminished performance metrics. This phenomenon underscores the importance of considering both the capabilities and limitations of pre-trained models when designing and implementing knowledge graph construction methodologies.

In addition to evaluation measures, transparency and time are critical factors in assessing the effectiveness of the M-KGA approach. As depicted in Figs. 2 and 3, transparency refers to the clarity and comprehensibility of the constructed knowledge graph. The node-based method excels in transparency by establishing direct connections between nodes, thereby presenting a clear and intuitive representation of relationships. In contrast, the cluster-based approach may exhibit less transparency, as it tends to add fewer relationships, resulting in a less explicit depiction of connections. Time, on the other hand, pertains to the efficiency of the knowledge graph construction process. The cluster-based approach demonstrates an advantage in terms of time efficiency, requiring less computational resources and processing time compared to the node-based method. However, this efficiency comes at a cost, as the cluster-based approach may sacrifice performance metrics such as accuracy, F1 score, recall, and precision, as previously discussed. While the cluster-based approach offers a quicker construction process, it may compromise transparency and performance. In contrast, the node-based method prioritizes transparency and performance, albeit at the expense of increased computational complexity and time consumption. Thus, the choice between these approaches should be carefully considered based on the specific requirements and priorities of the knowledge graph application.

5 Conclusion and Future Work

In conclusion, this study introduces the Medical Knowledge Graph Automation (M-KGA) approach, which aims to address the challenges associated with automating the construction of knowledge graphs (KGs) and enhancing their completeness. Leveraging user-provided medical concepts and BioPortal ontologies, M-KGA enriches the semantic content of KGs using pre-trained embeddings, thereby facilitating a more comprehensive representation of structured medical knowledge. Our approach incorporates two distinct methodologies, namely a cluster-based approach and a node-based approach, to uncover hidden connections within the knowledge graph. Through rigorous testing involving 300 medical concepts, our M-KGA framework demonstrates promising results, showcasing its potential to overcome the limitations of existing knowledge graph automation techniques. The performance metrics and graph visualizations presented in this study underscore the effectiveness of our approach in enhancing the transparency and accuracy of knowledge graphs, particularly in the medical domain. Looking ahead, future work will focus on addressing scalability issues associated with the cluster-based method, aiming to improve its performance. Additionally, we plan to explore retrieval augmented generation (RAG) with Large Language Models (LLMs), for knowledge graph development and performance comparison with our current approach. By continuing to innovate and refine our approach, we aim to further advance the field of knowledge graph automation and contribute to the development of more comprehensive and accurate representations of structured knowledge in healthcare domain.

Acknowledgement. This work is supported by the National Science Foundation under Grant No. 2431840.

References

1. Smolensky, P.: Connectionist ai, symbolic ai, and the brain. *Artif. Intell. Rev.* **1**(2), 95–109 (1987)
2. Zou, X.: A survey on application of knowledge graph. In: *Journal of Physics: Conference Series*, vol. 1487, p. 012016. IOP Publishing (2020)
3. Hao, X., Ji, Z., Li, X., Yin, L., Liu, L., Sun, M., Liu, Q., Yang, R.: Construction and application of a knowledge graph. *Remote Sens.* **13**(13), 2511 (2021)
4. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint [arXiv:1902.07669](https://arxiv.org/abs/1902.07669) (2019)
5. Noy, N.F., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Res.* **37**(suppl_2), 170–173 (2009)
6. Alsentzer, E., et al.: Publicly available clinical bert embeddings. arXiv preprint [arXiv:1904.03323](https://arxiv.org/abs/1904.03323) (2019)
7. Wang, Q., et al.: Covid-19 literature knowledge graph construction and drug repurposing report generation. arXiv preprint [arXiv:2007.00576](https://arxiv.org/abs/2007.00576) (2020)
8. Li, L., et al.: Real-world data medical knowledge graph: construction and applications. *Artif. Intell. Med.* **103**, 101817 (2020)

9. Sun, H., et al.: Medical knowledge graph to enhance fraud, waste, and abuse detection on claim data: model development and performance evaluation. *JMIR Med. Inf.* **8**(7), 17653 (2020)
10. Zhang, H., Wang, X., Pan, J., Wang, H.: Saka: an intelligent platform for semi-automated knowledge graph construction and application. In: *Service Oriented Computing and Applications*, pp. 1–12 (2023)
11. Malik, K.M., Krishnamurthy, M., AlObaidi, M., Hussain, M., Alam, F., Malik, G.: Automated domain-specific healthcare knowledge graph curation framework: subarachnoid hemorrhage as phenotype. *Expert Syst. Appl.* **145**, 113120 (2020)
12. Alam, F., Giglou, H.B., Malik, K.M.: Automated clinical knowledge graph generation framework for evidence based medicine. *Expert Syst. Appl.* **233**, 120964 (2023)
13. Miranda-Escalada, A., Gonzalez-Agirre, A., Armengol-Estepa, J., Krallinger, M.: Overview of automatic clinical coding: annotations, guidelines, and solutions for non-english clinical cases at codiesp track of clef ehealth 2020. In: *CLEF (Working Notes)*, vol. 2020 (2020)