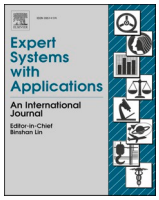
Expert Systems With Applications 233 (2023) 120964

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa



Automated clinical knowledge graph generation framework for evidence based medicine

Fakhare Alam [a], Hamed Babaei Giglou [b], Khalid Mahmood Malik [a,*]

[a] Department of Computer Science & Engineering, Oakland University, 115 Library Drive, Rochester, MI 48309, USA
[b] TIB Leibniz Information Centre for Science and Technology, Welfengarten 1 B, 30167 Hanover, Germany

ARTICLE INFO

Keywords:
Knowledge Graph
Healthcare Knowledge Graph
Deep Learning
Ontology
COVID-19
Cerebral Aneurysm
PICO
Evidence Based Medicine
Contextualization

ABSTRACT

To practice the evidence-based medicine, clinicians are interested to find the most suitable research for the clinical decision making. The use of knowledge graphs (KGs) in evidence-based clinical decision support systems is becoming increasingly popular. However, existing KG construction frameworks are not fully automated and contextualized, thus unable to adapt to new domains and incorporate constantly changing information into their knowledge base, resulting in loss of relevance over time. Furthermore, existing KGs construction frameworks don't generate KG that provide relevant information within an acceptable response time for evidence-based practitioners because the organization of constructed subgraphs is neither topic-specific nor evidence-based PICO (Participants/Problem P, Intervention-I, Comparison C, Outcome O) query-friendly. By employing concept extraction, semantic enrichment, optimized clustering, and state of art Recurrent Neural Networks (RNNs) with BioBERT based encoded representation to categorize PICO elements and predict relationships between concepts using huge corpus of publicly available literature on COVID-19 and cerebral aneurysm, this paper proposes a topic specific, PICO enabled, and fully automated framework to curate information and create KG of different clinical domains. The evaluation shows that the proposed framework achieves significant improvement over baseline models and has 93 %, and 82 % accuracy on aneurysm and COVID data set respectively for PICO classification. Also, the relationship extraction module has an accuracy of 96 % with precision and recall being 92 %, and 90 % respectively.

1. Introduction

The rapid growth of Knowledge Graph (KG) in recent years has been indicative of a resurgence in knowledge engineering. In many real-world applications, KGs have been found to be the primary source of information extraction, including text interpretation, recommendation systems, and answering natural language questions. A number of knowledge bases have been created, covering common sense knowledge such as Cyc (Lenat,1995), conceptNet (Liu & Singh, 2004), lexical knowledge such as WordNet (Oram, 2001),and BabelNet (Navigli & Ponzetto, 2012), encyclopedia knowledge such as Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008) DBpedia (Auer, et al., 2007), YAGO (Suchanek, Kasneci, & Weikum, 2007), WikiData (Vrandečić & Krötzsch, 2014), CN-Dbpedia (Xu, et al., 2017), probabilistic taxonomy knowledge like Probase (Wu, Li, Wang, & Zhu, 2012) and geographical knowledge GeoNames (geonames, 2022) Incorporating the above-mentioned open KGs has enabled a new form of knowledge storage that is capable of capturing semantically related large datasets. KGs that are generic in nature, as well as open world KGs, cannot address all the real-world issues. There is a recent interest in developing KGs specific to a wide variety of specific domains such as education, science, automotive, and healthcare (Du & Li, 2020; Wang, et al., 2020; Zhu, et al., 2020; Mohamed, Nováček, & Nounu, 2020; Li, et al., 2020) and specific problems such as impact analysis (Johnson, et al., 2022), root cause analysis (Wang, et al., 2021; Qiu, Du, Yin, Zhang, & Qian, 2020), and feature engineering. Domain and problem specific KGs are important because they include entities that are relevant to the domain and are semantically related to it.

The application of evidence-based medicine is crucial for delivering high-quality care in healthcare such as nursing, family medicine, etc., since it allows clinicians, physicians, and patients to evaluate available research and evidence, and apply data-backed solutions to make quantitative and qualitative decisions based on scientific results. Additionally, this knowledge can be used to provide evidence-based treatment

* Corresponding author.
E-mail addresses: fakharealam@oakland.edu (F. Alam), hamed.babaei@tib.eu (H.B. Giglou), mahmood@oakland.edu (K.M. Malik).

https://doi.org/10.1016/j.eswa.2023.120964
Received 7 November 2022; Received in revised form 29 June 2023; Accepted 6 July 2023
Available online 16 July 2023

options, improve treatment effectiveness, and ultimately help find a way to mitigate the disease by providing healthcare providers with a better understanding of the disease and its symptoms. However, updating the evidence-based decision support system using continuous evolving unstructured voluminous literature is challenging. The use of KG in evidence-based clinical decision support systems could provide an ability to store and process scientific evidence and clinical results, enable in-depth understanding of symptoms, related causes, potential diagnoses, enable historical data analysis about specific diseases, and assist experts when confronted with complex patient cases. However, it can be a herculean task to consume this vast amount of published knowledge and present it in machine understandable format and KGs have been proven to be an effective utility in this arena and are rooted in many healthcare applications such as patient diagnosis (Xi, Ye, Huang, & Li, 2021), possible disease treatments (Dai, Guo, Guo, & Eickhoff, 2021; Malik, et al., 2020), associate relationship between biomolecules and diseases(Yu, et al., 2021), for better data representation and knowledge inference by mapping relationships between the enormous variety and structure of healthcare data. Also, structuring this large graph-based data should make the query processing easy. This paper argues that curation of knowledge from unstructured data and structuring it using knowledge graph-based topic modeling and PICO elements could solve the challenges of evidence-based decision support systems.

The construction of knowledge is currently carried out manually or in a semi-automated manner, which requires significant effort and expertise. Identification of knowledge sources, disambiguation, and recognition of concepts in context, semantic enrichment and relationship determination of concepts, and inferential reasoning are some of the challenges associated with the creation of a knowledge base. The automatic extraction of coherent knowledge and construction of KGs from the different forms of data is challenging and a long-standing goal in artificial intelligence research. Data and entities in original data sources continuously change and evolve over time as new knowledge is added through articles, scientific manuals, and procedures such as recipes, how-to guides, etc. The various factors that define entities and relationships change over time, the ignorance of temporal information and the dynamic nature of the entities can lead to incorrect information extraction and poor decisions being made as a result of these factors. Therefore, automated construction of KGs must be more resilient and robust in order to accommodate continuous influxes of new knowledge, include temporal and contextual information, and make implicit information explicit.

The use of KGs on published literature to distill usable information that neuro-symbolic models and expert-based systems could use is one of the most promising approaches to the data consumption problem; and also, it provides more explanations for AI techniques such as machine learning and deep learning. Published clinical knowledge is an enormous amount of data and the knowledge is constantly evolving, where physicians are constantly attempting to find new evidence of diagnosis, treatment, and eventually cure. KG presents a light way of organization of biomedical facts for different purposes such as question answering, search engines, recommender systems, or medical assistants and it powers AI-driven discovery related to different drugs and diseases. At present, healthcare KGs are constructed in a static and manual or semi-automated fashion, and their relevance may erode over time as the underlying contextual and temporal information keeps changing with the discovery of new facts, leading to new understanding of complex phenomena. Additionally, existing healthcare KGs have issues pertaining to retrieval of relevant information within acceptable response time (Chen, Cui, Liu, Wu, & Wang, 2020), as the subgraphs are not optimized with respect to particular topics such as PICO elements (Participants/Problem P, Intervention-I, Comparison C, Outcome O). Enriching the subgraph with specific problems, intervention, and outcomes is key to delivering relevant search results and it reduces the search time too. However, the PICO classification is challenging due to the unavailability of clean, curated and preprocessed text corpus specific to the domain. Additionally, as knowledge evolves, the accuracy of entities and relationships among them changes, making it difficult to keep KGs contextual and current. As a result of continuous evolution, it is also more difficult and challenging to maintain the continuity of existing relationships, construction of new relationships from new knowledge, and embed it into existing KGs because there must be a mechanism to store temporal information. With the rise of embedding specific to biomedical domains such as BioBERT (Lee, et al., 2020), there is a need to develop deep learning-based techniques to classify KGs into PICO specific subgraphs and dynamically predict relationships between entities using contextual encoding layers and deep learning-based models.

This paper attempts to solve the above-mentioned challenges by proposing a PICO enabled automated knowledge graph curation framework that has the capability of handling not only the constantly changing information, entities, and relationships, but also having the capability to ingest and incorporate new data sources as they become available. The implementation methodology in the framework focuses on incorporation of basic KG qualities such as accuracy, consistency, completeness, timeliness, availability, and scalability (Wilkinson, et al., 2016), and aims to provide an automated framework for curation of KGs that is highly efficient and yet scalable. The use of KGs to assist clinicians and scientists has already been widely demonstrated by the authors (Shen, et al., 2020; Reese, et al., 2021), but there is still more to be done.

In the current study, we propose an automated framework for generating dynamic knowledge graphs for the healthcare domain by taking COVID-19 and cerebral/intracranial aneurysms as an example. In this framework, we addressed the identification and disambiguation of concepts using well known COVID ontologies as well as aneurysm ontologies, hierarchical relationships are determined using ontologies and deep learning-based machine learning models with an emphasis on contextual embedding using domain specific representation layer. We semantically enriched the entities by using ontology definitions, labels, and synonyms to provide context. Integrating multiple ontologies contributes to consistency, completeness, and diversity. An easy to update mechanism makes this framework scalable to new concepts, relationships, and ingest new information.

The main contribution of this paper can be summarized as follows:

- First, we created a framework for building initial schema using identification of semantically enriched clinical concepts by utilizing multiple generic, COVID-19, and aneurysm ontologies. As a result of merging information from multiple ontologies, a holistic approach is provided and self-learning is induced, which is ideal for a wider range of data analysis, as well as for integrating any newly published articles or publications. To ensure the accuracy of the data, only peer-reviewed and published ontologies are considered. The proposed method ensures that multiple clusters will emerge automatically and that KG updates within each cluster based on more data and information.
- Second, an optimized clustering-based machine learning model to identify important topics in the research papers. This model takes vectorized input of the research papers and uses k-means clustering to identify prominent topics present. The numbers of topics are optimized using distortion and inertia. The searchability of KGs are further increased by categorizing each sub-KGs into Aim(A), Problem (P), Intervention (I), Comparison (C), and Outcome (O).
- Third, a contextual deep learning model that uses Recurrent Neural Networks (RNNs) based Long Short-Term Memory (LSTM), to classify research paragraphs by Aim, Patient/Population, Intervention, Comparison, and Outcome categories in addition to providing PICO classified and BioBERT encoded dataset extracted from publicly available COVID-19 research papers.
- Fourth, we present a Bio-BERT encoded, an optimized Bi-LSTM model that utilizes CNN layers to extract relationships between entities.

Lastly, we present a methodology in which concepts, taxonomic relationships, and non-taxonomic relationships can be combined in order to create an accurate, flexible, and scalable framework for building a KG, and has an ability to ingest the continuous changing knowledge.

The remainder of the paper is structured as follows. Section 2 reviews previous literature on healthcare knowledge graph, KG construction methods, and relationship extraction methods. Section 3, material & methods provides the details about the dataset, the framework of automated KG, followed by results & evaluation in section 4. Section 5, Section 6, and Section 7 highlights contributions, limitations, and future research directions of this paper respectively.

