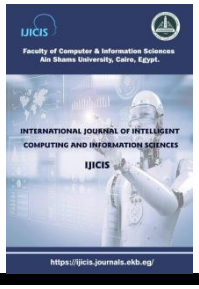
IJICIS, Vol.23, No.1, 1-17
DOI: 10.21608/ijicis.2023.179031.1236

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/



WEIGHTED ENTITY LINKING AND INTEGRATION ALGORITHM FOR MEDICAL KNOWLEDGE GRAPH GENERATION

Noura Maghawry — The British University in Egypt — noura.elmaghawry@bue.edu.eg
Manal Tantawi — The British University in Egypt — samy.ghoniemy@bue.edu.eg
Eman Shabaan — Ain Shams University — eman.shaaban@cis.asu.edu.eg
Karim Emara — Ain Shams University — karim.emara@cis.asu.edu.eg

Received 2022-12-05; Revised 2022-12-23; Accepted 2023-01-14

Abstract: *Semantic data integration is the process of interrelating information from multiple heterogeneous resources. There is a need for representing data concepts and their relationships to eliminate heterogeneity among different data sources in healthcare management systems. Standardized medical ontologies provide predefined medical vocabulary serving as a stable interface for concepts related to medical data sources. However, different ontologies have different concepts although these concepts have logical relations between them such as the Human Disease Ontology and the Symptoms ontology. There aroused a need for a knowledge graph providing a reliable knowledge base for any intelligent healthcare expert advisor disease prediction system. The knowledge graph provides a model for linking and integrating different concepts having logical relationships such as diseases and their symptoms. Medical online website and encyclopedia provides a reliable source for building such a knowledge graph. The knowledge graph is enriched with social networks data where information extracted reflects a major source of data based on user experiences. The paper proposes a framework for constructing a disease-symptom entity linked knowledge graph based on online medical encyclopedia and social networks user experiences. Entity linking such an integrated knowledge graph with standardized medical ontologies makes it a reliable knowledge base for a standard system that could be used by social networks user and the professional staff.*

Keywords: *Medcial Ontologies, Knowledge graph construction, Entity linking, Semantic Data Integration.*

1. Introduction

Healthcare has been provided mainly in a reactive manner. With the progress in the field of Artificial intelligence and the widespread of use of social networks, the world has offered more opportunities for healthcare systems to observe individual physiological signals and provide healthcare in a more proactive manner [1]. Adding social network users' experiences brings a challenging level for personalized healthcare and intelligent expert advisor monitoring systems. Nowadays, intelligent healthcare expert advisor systems need to be built on different heterogenous data sources, involving

medical facts and data gathered from users' experiences which would enrich these facts with more information. To gather and analyze these heterogenous data, systems should utilize semantic integration to transform data into knowledge and intelligence within the healthcare field [2,3]. Relying on semantic-based techniques to manage heterogeneous data in the healthcare domain provides medical doctors and normal users an early disease prediction if any disease ailment is detected. Intelligent healthcare systems need to rely on trustful resources and integrate these resources as a knowledge base for such systems. One of these resources should be medical facts gathered from reliable resources and linked to standard terms and terminologies that both social networks' users and professional staff could understand. Standardized medical ontologies provide a reliable and standard representation of the concepts of medical terminologies and the relation between them. However, these ontologies have different unrelated concepts although these concepts have logical relationships such as the relationship of a disease with its symptoms.

There are several standard online ontologies representing concepts of medical domain. The importance of an ontology provides a predefined and rich vocabulary serving a stable interface for concepts of the data sources and at the same time is independent of the database schemas [4]. In this paper, the framework proposed focus on