2. Related works

2.1. Knowledge graph in healthcare

In numerous studies, KGs have been constructed and used for clinical research and healthcare. KG has a great deal of potential for diagnosing diseases (Chai, 2020), especially in terms of identifying disease-symptom relationships (Pham, 2022). Healthcare specific knowledge bases have gained traction as a result of the COVID-19 pandemic, and efforts are being made to establish, enhance, and create disease specific knowledge bases since they provide a means of mining, storing, and analyzing vast amounts of multimodal, heterogeneous data. Table 1 shows the review of recent healthcare knowledge generation by analyzing them on methodology such as reasoning, ML based, data sources and its limitations in terms of generalizability, PICO enabled, and topic specificity.

2.2. Knowledge graph construction methodology and data sources

In traditional KG construction, Named Entity Recognition (NER) (Mohit, 2014), entities disambiguation, and Relation Extraction (RE) methods are involved, and it also depends on the underlying schema. Schema options include ontology schemas (Kuhn, Mischkewitz, Ring, & Windheuser, 2016; Suchanek, Kasneci, & Weikum, 2007), schema free approaches, or hybrid approaches that combine schema-free and schema-based approaches. In recent times, KGs have been constructed using big data mining and Natural Language Processing (NLP) techniques, such as association mining, deep learning methods such as Convolution Neural Network (CNN), Recurrent Neural Network (RNN), machine learning models such as Random Forest, Support Vector Machines, Logistic Regression, XGBoost, Adaptive Boosting, and extract knowledge from different sources, including silos of data and entity repositories (Smirnova & Cudré-Mauroux, 2018; Al-Moslmi, Ocaña, Opdahl, & Veres, 2020; Ji, Pan, Cambria, Marttinen, & Philip, 2021). However, these methods need domain expert to validate the mined rules. There are also some built-in models available such as IBM Watsons (Chen, Argentinis, & Weber, 2016), AllenNLP (Gardner, et al., 2018), and CoreNLP (Manning, et al., 2014).

In healthcare, construction of domain specific KGs using the above-mentioned approach is also on the rise as it offers an effective way to extract meaningful information from different modalities and variety of data. (Gatta, et al., 2017) proposed an approach to mine medical data and store it in directed graphs for easy visualization and interpretation. (Mohammadhassanzadeh, Abidi, Van Woensel, & Abidi, 2018) integrated ontology-based knowledge with reasoning for completeness by integrating various knowledge bases such as DrugBank, Disease Ontology, and semantic databases such as MEDLINE. (Rastogi & Zaki, 2020) constructed the personalized KGs by using contextual information and combining it with knowledge bases. There has been effort made on constructing disease specific KGs to extract health-related symptoms and risk factors. (Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017) constructed KG by extracting disease and symptoms from Electronic Health Records (EHRs) and combining it with Google Health Knowledge Graph (GHKG). Likewise (Huang, Yang, Harmelen, & Hu, 2017)

Table 1
Healthcare Knowledge Graph Review Summary- Implementation Methodology and Data Sources.

Reference	Description	Data Sources	Limitations
(Dai, Guo, Guo, & Eickhoff,2021)	Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings	DeepDDI and Decagon	. Semi-Automated . Not PICO enabled . Less generalizable
(Yu, et al., 2021)	multi-typed drug interaction prediction via efficient knowledge graph summarization	DrugBank Dataset	. Semi-Automated . Not PICO enabled . Less generalizable
(Xi, Ye, Huang, & Li, 2021)	Breast Cancer diagnosis from clinical ultrasound reports	Collected by clinicians Cancer Center of SUn Yat-sen University	. Automated manual feature creation . Not PICO enabled . Less generalizable
(Mohammadhassanzadeh, Abidi, Van Woensel, & Abidi, 2018)	Investigating plausible reasoning over knowledge graphs for semantic based health data analytics	Medline	. Automated . Not PICO enabled . Less generalizable
(Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017)	Identification and classification diseases and symptoms across the collected medical records using classifiers such as Logistic Regression (LR), Naive Bayes (NB), Bayesian Inferences	Custom Data, Google Healthcare Knowledge Graph (GHKG)	. Automated human in loop . Not PICO enabled . Less generalizable
(Liu, Yin, Wang, Liu, & Ni, 2021)	Multiple criteria decision-making approaches for healthcare management applications	PubMed, DrugBank, DrugBook, and UMLS	. Semi-Automated . Not PICO enabled . Less generalizable
(Yu, et al., 2017)	Knowledge graph for traditional Chinese medicine health preservation	TCM data	. Semi-Automated . Not PICO enabled . Less generalizable
(Postiglione, 2021)	Italian healthcare knowledge graph	EHRs	. Manual . Not PICO enabled . Less generalizable
(Gyrard, Gaur, Shekarpour, Thirunarayan, & Sheth, 2018)	Personalized health knowledge graph	EHRs, weather data,	. Manual . Not PICO enabled . Generalizable
(Li, et al.,2021)	Real- world data medical knowledge graph	EMRs	. Manual . Not PICO enabled . Generalizable

constructed KGs for depressions by utilizing PubMed, DrugBank and United Medical Language System (UMLS) library. Since the entities and relationship between them are changing, (Ma, Tresp, & Daxberger, 2019) attempted to construct a temporal KG for cognitive episodic memory by generalizing four static KGs vectors to 4-dimensional temporal KGs. (Zhang, et al., 2020) present a platform named Health Knowledge Graph Builder that could be used to build disease specific KGs considering multiple sources of data including those of clinician feedback. The framework utilizes data from medical records to inspect new concepts/relationships and allows clinicians to annotate unstructured data based on existing clinical KGs. While this approach provides automated and manual knowledge representation and allows to add new diseases to existing KGs, the dataset used was limited in quantity. Also, the dependency on several sources and bias caused due to manual intervention might reduce the extensive usability of this framework. The work in (Chen, Agrawal, Horng, & Sontag, 2020) provides a case study on how to evaluate the accuracy of health KG models for disease-symptom relationships. Unstructured datasets from de-identified patient EHRs (273,174 EHRs) are extracted and each chronological patient file is identified as an object entity, GHKG is used as the comparison entity to evaluate disease-symptom edge on the KG. While the approach aims at availing accurate information on healthcare KGs, their study is limited in comparison with GHKG and unclear on the inferencing of outliers. Likewise (Li, et al., 2021) presents a series of methods to create a medical KG using a quadruplet structure with data extracted from large EHRs, laboratory reports, and radiology reports. NER is used to identify 9 different objects pertaining to diseases, symptoms, medicine, gender, age, exam, lab exam, lab item and surgery. Relationship is extracted by building triplets of two entities with a relation, and every relation is associated with 4 properties providing the quadruplet structure. The work provides a novel approach of quadruplet medical KG however, the KG build process is variable intensive and manual. (Wise, et al., 2020) present a COVID-19 Knowledge Graph (CKG) for retrieval of semantically similar articles. The proposed CKG uses a hybrid approach by combining semantic information with the topological document information. (Chen, Ebeid, Bu, & Ding, 2020) utilizes the rich CORD-19 dataset and maps the insights obtained from PubMed dataset using machine learning to present a KG that helps identify COVID-19 experts and concepts from published literature. (Wang et al., 2020) propose COVID-KG, a knowledge discovery framework that uses semantic representation and ontologies to provide multimedia analysis of text and image data from literatures to aid in clinical Q&As and generate reports associated with COVID-19 drug administration based on a case study. The authors in (Michel, et al., 2020) have successfully constructed the COVID knowledge graph by leveraging the CORD-19 data abstracts, publicly available data sources such as DBpedia, Wikidata, and utilizing Argumentative Clinical Trial Analysis (ACTA) tool. ACTA tool has a limitation in that it is particularly effective for clinical trial-related documents. However, the CORD data encompasses a broader range of research studies and information, including clinical trials.

2.3. Knowledge graph relationship extraction methodology

Most of the work in Relationship Extraction (RE) employed traditional Biological Expression Language (BEL) (Domingo-Fernández, et al., 2021), machine learning and deep learning approaches in supervised and distant supervision (Pawar, Palshikar, & Bhattacharyya, 2017) fashion. However, recently the pre-trained language models (PrLMs) (Aydar, Bozal, & Ozbay, 2020) achieved state-of-the-art performances in many tasks of NLP including domain-specific REs. (Pawar, Palshikar, & Bhattacharyya, 2017) concluded that semi-supervised approaches are well suited for open domain RE systems due to their ease of scaling and extensibility in finding new relations.

The earlier approaches heavily relied on hand-crafted feature engineering (Pawar, Palshikar, & Bhattacharyya, 2017). These methods could be categorized into lexical (Daelemans & Bosch, 2005; Mintz, Bills, Snow, & Jurafsky, 2009; Santus, Biemann, & Chersoni, 2018; Brychcín, Hercig, Steinberger, & Konkol, 2018) and word embedding (Lai, Leung, & Leung, 2018; Zeng, Lin, Liu, & Sun, 2016; Speer & Lowry-Duda, 2018) features. (Buscaldi, Schumann, Qasemizadeh, Zargayouna, & Charnois, 2017) reported the popularity of SVM in machine learning models at the SemEval-2018 relationship extraction task. The machine learning based RE methods mostly suffers from feature quality and is unable to generalize well on many domains because of feature distinctions in different domains.

The latest studies in deep learning focus on extraction of deep relational features using CNN and RNN-based approaches, although attention mechanisms have been combined with CNN or RNN methods in a few cases. CNN-based methods as implemented by (Liu, Sun, Chao, & Che,2013; Zeng, Liu, Chen, & Zhao, 2015; Lin, Shen, Liu, Luan, & Sun, 2016) extract global features of relations with a sentence as high-level features for classification. RNN based methods as implemented by (Ji, Liu, He, & Zhao, 2017; Jat, Khandelwal, & Talukdar, 2018; Du, Han, Way, & Wan, 2018; Zhang & Wang, 2015) use forward and backward propagations to obtain sequential features by considering past and future sequences. (Wang, Cao, De Melo, & Liu, 2016) proposed CNN architecture, which relies on two levels of attention to better discern patterns in heterogeneous contexts. In addition, (Lee, Seo, & Choi, 2019) proposed a new end-to-end RNN model that incorporates entities and their latent types as features for RE. Similarly, (Xiao & Liu, 2016) introduced a hierarchical RNN using BiLSTM, where it is capable of extracting information from raw sentences for RE. Attention based mechanisms allow more accurate feature extraction from concepts and can predict relations by analyzing the whole sentence, however it suffers from an out-of-vocabulary and local contextual representation issues since it is trained in a specific training set.

Recently researchers are also using transfer learning model such as PrLMS for relationship extraction. The basic architecture of PrLMs includes a stack of encoders, fully connected to a stack of decoders. Each encoder consists of a self-attentive component and anticipatory network blocks. Each decoder consists of a self-attention component, an encoder-decoder attention component, and an anticipation component block (Harnoune, et al., 2021). Domain-specific LMs, such as Bidirectional Encoder Representations from Transformers (BERT), Transformer-XL, and OpenAI's GPT-2 have been used for RE tasks in multiple domains. For example, (Wu & He, 2019) proposed a model that exploits both the pre-trained BERT LM and information from the target entities for the relationship classification task. They fine-tuned BERT for the RE task and achieved higher results than CNN or RNN-based approaches. The single BERT-based model in joint entity-relation extraction has been investigated by (Eberts & Ulges, 2021), where they presented a BERT-based RE framework with a softmax classifier and showed that joint models with PrLMs are performing well. (Harnoune, et al., 2021) concluded that the BERT variant such as BioBERT allows faster learning process by using RNN and CNN based architecture in medical RE tasks. However, there are still areas for improvement. In another study, (Kim, Yun, & Kim, 2021) considered the BERT model for unsupervised relation extraction and introduced Co-BERT, an open information extraction service that uses the power of BERT models with minor adjustments in self attention. Authors conclude that the attention mechanism in fine-tuning LMs for domain-specific tasks plays a critical role, and this is the reason why BERT model representations perform well for domain specific relationship prediction.

Based on above literature review, it is obvious that existing approaches in curation and generation of KGs has following limitations:

a. KGs do not have the ability to handle the continual influx of new data sources and knowledge and lack the ability to continuously update contextual and temporal information.
b. Domain specific KGs cannot be generalized and have weaknesses in terms of concept extraction, relationship extraction, quality, and completeness.

c. The current KGs are not PICO-specific; having a PICO-enabled subgraph will make searching and indexing in KGs easier.

3. Material and methods

3.1. Datasets

In the current study, the curation framework for knowledge graphs has been implemented and tested on two data domains specific to the field of health care, COVID-19 and subarachnoid hemorrhage. In addition, it uses relationships extraction datasets to extract triplets and identify relationships between concepts. The framework also uses generic and specific ontologies from BioPortal to extract hierarchical relationships per concepts.

3.1.1. COVID 19 dataset

In the current framework, the COVID-19 open research dataset (CORD-19) (Wang & Kohlmeier, 2020) dataset is being used to gather research papers that are related to COVID-19 and contain information about the COVID -19. The CORD-19 database contains more than a million scholarly articles on COVID-19, SARS-CoV-2, and other coronaviruses.

We used CORD-19 as our initial dataset to collect textual information about covid using abstract text of the paper and full paper text if it was available, and later categorize the sentences into PICO with Aim category using the parsing framework presented by (Afzal, Alam, Malik, & Malik, 2020). This dataset along with BioBERT encoded representations is publicly available to consume for further research in GitHub repository ("GitHub - PICO Dataset," n.d.).

3.1.2. Cerebral aneurysm dataset

A cerebral aneurysm, also known as a brain aneurysm or Intracranial aneurysm, occurs when an area of an artery in the brain becomes weak or thin, causing it to expand and fill with blood, resulting in a bulge. These aneurysms have the potential to rupture, leading to a life-threatening hemorrhage. It is crucial to possess a data-driven understanding in order to gain insights into the likelihood of aneurysm rupture, as even experienced physicians are unable to accurately predict the probability of rupture (Malik, Anjum, Soltanian-Zadeh, & Malik, 2018).

In order to test the framework's robustness and scalability on another dataset, we used the publicly available PICO classified dataset (Afzal, Alam, Malik, & Malik, 2020) that was curated from biomedical literature and developed based on a study of the quality of the research papers, using ensemble machine learning models and PICO classification using deep neural network. This dataset contains more than 170,000 abstracts related to subarachnoid hemorrhage and classified into A-Aim, P-Patient/Problem, C-Comparison, O-Outcome. Table 2 shows the frequency per each PICO category.

3.1.3. Relationship extraction dataset

BioRel (Xing, Luo, & Song, 2020) is an imbalanced dataset for biomedical Relation Extraction (RE), offering reasonable performance as a starting point. The first step in this dataset was to identify Medline sentences mentioning entities, followed by MetaMap (Aronson, 2001) references linked to UMLS. In the end, each pair of entities in the sentence has its own relation label. A Multi-Instance-Learning (MIL) dataset, BioRel is divided into train, development, and test sets.

Each set consists of n bags $\{\langle h_1, t_1, r_1\rangle, \langle h_2, t_2, r_2\rangle, \ldots, \langle h_n, t_n, r_n\rangle\}$, where h_i, t_i, and r_i are head entity, tail entity, and relation type. Each sentence consists of a sequence of k words $\{w_1, w_2, \ldots w_k\}$. For the task of RE in this research, we have used only head h and tail t entities and relation types. For example:

Three cases of **acute myeloid leukemia** developing after treatment of renal disease with **cyclophosphamide** have been studied.

Where 'acute myeloid leukemia' and 'cyclophosphamide' are head and tail entities. For this instance, in a bag, the relation type identified is **'treated by'**. The dataset contains 533,560 sentences, 26 million words, 69,513 entities, and 125 types of relationships. Table 3 presents the frequencies of the train, development, and test sets for the dataset with respect to sentences, head entities, and tail entities.

3.2. Automated curated knowledge net framework

A Knowledge Graph is a large semantic network, defined as a directed labeled graph *G:(V, E)* with a set of nodes V: $\{v_1, v_2 \ldots v_n\}$ and edges E: $\{e_1, e_2 \ldots e_n\}$. A node is a real-world entity, such as an object, an event, a situation, a place or a person, while a boundary defines a relationship among them. Specifically in the healthcare, nodes can be diseases, drugs, or symptoms and edges could be defined as *'cures','treats' 'symptoms of'*. KGs organize raw information in a structured form, integrate knowledge across different sources by defining descriptions of entities and relationships that are both understandable by humans and machines. It stores rational facts and information in a centralized fashion, bridges data silos, and generates canonical business knowledge models that can incorporate all data formats such as structured, semi-structured, and unstructured.

To address these unmet needs, an automated deep-learning curated knowledge network is proposed. This framework utilizes Natural Language processing (NLP) techniques to cleanse text data, followed by a clustering approach to identify different topics in the article corpus. Next, we used a deep learning-based classification model to categorize text into PICO elements. The classified text was then used in the KG modeling employing concept extraction using Ontology Based Information Extraction (OBIE) techniques, relationship extraction using Biomedical BERT(BioBERT) and CNN and Bidirectional LSTM (Bi-LSTM).

The proposed KG construction framework includes five major components: Data Preprocessing, Clustering, PICO classification, Relationship Extraction, and Knowledge Graph Generator. The main tasks of the data processing module are cleaning, stop words removal, tokenization, and generating vector representations. The clustering module includes generation, identification and optimization of clusters in the text corpus by using semantic enrichment, vector embedding and dimensionality reduction techniques. The PICO classification modules classified each paragraph into PICO elements and an additional category Aim by using RNNs based LSTM classifiers. The relationship extraction module utilizes generic and specific ontologies to extract hierarchical relationships and uses deep learning-based models to extract non-taxonomic relationships between concepts. Lastly, the knowledge generator module combines triplets, hierarchical relationships, and creates initial KGs that are optimized to remove ambiguous relationships and add missing links. Fig. 1. shows the different components of automated KG construction

Table 2
PICO elements distribution frequencies in aneurysm dataset.

PICO Category	PICO Elements Frequency
A	31,676
P	49,308
I	11,711
R	30,515
O	50,072

Table 3
Distribution of sentences, head entities, and tail entities in train, development, and test splits for BioRel Dataset.

Unique Count	Train	Development	Test
# Sentences	534,277	114,506	114,565
# Head Entities	36,961	16,550	16,583
# Tail Entities	37,607	16,902	17,011

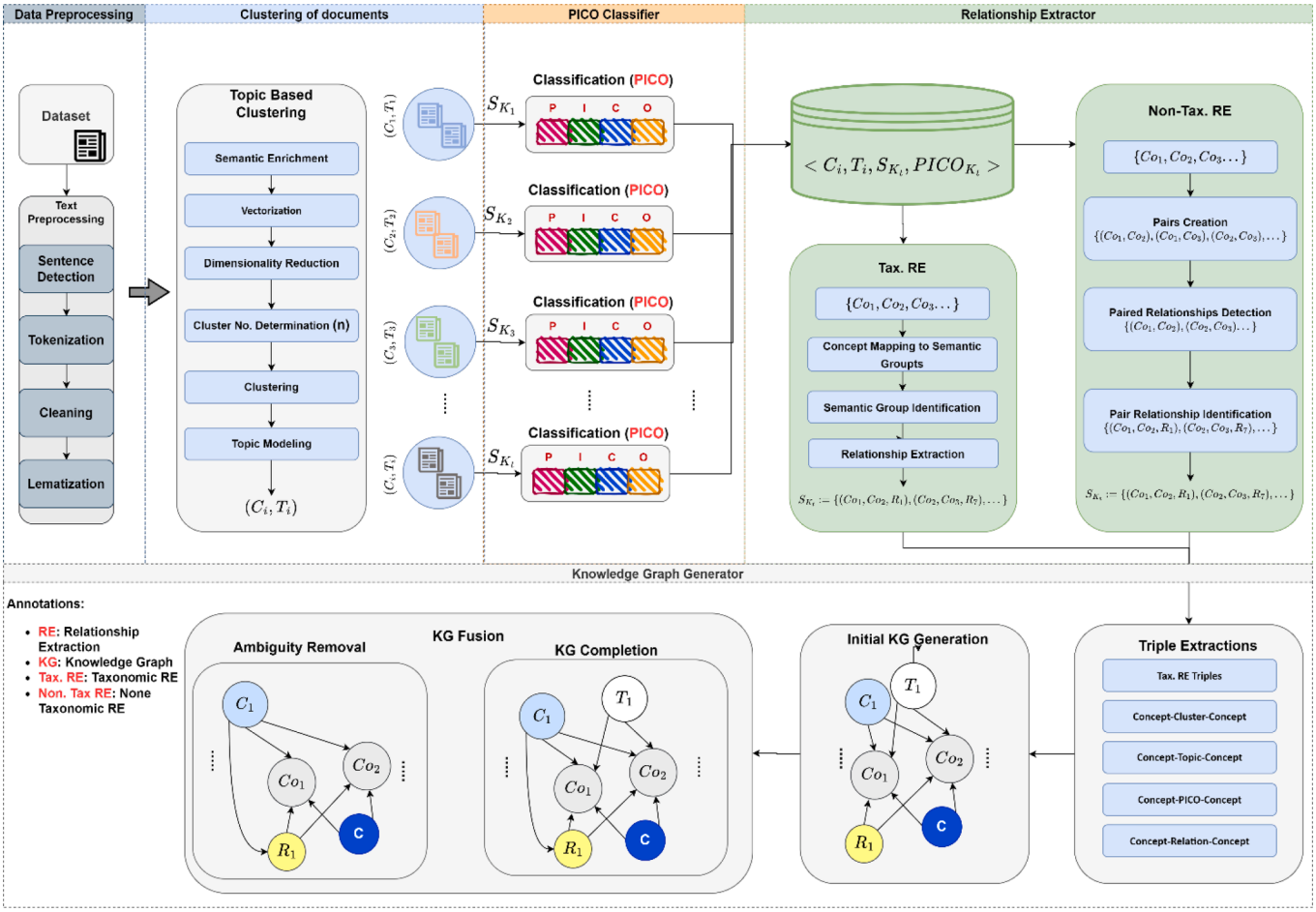


Fig. 1. Functional View of proposed Automated Knowledge Graph Curation Framework with five major components; Data Preprocessing, Clustering, PICO classification, Relationship Extraction, and Knowledge Graph Generator.

framework.

3.3. Data preprocessing

Data preprocessing is the process of transforming raw data into a format that is easily understood by machine learning algorithms. The COVID-19 dataset, as well as the aneurysm dataset contain texts that come from a variety of data sources. Thus, this step of preprocessing eliminates the inconsistencies and duplicates in the data, performs the text cleaning for NLP, and vectorizes the text into an embeddable format as well.

3.3.1. Text cleaning

In the text cleaning process at first, we removed special characters, HTML tags, references, tables, and images in the full text of the paper, then we followed up by detecting sentences and removing punctuation and stop words from the papers. To remove stop words, we enhanced the Python NLTK library (Bird, et al.,2009) stop words with medical-specific stop words. For the purposes of identifying special characters, symbols, and URLs, we used regular expressions. After the text cleaning steps were completed, the final cleaned output was ready for tokenization.

3.3.2. Concept Extraction, tokenization and embedding layer

The tokenization splits the sentences from the research articles into individual tokens. To extract the concept from the texts, we used the BioPortal API (Noy, et al., 2009) annotator method by using underlying generic and specific ontologies. These concepts per text are saved for further enrichment. For machine learning based classifiers, we used keras tokenizer (Team, K. (n.d.)) and an embedding layer in keras to generate vector embedding. We also iterated with BioBERT tokenization and embedding to generate a vector of 768 dimensions and fed it to machine learning based classifiers.

3.4. Clustering of documents

The purpose of clustering is to group similar research articles and extract related knowledge under one node in the KGs to make it more coherent. As the initial corpus of research articles about coronavirus and subarachnoid hemorrhage is very large, a wide variety of topics are included within the articles. These topics comprised symptoms and diseases, corona vaccines, genetic information, risk factors, treatment methods, shapes of aneurysms, and genetic information and genetic screening. In order to make KGs better structured and easier to search for by following the different nodes specific to each topic, it was necessary to cluster these documents automatically. The source code for clustering is available on GitHub repository ("GitHub - Evidence Based KG.," n.d.) and the step-by-step process of clustering is briefly described below.

3.4.1. Semantic enrichment

The purpose of semantic enrichment is to contextualize the extracted concepts and improve the search and analytics in the KG. Adding metadata and a definition to each of the entities in KGs helps to link them together and create relationships.

The concept extracted from the tokenization step has been further enriched using BioPortal API. We enriched extracted concepts with definitions, synonyms and labels available in the respective ontologies.

Additionally, it strengthened the KGs in terms of robustness by providing contextual meaning to existing concepts.

3.4.2. Vector embedding

The use of embeddings makes machine learning models more efficient and easier to utilize. With the aid of in-built tokenizers and the addition of embedded layers (Ketkar, 2017), the semantically enriched and tokenized concepts were converted to vector representations. Tokenized sentences are converted to vectors of 250 dimensions each based on the tokenization process. We also utilized BioBERT to generate the 768-dimensional vector representation of the texts.

3.4.3. Dimensionality reduction

Principal Component Analysis (PCA) is a method of unsupervised dimensionality reduction in which features are constructed by considering the combination of linear (linear PCA) or nonlinear (kernel PCA) features. Using linear PCA method, we were able to visualize the data as two components and used the dimensionally reduced vector into clustering so that it could be segmented into different topics based on its components.

3.4.4. Cluster identification

The PCA components after dimensionality reduction are fed to a clustering model to identify different clusters pertaining to whole documents. We used the K-means algorithm for clustering and used the elbow method to determine the number of clusters dynamically. Each specific cluster is now mapped back to original texts, and manually reviewed to specify the name of each topic.

3.4.5. Topic modeling

Once the clusters have been identified and the documents have been mapped to each identified cluster, the next step is to obtain a deeper understanding of the meaning of each cluster by selecting the most significant words within it. The documents within each cluster are modeled using Latent Dirichlet Allocation (LDA) algorithm (Blei, et al.,2003,). In LDA topic modeling, each research article is described by distribution of topics and each topic is described by distribution of words.

3.5. PICO classifier

The purpose of the PICO classifier is to loop through each of the topics and categorize the texts into the various PICO elements along with any additional Aim categories that may be present. This categorization creates a subgraph specific to each element, increases its relevance, helps in indexing and improves the response time to query. The PICO classification is done for each identified topic in the cluster identification stage. This is achieved by segregating each topic and its corresponding vectorized corpus input and feeding it into the RNN based PICO classifier model as inputs. The model is composed of five layers - the input layer, the embedding layer, the logical hidden layer, the classification layer, and the output layer. In this approach, the hidden layer is based on Long Short-Term Memory (LSTM). The LSTM contains 100 units, considering X_t as data that entered the memory cell for training, it has three gates to control the flow of information in the network. The forget gate f_t controls how much of the previous state can be retained. The f_t defined as follows:

$$f_t = \sigma\left(W^f x_t + V^f h_{t-1} + b_f\right) \tag{1}$$

The input gate i_t in LSTM cell determines whether to use the current input to update the information of the LSTM or not using the following formulation:

$$i_t = \sigma(W^i x_t + V^i h_{t-1} + b_i) \tag{2}$$

The output gate o_t determines which parts of the current cell state need to be outputted to the next layer for iteration via:

$$o_t = \sigma(W^o x_t + V^o h_{t-1} + b_o) \tag{3}$$

Where W, V, and b are the weight matrix and biases, respectively. Also,σ is the sigmoid activation function which is defined as follows to control the weight of the message passing through.

$$\sigma(X) \;=\; 1/\left(1 + e^{-X}\right) \tag{4}$$

At the end the output h_t in each cell can be retained as:

$$c_t = f_t \otimes c_{t-1} + i_t \otimes tanh(W^c x_t + V^c h_{t-1} + b_c) \tag{5}$$

$$h_t = o_t \otimes tanh(c_t) \tag{6}$$

Where $\otimes$ is the dot product.

3.6. Relationship extractor

This component was designed to identify taxonomic and non-taxonomic relationships between concepts that pertain to PICO elements pertaining to different topics. The detailed process of relationship extraction is given in the following subsections.

3.6.1. Taxonomic relationship extraction

The taxonomic relationship, also referred to as the "IS-A" relation, is one of the most important components of taxonomies, semantic hierarchies, and knowledge graphs. We used the Is-A relationship from the specific and generic ontology to relate different concepts, merge concepts and build KG. For the COVID-19 dataset, we used specific ontologies such as COVIDCRFRAPID, COVID-19, CODO, COVID19, IDO-COVID-19 and generic ontologies SNOMEDCT, MEDDRA. For the aneurysm dataset, we used SNOMEDCT, MEDDRA, NBO, NIFSTD ontologies. The IS-A relationship is extracted by using maximum 3 levels up if available and constructing subgraphs in top-down fashion. The subgraph of concepts is further merged and joined through common entities. For example, one of the sentences in the aneurysm dataset is:

This Mendelian randomization study suggests that high blood pressure is a major risk factor for intracranial aneurysm.

The initial concept extraction contains two concepts *High Blood Pressure* and *intracranial aneurysm,* but after semantically enriching by extracting synonyms, definition and labels the final concepts identified are *High Blood Pressure*, intracranial *aneurysm*, *hypertension*, and *brain aneurysm*.Fig. 2**.** shows the illustration of the merger of concepts based on taxonomic relation extraction and generation of complete subgraphs for the whole sentence.

3.6.2. Non-Taxonomic relationship extraction

A non-taxonomic relationship is a kind of 'part-of' relationship, in which one concept is a part of another. In order to enrich the KGs with non-taxonomic relationships, the main objective is to generate relationship triplets and to identify non-taxonomic relationships. Creation of non-taxonomic relationship between concepts is a two-step process, identifying the concepts with the relationship, followed by identification of the type of pairs with the relationship.

The non-taxonomic RE for KG generation formally defined as below:

Let's consider concepts {Co_1, Co_2, Co_3, …., Co_n} as inputs where n is the number of concepts in the list for a triple generation. The aim is to build a model f: (Co_i, Co_j) → argmax (R^r) where r is the dimension of the output relationships R and argmax is the function that outputs the relationship with maximum probability, i and j is the index of input concepts.

The following algorithm illustrates the main steps in establishing non-taxonomic relationships.

Algorithm 01: Pseudo code to extract non-taxonomic relationships.

1. Inputs: {Co_1, Co_2, Co_3, …., Co_n}
2. Pairs creation: {(Co_1, Co_2), (Co_1, Co_3), (Co_1, Co_4), …. (Co_1, Co_n), …., (Co_{n-1}, Co_n)}

(continued on next page)

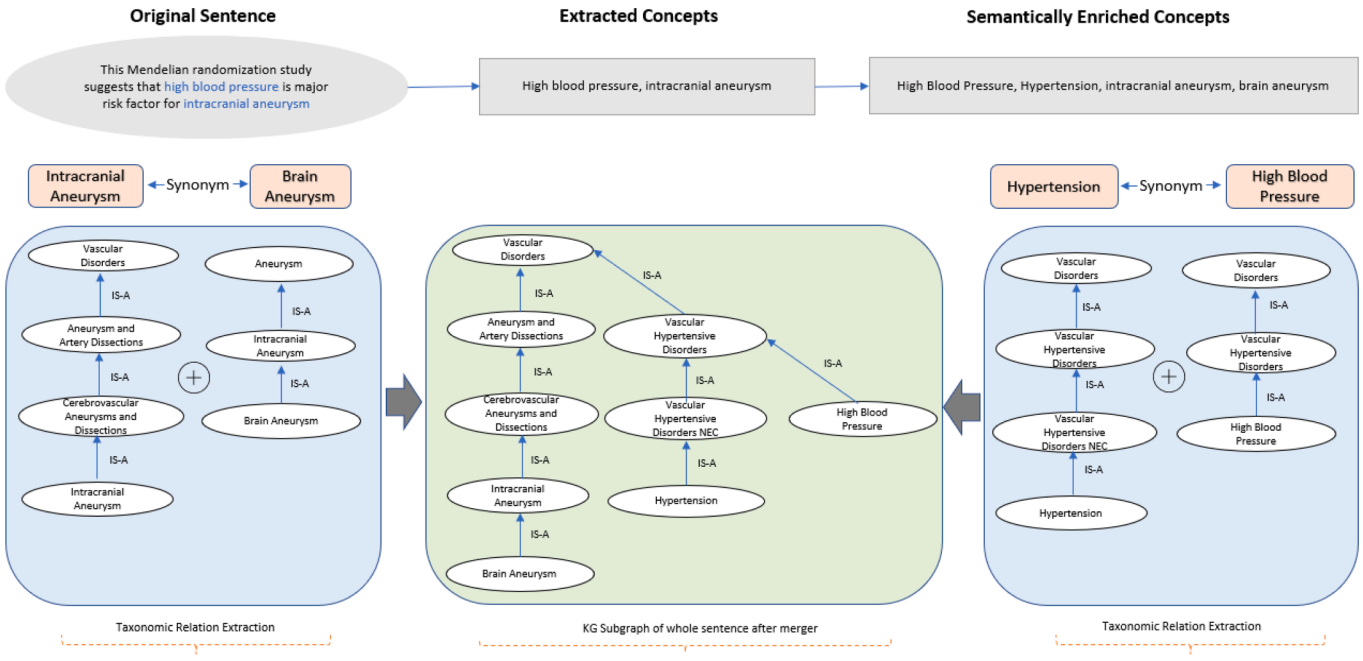


Fig. 2. Sample taxonomic relationship extraction and merger.

(*continued*)

3. Paired relationships detection using model f, if argmax (R^r) < e (where e is the threshold for non-relationship pairs) then the probability for all relationship types is in the minority so the pair of (Co_i, Co_j) do not contain relationship: {(Co_1, Co_2), …. (Co_1, Co_n), …., (Co_{n-1}, Co_n)}
4. Pair relationship identification using model f: {(Co_1, Co_2, R_1), …. (Co_1, Co_n, R_7), …., (Co_{n-1}, Co_n, R_{125})}
5. Generate Output

In earlier approaches, feature-based and deep learning-based methods were tried, but the biggest problem was their representations and domains of use. The direct application of state-of-the-art NLP methodologies to biomedical text mining has some limitations. Although vector representations such as Word2Vec, GLoVe by Stanford, ELMo by AllenNLP, and BERT have been examined in many general domains, it is difficult to estimate their performance on biomedical datasets. The word distributions of general and biomedical corpus also differ significantly, which can present problems for biomedical text mining. With the advancement of NLP and the high performance of the PrLMs, it was demonstrated that they were capable of performing well in a specific domain such as medical texts because the attention mechanism enables them to consider context. One such model, BioBERT, which is a domain-specific language representation for the biomedical domain, has been taken into consideration as producing a representation layer to the RE extraction model.

The non-taxonomic relationship model includes deep learning and PrLMs using the following layers: representation layer, biomedical BERT layer, CNN layer, Bi-LSTM layer, and classifier for biomedical RE. The BioBERT model is based on the attention mechanism and transformer coding structure. BioBERT is used to obtain feature representations containing global semantic and contextual feature information of biomedical text, CNNs has a good performance on features extraction by convolution kernel, which improves the accuracy of feature descriptors. Here CNNs used to obtain features at different levels and combine attention to obtain weighted local features, and fuse global contextual representation with weighted local features. This allows to Bi-LSTM layers to better capture the sequential representation of inputs for better classifications. The detailed architecture for the model is shown in Fig. 3.

3.6.2.1. Representation layer. In the representation layer, words are split into their full forms or word pieces according to a set of rules, which are then encoded into numerical vectors. The representation consists of tokenization and concatenation of tokens. Each concept is tokenized using the BioBERT tokenizer. It produces the token, position, and segment embeddings.

$$Tokenizer(Co) = Co \rightarrow (W_k, P_k, S_k) and S_k, P_k, W_k \in \mathbb{R}^{200}, k \in [1, 200] \quad (7)$$

Where, W, P, and S are token, position, and segment embeddings respectively. Next, embeddings are concatenated to produce the input sequence to the BioBERT module.

$$E_n = Tokenizer(Co_i) \oplus Tokenizer(Co_j), where E_n \in \mathbb{R}^{400} \quad (8)$$

3.6.2.2. BioBERT layer. Biomedical LMs are a natural choice as vectorization of inputs into semantic vectors since the proposed knowledge graph consumes biomedical text as input. A general-purpose language model, Bidirectional Encoder Representations from Transformers (BERT) achieves state-of-the-art performance in a wide range of tasks. The BioBERT is a biomedical variant of BERT that has been trained on medical datasets such as PubMed and PMC. Like BERT, BioBERT takes input sequence tokens and calculates embeddings for them. Our proposed method uses pre-trained BioBERT as the first layer, and we have fine-tuned it for the RE task using feature extractor layers on top to update the LM weights appropriately.

The $V_{BioBert}$ represents the outputs of the BioBERT encoder which will be fed into the next layers for local feature extraction.

$$V_{BioBert} = Encoder_{12}(Encoder_{11}(.....(Encoder_1(E_n))....)), where V_{BioBERT} \in \mathbb{R}^{768} \quad (9)$$

3.6.2.3. Convolutional neural network layer. CNNs (Yamins, Hong, Cadieu, & DiCarlo, 2013) are neural networks which extract local features. This process allows the extraction of high-level features and concept sharing. Thus, the network learns the importance of both concepts regarding relation types. The convolutional layer was followed by a ReLU activation function (Agarap, 2018) on the extracted local features to produce the nonlinear representations. Ultimately, this layer applies a 1D convolution to an input composed of several input planes.

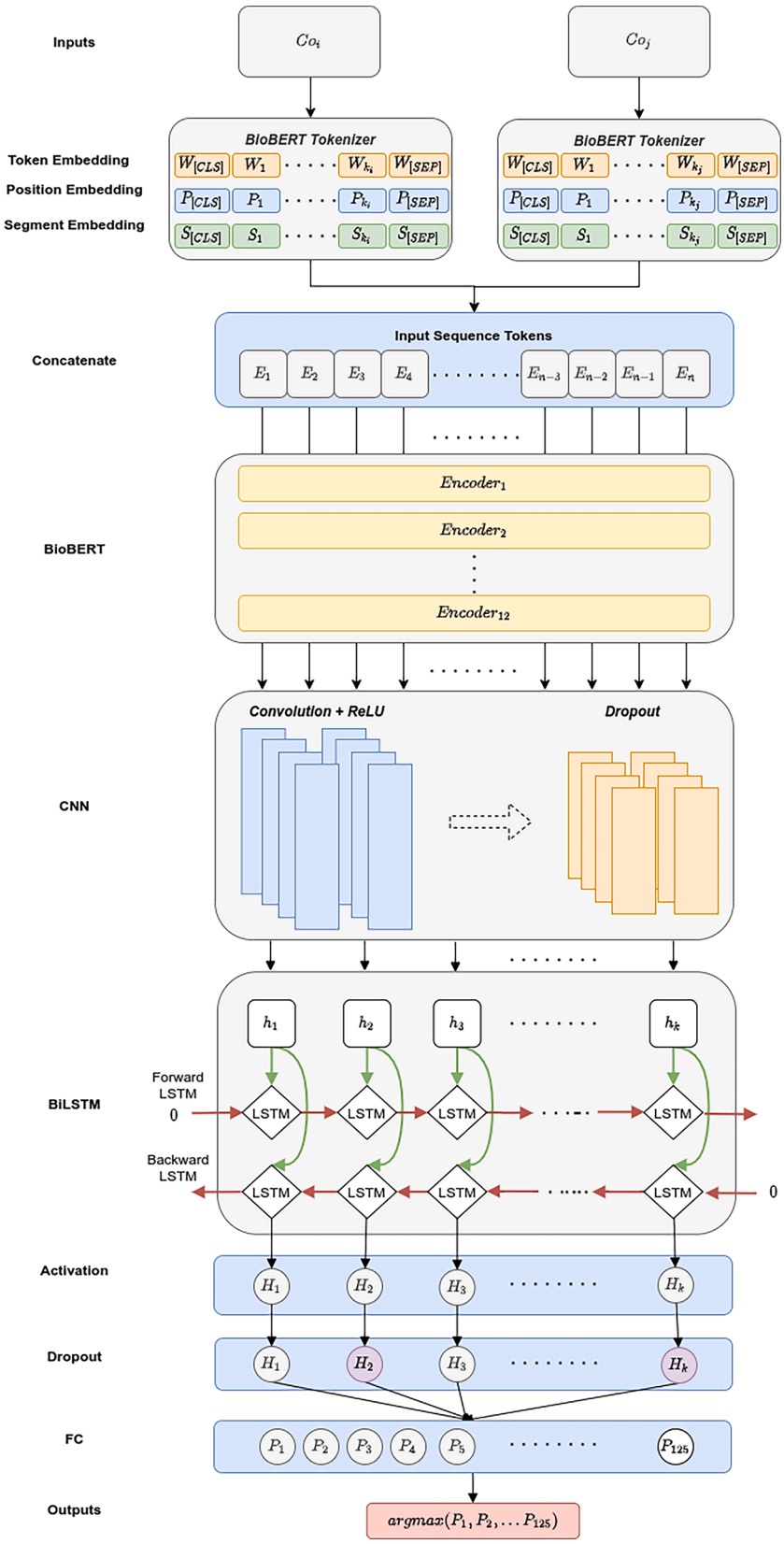


Fig. 3. Architecture diagram of Relationship Extraction Module.

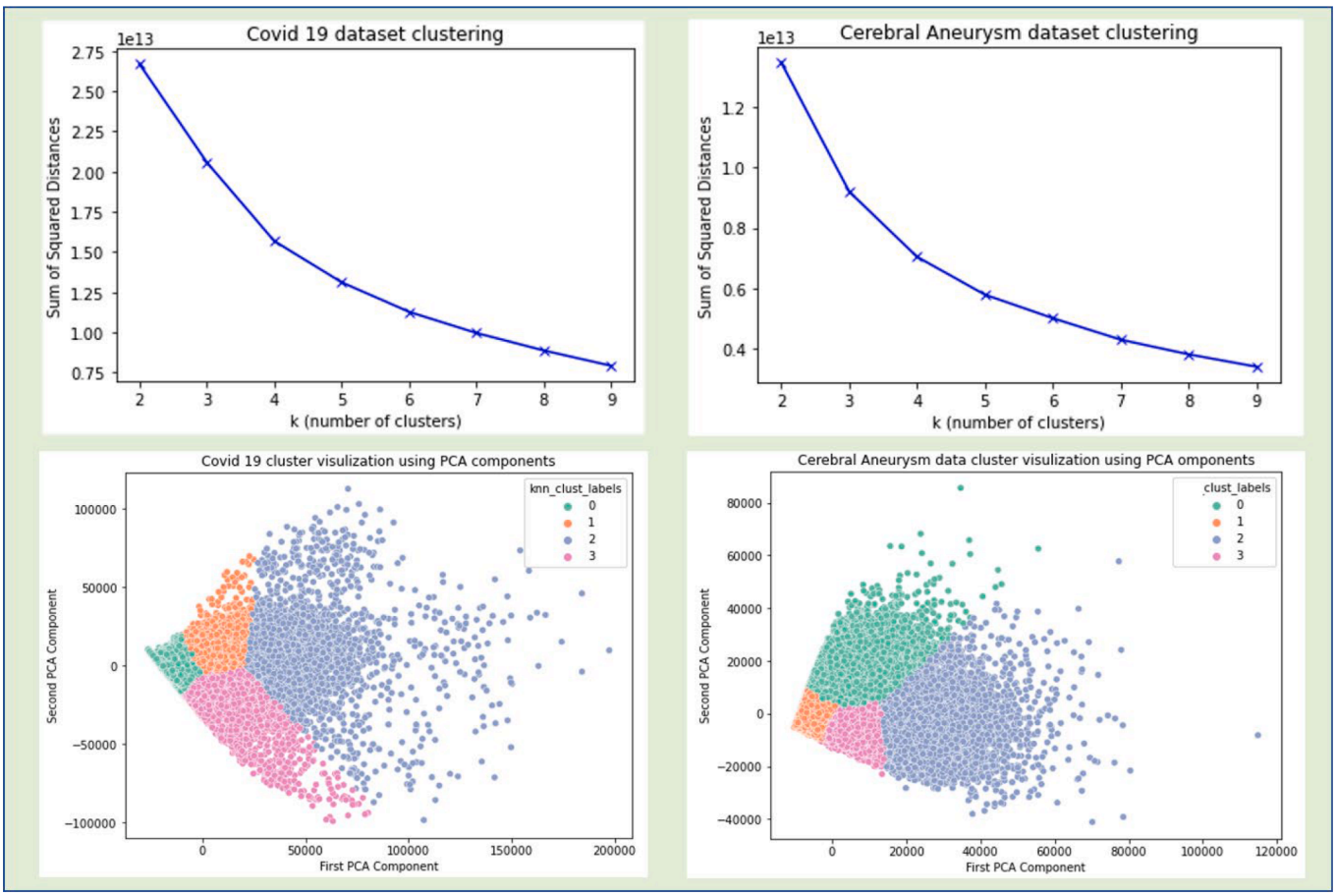


Fig. 4. Cluster optimization and selection using K-means clustering and elbow method.

$$out(N_i, C_{out_j}) = bias(C) + \Sigma^{C_{in}-1}{}_{k=0}(weight(C_{out}, k) \star V_{BioBert}(N_i, k)) \quad (10)$$

Where $\star$ is the cross-correlation operator, N is the batch size, C is a number of channels (C_{in}, C_{out} are number of input and output channels respectively), L is a length of the signal sequence where it is equal to

$$L_{out} = \lfloor (L_{in} + 2*padding - dilation*(\text{ker}nelsize - 1) - 1)/stride + 1 \rfloor \quad (11)$$

Where padding is the same, dilation is 1 (the spacing between kernel elements), kernel size is 3, and the stride is 1. Next, a nonlinearity is applied to $out(N_i, C_{out_j})$ to capture the final local features.

$$CNN_{out} = ReLU(out(N_i, C_{out_j}))whereReLU(X) = Max(X, 0) \quad (12)$$

In the end, dropout randomly zeroes some of the elements of the inputs with probability p = 0.3 using samples from a Bernoulli distribution.

3.6.2.4. Bidirectional LSTM. A bidirectional LSTM on top of a CNN layer eliminates the need for more feature engineering techniques by capturing local features and extracting them for long dependency features. On top of the CNN layer is a BiLSTM layer that is composed of two LSTM networks that can read input texts both forward and backward. The forward LSTM processes information from left to right ($H\rightarrow$) and the backward LSTM processes information from right to left ($H\leftarrow$), Next, the output of forward and backward LSTMs is combined and fed into the next layer. The following equation represents steps:

$$H_t\rightarrow = LSTM(CNN_{out}, H_{(t-1)}\rightarrow \quad (13)$$

$$H_t\leftarrow = LSTM(CNN_{out}, H_{(t-1)}\leftarrow \quad (14)$$

$$BiLSTM_{H_t} = [H_t\rightarrow : H_t\leftarrow]whereBiLSTM_{H_t} \in \mathbb{R}^{n*400} \quad (15)$$

3.6.2.5. Classifier. The CNN-Bi-LSTM layers contain enriched representations of input concepts; the classifier tries to learn these representations with respect to relation types. Backward propagation allows each layer of the network to update its weights. The classifier consists of hidden states h_t as inputs to the ReLU activation function followed by a dropout ($p = 0.3$) each neuron will be zeroed out independently on every forward call. As a next step, we apply a linear y transformation to hidden states, which produces 125 output features from 768 input features. Lastly, the outputs are fed into the argmax function to output the class with maximum probability as the relation type of the input concepts. The equation is described conceptually below and source code of implementation is available on GitHub repository ("GitHub - Relationship Extraction.," n.d.)

$$CNNBiLSTM_{out} = ReLU(BiLSTM_{H_t}) \quad (16)$$

$$Y := bias + CNNBiLSTM_{out} * weight^T \quad (17)$$

$$argmaxY(p) := argmax_{p \in Y} p = \{p \in Y : Y(p) \geq Y(i) foralliinY\} \quad (18)$$

4. Results and evaluation

4.1. Evaluation metrics

The knowledge framework is evaluated on both COVID and aneurysm dataset. Different metrics have been used to evaluate each component of the automated knowledge framework. After calculating the average distance to centroid across all data points across different

cluster values, we applied the elbow method to optimize clustering using the COVID-19 and cerebral aneurysm datasets.

PICO classification models and relationship extractor models are evaluated and compared to baseline machine learning based classifiers using metrics accuracy, precision, recall, and f1 score. In general, the accuracy metric measures the ratio of correct predictions over the total number of instances evaluated.

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (19)$$

The precision is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class.

$$Precision = TP/(TP + FP) \quad (20)$$

The recall is used to measure the fraction of positive patterns that are correctly classified.

$$Recall = TP/(TP + TN) \quad (21)$$

The F1 score metric represents the harmonic mean between recall and precision values.

$$F1 = 2*(Precision*Recall)/(Precision + Recall) \quad (22)$$

Where TP, TN, TP, FN are defined as follows:

- True Positive (TP): Predicted *positive* and actual also *positive*
- True Negative (TN): Predicted *negative* and actual also *negative*
- False Positive (FP): Predicted *positive* but actual *negative*
- False Negative (FN): Predicted *negative* but actual is *positive*.

4.2. Cluster Model

Fig. 4 shows the clustering optimization result of COVID-19 and brain aneurysm dataset. For both the dataset, the sum of squared distances (inertia) falls suddenly at $k = 4$, so the optimized number of clusters is determined as 4 and hence four topics are identified. After reducing the texts represented in vectors to two components, it is evident that there is a clear distinction between optimized cluster numbers.

4.3. Concept Extraction

We evaluated the concept extraction module using randomized samples of 100 abstracts from the corpus. We manually annotated these abstracts, extracted concepts, and compared the results with concepts extracted from BioPortal API, receiving precision 91 %, recall 78 % and F1 Score 84 %. In both datasets, we used generic and specific ontologies, which allowed us to achieve good precision and make sure that no irrelevant terms were extracted.

4.4. PICO Classification

The aim of this experiment was to quantify the comparative analysis of the proposed LSTM model with other baseline classifiers such as Logistic Regression (LR), k-Nearest Neighbors(kNN), Adaptive Boosting, Gradient Boosting, and Multi-Layer Perceptron (MLP) models on both the covid, and aneurysm dataset using keras embedding layer and BioBERT embedding layer for initial input texts. The LR model was initialized with base configuration, KNN model was started with five neighbors and Euclidean as distance measure, Gradient Boosting model was initialized with learning rate = 0.10 and keeping rest of the parameters as default, and lastly MLP was initialized with hidden layer of sizes 100, 70, 30, 20 and 10 neurons, and maximum iteration of 1000. Results of these models are compared on four metrics: precision, recall, F1-score and accuracy as shown in Table 4 and Table 5 on COVID and aneurysm datasets. The proposed model is an RNN based LSMT model using BioBERT tokenization and embedding layer as an input with 82 % of precision, recall, accuracy and F1-score on COVID-19 dataset and 93 % precision, recall, accuracy and F1-score on aneurysm dataset.

Table 4
Comparative results of deep learning with traditional machine learning models for different representation layer on COVID dataset.

Classifier	Embedding Layer	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Keras	38	37	33	42
	BioBERT	77	77	77	77
k-Nearest Neighbors (kNN)	Keras	41	39	39	39
	BioBERT	75	74	74	74
Adaptive Boosting	Keras	37	45	39	45
	BioBERT	58	58	58	58
Gradient Boosting	Keras	48	48	42	48
	BioBERT	64	63	62	63
Multi-Layer Perceptron (MLP)	Keras	8	28	12	28
	BioBERT	63	64	64	64

Table 5
Comparative results of deep learning with traditional machine learning models for different representation layer on aneurysm dataset.

Classifier	Embedding Layer	Precision	Recall	F1-Score	Accuracy
Logistic Regression	Keras	36	42	36	42
	BioBERT	76	74	75	77
k-Nearest Neighbors (kNN)	Keras	35	35	34	35
	BioBERT	80	78	78	80
Adaptive Boosting	Keras	49	50	47	50
	BioBERT	64	66	66	66
Gradient Boosting	Keras	57	49	45	49
	BioBERT	64	61	59	61
Multi-Layer Perceptron (MLP)	Keras	37	29	13	29
	BioBERT	83	83	83	85
Long Short-Term Memory (LSTM)	Keras	85	84	84	84
	BioBERT	93	93	93	93

4.5. Non-Taxonomic relationship extraction evaluations

The aim of this experiment was to quantify the comparative analysis of the proposed RE model with other baseline machine learning model such as Logistic Regression (LR), Multi-Layer Perceptron (MLP) with Latent Semantic Analysis (LSA) encoding, BERT and BioBERT Model with encoding using softmax classifier.

4.5.1. Logistics Regression with Latent semantic analysis

This model uses Latent Semantic Analysis (LSA) (Dumais, 2004) as a representation layer and Logistic Regression (LR) as a classifier. LSA embeds documents in a vector space using a bag of words method. This LSA is based on Term Frequency-Inverse Document Frequency (TF-IDF) and Singular Value Decomposition (SVD). The inputs were represented by 100 unigrams based on TF-IDF with SVD n-components of 50.

4.5.2. Multi-Layer Perceptron with Latent semantic analysis

This model uses LSA as a representation layer and MLP (Multi-Layer Perceptron) as a classifier. The MLP classifier optimizes the log-loss function using adam optimizer where we designed three fully connected dense layers with 500, 300, and 250 neurons in each layer, respectively. The model trained over 50 epochs, batch size of 200 and learning rate of 0.001.

4.5.3. Hypertuned BERT with softmax Layer

A fine-tuned BERT in the usual multi-classification is one of the

baselines in comparison to modified/integrated transformer models. A basic BERT-based model as a baseline for the RE task consists of the BERT, linear, dropout, and linear layers in sequential form (the BERT → Linear1 → Dropout → Linear2 scheme). A BERT outputs a CLS token as an input for the next linear layer, where it takes input features in size 768. Next, a dropout, a regularization method that approximates training a large number of neural networks with different architecture in parallel with a probability of 0.3 has been added. In the end, the second linear layer inputs feature in size 768 and outputs predictions in size of 125 (the number of unique relationships in the whole dataset). The argmax function is applied to the outputs of the second linear layer to the output relationship for input concepts. The training was done using the following parameters: a batch size of 4, Adam optimizer with a learning rate of 10e-05, 2 epoch numbers, and cross-entropy loss function.

4.5.4. BioBERT with softmax layer

It has been trained in the same fashion as we presented in fine-tuning the BERT model with the following setting: BioBERT → Linear1 → Dropout → Linear2. In the first layer instead of the BERT model BioBERT, a medical domain transformer model has been utilized.

The RE model is evaluated against above mentioned baseline models with varying embedding layers and results are shown in Table 6. The relation prediction classification is highly imbalanced, so traditional machine learning models like LR with LSA embedding are not able to capture high performance features. A simple neural network model with LSA embeddings performs better than LSA-LR and it shows that deep learning models are able to capture important features. Also, we observed that domain specific Language Models (LMs) perform better than general purpose models such as BERT.

In the proposed CNN-BiLSTM with BioBERT embedding, we obtained 2 % higher F1-score since CNNs are able to learn higher level features. Furthermore, traditional machine learning models do not perform as well as the proposed method since they are not able to capture high performance features.

4.6. Case Query: Aneurysm and COVID treatment evidence

To test the proposed automated KG generation framework and embedded evidence within it, we executed a simple query relating to COVID-19 and aneurysm: “Aneurysm treatment” and “COVID-19 treatment”. In the subgraph as shown in Fig. 5., it searches through the entire KG and traverses through specific topic *treatment* that was identified through clustering and topic modeling in the framework and provide evidence relating to problem, interventions, comparisons, and outcomes. In the example subgraph for aneurysm treatment, intervention (I) for medium aneurysm at Anterior Communicating Artery (ACoA) with hypertension symptom is identified as endovascular treatment with 88.4% good outcome. The control measure i.e., comparison (C) determined in the KG is surgical clipping with 73.2 % positive outcome (O). Similarly, for COVID-19 the treatment for a patient of age between 56 yrs. and $\leq$ 73 yrs. with diabetes and hypertension, immunomodulators and anticoagulant are identified as intervention (I) with outcome as 2–3 weeks of hospital stay (O). The control measure (C) identified as use of antibiotic and antiviral medication with 3–4 weeks of hospital stay (O).

Table 6
Comparative results of proposed BioBERT encoded CNN based BiLSTM model with other baseline models.

Classifier	Precision	Recall	F1-Score	Accuracy
LSA-LR	16	8	9	45
LSA-MLP	60	34	40	66
BERT	83	74	76	93
BioBERT	93	86	88	95
BioBERT-CNN-BiLSTM	92	90	90	96

5. Discussion

Knowledge Graphs are expected to be an important technology to transform evidence-based medicine. As the underlying data source and inherent information keeps changing in literature, the automated knowledge graph generation framework for evidence-based medicine is unable to generate knowledge that is easy to query. With the proposed architecture of automated KG construction, we made it flexible and adaptable to new sources of information by enabling automatic clustering of new information and assigning it to existing clusters or even creating new clusters followed by PICO classification which in turn makes deployment and storage of information easier and smarter. Authors in (Michel, et al., 2020) created KGs and incorporated PICO elements in it, however the KG construction tool (ACTA) has limitation on working some other types of studies apart from clinical trials. Moreover, KGs lack topic specificity, resulting in delays in query response time, and PICO elements are scattered across KGs, making it challenging to search for specific elements related to a particular topic. Another limitation is that the classification of relations is restricted to the PICO element level, identifying only three types of relations (support, attack, and no relations), rather than considering the identification of relationships (taxonomic, non-taxonomic) between extracted entities. To the best of our knowledge, this is the first attempt to create a framework for topic specific, PICO enabled KG generation framework that is generalizable to different clinical domains. The responsiveness of KGs is an important factor for its performance, (Su, et al., 2016) also mentioned that response within acceptable time generates valuable and timely insights, they try to optimize the to structure the query process to achieve this. In the current framework, we improved response by optimization of subgraphs using topics and enablement of PICO elements. (Özcan, Lei, Quamar, & Efthymiou, 2021) concluded that semantic enrichment enables deep reasoning capability; we achieved it by assimilation of information from generic and specific ontologies. Most of the existing work such as (Rotmensch, Halpern, Tlimat, Horng, & Sontag, 2017; Kamdar, et al., 2021) tried to generate KGs using single iteration of annotation of concepts, this framework uses semantically enriched concepts as final one and creates KGs, and thus increased it semantically by many folds and improves its searchability aspects and ability to provide specific and precise information in response to related queries. The robustness of KGs is further enhanced by inclusion of both taxonomic and non-taxonomic relationships and scaling it both ways horizontally and vertically.

Finally, the use of domain specific and language models for high performance feature extraction of underlying text data improves the performance in ML models, specifically PICO classification and RE model. BioBERT embedded layer and LSTM model improves the performance PICO classification task by 11 % in terms of accuracy for the both COVID-19 dataset and cerebral aneurysm dataset. The RE model performance also improves significantly when BioBERT is used in conjunction with Bi-LSTM and CNN.

6. Limitation and future directions

The architecture of the knowledge graph framework presented here is tested on two types of healthcare datasets and is flexible enough to be applied to other domains. Future work will focus on the application of the architecture framework on areas such as food supply chain, dietary recommendations, agriculture, fisheries etc. As of now, the RE model for non-taxonomic relationship has the limitation of predicting only 125 different types of relationship, and in future with more data and manual annotation of relationship, this could be increased significantly. In addition, the concept extraction module is evaluated on a randomized sample of 100 abstracts due to the effort needed in manual annotation, so adding more abstracts could improve the evaluation. A further improvement to the KG completion process could be made through embedding and link prediction.

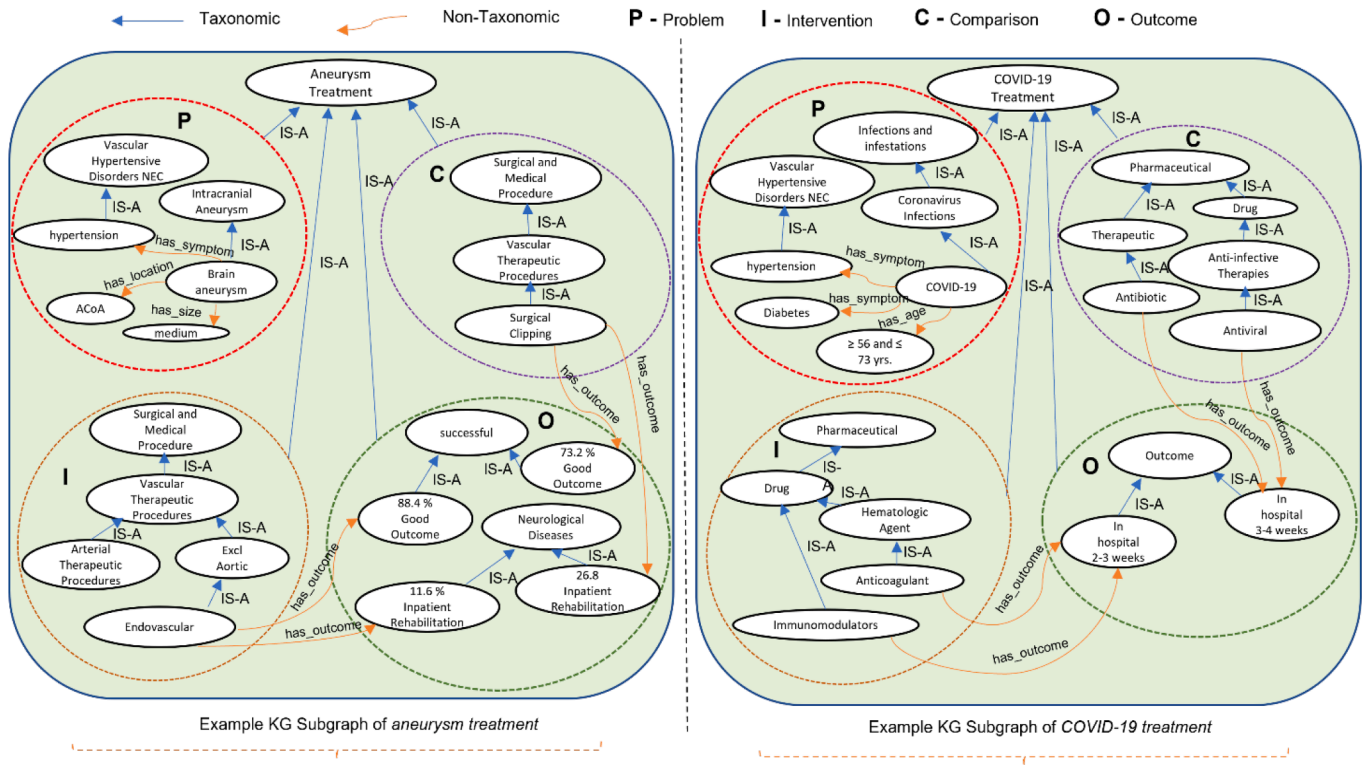


Fig. 5. An excerpt/sub-graph of aneurysm/covid treatment evidence.

7. Conclusion

The use of KGs in the healthcare domain has shown to be a successful method for mapping relationships between data that are extremely diverse and structured. In addition, it provides uncanny capability to model latent relationships between information sources and capture linked information which other data models fail to do. By designing the framework of automated KG curation and creating a pipeline to assimilate data from large corpus research papers, we employed a set of methods for clustering of documents and its optimization using unsupervised learning algorithms such as k-means, concept extraction and its enrichment using well known BioPortal API, PICO elements along with additional Aim category classification using RNNs and BioBERT encoding, taxonomic relationships extraction using ontologies, and non-taxonomic relationship extraction using deep learning architecture of Bi-LSTM and CNN with BioBERT representation layer. With this framework, we have provided other researchers with the freedom to build on, extend, and even use sub-parts of it in a variety of applications such as food preparation, recommendation systems, and fisheries industry etc. Furthermore, we developed a specialized dataset of PICOs (Patient/Problem, Intervention, Comparison, and Outcomes) for use in the COVID-19 domain.

In the proposed framework, the variations of baseline models and embeddings are used to test its different components. The concept extraction using BioPortal API has acceptable precision and recall, the clustering approach is optimized using the elbow method utilizing SSE metric, PICO classification using LSTM and BioBERT embedding shows 11 % improvement in accuracy over the baseline models on both the dataset. The proposed relationship extraction model using CNN, Bi-LSTM, and BioBERT encoding shows improvement in recall by 4 % on other baseline models.

CRediT authorship contribution statement

Fakhare Alam: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing, Validation, Visualization. **Hamed Babaei Giglou:** Software, Visualization, Writing – review & editing, Formal analysis. **Khalid Mahmood Malik:** Funding acquisition, Project administration, Resources, Investigation, Validation.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Khalid Mahmood Malik reports financial support was provided by Deputyship for Research & Innovation, Ministry of Education, Saudi Arabia.

Data availability

Data will be made available on request.

Acknowledgements

This material is based upon work supported by Deputyship for Research & Innovation, Ministry of Education, Saudi Arabia, Grant/Award Number: 959.

References

Afzal, M., Alam, F., Malik, K. M., & Malik, G. M. (2020). Clinical context–aware biomedical text summarization using deep neural network: Model development and validation. *Journal of Medical Internet Research*.

Agarap, A. F. (2018). Deep learning using rectified linear units (relu). . *arXiv preprint arXiv:1803.08375*..

Al-Moslmi, T., Ocaña, M. G., Opdahl, A. L., & Veres, C. (2020). Named entity extraction for knowledge graphs: A literature overview. *IEEE Access*, 32862–32881.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. *The semantic web*, 722–735.

Aydar, M., Bozal, O., & Ozbay, F. (2020). Neural relation extraction: a survey. . arXiv e-prints, arXiv.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit.* ": O'Reilly Media, Inc.".

Blei, D., Ng, A., & Jordan, M. Latent Dirichlet allocation Journal of Machine Learning Research (3). 2003. URL: https://www. jmlr. org/papers/v3/blei03a.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. *ACM SIGMOD international conference on Management of data*, (pp. 1247-1250).

Brychcín, T., Hercig, T., Steinberger, J., & Konkol, M. (2018). Uwb at semeval-2018 task 10: Capturing discriminative attributes from word distributions. *Proceedings of The 12th International Workshop on Semantic Evaluation* , (pp. 935-939).

Buscaldi, D., Schumann, A. K., Qasemizadeh, B., Zargayouna, H., & Charnois, T. (2017). Semantic relation extraction and classification in scientific papers. *International Workshop on Semantic Evaluation*, 679–688.

Chai, X. (2020). Diagnosis method of thyroid disease combining knowledge graph and deep learning. *IEEE Access*, 149787–149795.

Chen, C., Cui, J., Liu, G., Wu, J., & Wang, L. (2020). Survey and open problems in privacy preserving knowledge graph: Merging, query, representation, completion and applications. *arXiv preprint arXiv*.

Chen, C., Ebeid, I. A., Bu, Y., & Ding, Y. (2020). Coronavirus knowledge graph: A case study. *arXiv preprint arXiv*.

Chen, I. Y., Agrawal, M., Horng, S., & Sontag, D. (2020c). Robustly extracting medical knowledge from ehrs: A case study of learning a health knowledge graph. In *Pacific Symposium on Biocomputing* (pp. 19–30).

Chen, Y., Argentinis, J. E., & Weber, G. (2016). IBM Watson: How cognitive computing can be applied to big data challenges in life sciences research. *Clinical Therapeutics*, 688–701.

Daelemans, W., & Bosch, A. (2005). Memory-Based Language Processing . *Studies in Natural Language Processing*.

Dai, Y., Guo, C., Guo, W., & Eickhoff, C. (2021). Drug–drug interaction prediction with Wasserstein Adversarial Autoencoder-based knowledge graph embeddings. *Briefings in Bioinformatics*.

Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., … Kodamullil, A. T. (2021). COVID-19 Knowledge Graph: A computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics, 37* (9), 1332–1334.

Du, J., & Li, X. (2020). A knowledge graph of combined drug therapies using semantic predications from biomedical literature: Algorithm development. *JMIR medical informatics,*.

Du, J., Han, J., Way, A., & Wan, D. (2018). Multi-level structured self-attentions for distantly supervised relation extraction. *arXiv preprint arXiv:1809.00699.*.

Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 188–230.

Eberts, M., & Ulges, A. (2021). An end-to-end model for entity-level relation extraction using multi-instance learning. *arXiv preprint arXiv:2102.05980.*.

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N., & Zettlemoyer, L. (2018). Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv*.

Gatta, R., Vallati, M., Lenkowicz, J., Rojas, E., Damiani, A., Sacchi, L., & Valentini, V. (2017). Generating and comparing knowledge graphs of medical processes using pMineR. *Knowledge Capture Conference* , (pp. 1-4).

geonames. (2022, 11 2). Retrieved 11 2, 2022, from geonames: http://www.geonames.org/.

GitHub - PICO Dataset. (n.d.). Retrieved November 01, 2022, from https://github.com/smileslab/EBM_Automated_KG/tree/main/Covid_PICO_Dataset.

GitHub - Evidence Based KG. (n.d.). Retrieved November 01, 2022, from https://github.com/smileslab/EBM_Automated_KG/tree/main/Knowledge_Generation.

GitHub - Relationship Extraction. (n.d.). Retrieved November 01, 2022, from https://github.com/smileslab/EBM_Automated_KG/tree/main/Relationship_Extraction.

Gyrard, A., Gaur, M., Shekarpour, S., Thirunarayan, K., & Sheth, A. (n.d.). Personalized health knowledge graph. *CEUR workshop proceedings* ..

Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z., & El Asri, B. (2021). BERT based clinical knowledge extraction for biomedical knowledge graph construction and analysis. *Computer Methods and Programs in Biomedicine Update*.

Huang, Z., Yang, J., Harmelen, F. V., & Hu, Q. (2017). Constructing knowledge graphs of depression. In *International conference on health information science* (pp. 149–161).

Jat, S., Khandelwal, S., & Talukdar, P. (2018). Improving distantly supervised relation extraction using word and entity based attention. *arXiv preprint arXiv*.

Ji, G., Liu, K., He, S., & Zhao, J. (2017). Distant supervision for relation extraction with sentence-level attention and entity descriptions. *Proceedings of the AAAI conference on artificial intelligence*.

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Philip, S. Y. (2021). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 494–514.

Johnson, J. M., Narock, T., Singh-Mohudpur, J., Fils, D., Clarke, K. C., Saksena, S., & Yeghiazarian, L. (2022). Knowledge graphs to support real-time flood impact evaluation. *AI Magazine*, 40–45.

Kamdar, M. R., Dowling, W., Carroll, M., Fitzgerald, C., Pal, S., Ross, S., & Samarasinghe, M. (2021). A Healthcare Knowledge Graph-based Approach to Enable Focused Clinical Search. *ISWC*.

Ketkar, N. (2017). Introduction to keras. *Deep learning with Python* .

Kim, T., Yun, Y., & Kim, N. (2021). Deep learning-based knowledge graph generation for COVID-19. . *Sustainability*.

Kuhn, P., Mischkewitz, S., Ring, N., & Windheuser, F. (2016). Type inference on Wikipedia list pages. *Informatik*.

Lai, S., Leung, K. S., & Leung, Y. (2018). SUNNYNLP at SemEval-2018 Task 10: A support-vector-machine-based method for detecting semantic difference using taxonomy and word embedding features. In *Proceedings of the 12th international workshop on semantic evaluation* (pp. 741–746).

Lee, J., Seo, S., & Choi, Y. S. (2019). Semantic relation classification via bidirectional lstm networks with entity-aware attention using latent entity typing. *Symmetry*.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT :A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 1234–1240.

Lenat, D. (1995). A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 33–38.

Li, L., Wang, P., Yan, J., Wang, Y., Li, S., Jiang, J., & Liu, Y. (n.d.). Real-world data medical knowledge graph: construction and applications. *Artificial intelligence in medicine*.

Li, N., Yang, Z., Luo, L., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2020). KGHC: A knowledge graph for hepatocellular carcinoma. *BMC Medical Informatics and Decision Making*, 1–11.

Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). Neural relation extraction with selective attention over instances. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Liu, C., Sun, W., Chao, W., & Che, W. (2013). Convolution neural network for relation extraction. In *International conference on advanced data mining and applications* (pp. 231–242).

Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT Technology Journal*, 211–226.

Liu, W., Yin, L., Wang, C., Liu, F., & Ni, Z. (2021). Multitask healthcare management recommendation system leveraging knowledge graph. *Journal of Healthcare Engineering*.

Ma, Y., Tresp, V., & Daxberger, E. A. (2019). Embedding models for episodic knowledge graphs. *Journal of Web Semantics*.

Malik, K. M., Krishnamurthy, M., Alobaidi, M., Hussain, M., Alam, F., & Malik, G. (2020). Automated domain-specific healthcare knowledge graph curation framework: Subarachnoid hemorrhage as phenotype. *Expert Systems with Applications, 145*.

Malik, K. M., Anjum, S. M., Soltanian-Zadeh, H., Malik, H., & Malik, G. M. (2018). A framework for intracranial saccular aneurysm detection and quantification using morphological analysis of cerebral angiograms. *IEEE Access, 6*, 7970–7986.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstration* (pp. 55–60).

Michel, F., Gandon, F., Ah-Kane, V., Bobasheva, A., Cabrio, E., Corby, O., ... & Winckler, M. (2020). Covid-on-the-Web: Knowledge graph and services to advance COVID-19 research. In The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II 19 (pp. 294-310). Springer International Publishing.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP* (pp. 1003–1011).

Mohamed, S. K., Nováček, V., & Nounu, A. (2020). Discovering protein drug targets using knowledge graph embeddings. *Bioinformatics*, 603–610.

Mohammadhassanzadeh, H., Abidi, S. R., Van Woensel, W., & Abidi, S. S. (2018). Investigating plausible reasoning over knowledge graphs for semantics-based health data analytics. In *27Th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE)* (pp. 148–153).

Mohit, B. (2014). Named entity recognition. *Natural Language Processing of Semitic Languages*, 221–245.

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 217–250.

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., & Musen, M. A. (2009). BioPortal: Ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*.

Oram, P. (2001). WordNet: An electronic lexical database. Christiane Fellbaum (Ed.). *Applied Psycholinguistics*, 131-134.

Özcan, F., Lei, C., Quamar, A., & Efthymiou. (2021). Semantic enrichment of data for AI applications. In *Proceedings of the Fifth Workshop on Data Management for End-To-End Machine Learning* (pp. 1–7).

Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction: A survey. *arXiv preprint arXiv*.

Pham, T. T. (2022). Graph-based multi-label disease prediction model learning from medical data and domain knowledge. *Knowledge-Based Systems*.

Postiglione, M. (2021). Towards an Italian Healthcare Knowledge Graph. In *International Conference on Similarity Search and Applications* (pp. 387–394).

Qiu, J., Du, Q., Yin, K., Zhang, S. L., & Qian, C. (2020). A causality mining and knowledge graph based method of root cause diagnosis for performance anomaly in cloud applications.

Rastogi, N., & Zaki, M. J. (2020). Personal Health Knowledge Graphs for Patients. *arXiv preprint arXiv*.

Reese, J. T., Unni, D., Callahan, T. J., Cappelletti, L., Ravanmehr, V., Carbon, S., & Mungall, C. J. (2021). KG-COVID-19: A framework to produce customized knowledge graphs for COVID-19 response. *Patterns*.

Rotmensch, M., Halpern, Y., Tlimat, A., Horng, S., & Sontag, D. (2017). Learning a health knowledge graph from electronic medical records. *Scientific reports*, 1–11.

Santus, E., Biemann, C., & Chersoni, E. (2018). Combining vector-, pattern-and graph-based information to identify discriminative attributes. *arXiv preprint arXiv: 1804.11251*..

Shen, I., Zhang, L., Lian, J., Wu, C. H., Fierro, M. G., Argyriou, A., & Wu, T. (2020). In search for a cure: Recommendation with knowledge graph on CORD-19. In *26Th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 3519–3520).

Smirnova, A., & Cudré-Mauroux, P. (2018). Relation extraction using distant supervision. *: A survey. ACM Computing Surveys (CSUR),* .

Speer, R., & Lowry-Duda, J. (2018). Luminoso at semeval-2018 task 10: Distinguishing attributes using text corpora and relational knowledge. *arXiv preprint arXiv: 1806.01733*.

Su, Y., Sun, H., Sadler, B., Srivatsa, M., Gür, I., Yan, Z., & Yan, X. (2016). On generating characteristic-rich question sets for qa evaluation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Suchanek, F. M., Kasneci, G., & Weikum, G. (2007). Yago: A core of semantic knowledge. In *16Th international conference on World Wide Web* (pp. 697–706).

Team, K. (n.d.). *Keras documentation: Text data preprocessing*. Retrieved November 01, 2022, from https://keras.io/preprocessing/text/.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 78–85.

Wang, H., Wu, Z., Jiang, H., Huang, Y., Wang, J., Kopru, S., & Xie, T. (2021). Groot: An event-graph-based approach for root cause analysis in industrial settings. In *36Th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 419–429).

Wang, L. L., & Kohlmeier, S. (2020). Cord-19: The covid-19 open research dataset. .

Wang, L., Cao, Z., De Melo, G., & Liu, Z. (2016). Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*..

Wang, L., Xie, H., Han, W., Yang, X., Shi, L., Dong, J., & Wu, H. (2020). Construction of a knowledge graph for diabetes complications from expert-reviewed clinical evidences. *Computer Assisted Surgery*, 29–35.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., & Onyshkevych, B. (2020). COVID-19 literature knowledge graph construction and drug repurposing report generation. *arXiv preprint arXiv*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 1–9.

Wise, C., Ioannidis, V. N., Calvo, M. R., Song, X., Price, G., Kulkarni, N., & Karypis, G. (2020). COVID-19 knowledge graph: accelerating information retrieval and discovery for scientific literature. *arXiv preprint arXiv*.

Wu, S., & He, Y. (2019). Enriching pre-trained language model with entity information for relation classification. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 2361–2364).

Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. In *ACM SIGMOD international conference on management of data* (pp. 481–492).

Xi, J., Ye, L., Huang, Q., & Li, X. (2021). Tolerating data missing in breast cancer diagnosis from clinical ultrasound reports via knowledge graph inference. In *27Th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 3756–3764).

Xiao, M., & Liu, C. (2016). Semantic relation classification via hierarchical recurrent neural network with attention. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1254–1263).

Xing, R., Luo, J., & Song, T. (2020). BioRel: Towards large-scale biomedical relation extraction. *BMC bioinformatics*.

Xu, B., Xu, Y., Liang, J., Xie, C., Liang, B., Cui, W., & Xiao, Y. (2017). CN-DBpedia: A never-ending Chinese knowledge extraction system. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 428–438).

Yamins, D. L., Hong, H., Cadieu, C., & DiCarlo, J. J. (2013). Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. *Advances in Neural Information Processing Systems*.

Yu, T., Li, J., Yu, Q., Tian, Y., Shun, X., Xu, L., & Gao, H. (2017). Knowledge graph for TCM health preservation: Design, construction, and applications. *Artificial Intelligence in Medicine*, 48–52.

Yu, Y., Huang, K., Zhang, C., Glass, L. M., Sun, J., & Xiao, C. (2021). SumGNN: Multi-typed drug interaction prediction via efficient knowledge graph summarization. *Bioinformatics*, 2988–2995.

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1753–1762).

Zeng, W., Lin, Y., Liu, Z., & Sun, M. (2016). Incorporating relation paths in neural relation extraction.

Zhang, D., & Wang, D. (2015). Relation classification via recurrent neural network. . *arXiv preprint arXiv:1508.01006*..

Zhang, Y., Sheng, M., Zhou, R., Wang, Y., Han, G., Zhang, H., & Dong, J. (2020). HKGB: An inclusive, extensible, intelligent, semi-auto-constructed knowledge graph framework for healthcare with clinicians' expertise incorporated. *Information Processing & Management*.

Zhu, Y., Che, C., Jin, B., Zhang, N., Su, C., & Wang, F. (2020). Knowledge-driven drug repurposing using a comprehensive drug knowledge graph. *Health Informatics Journal*, 2737–2750.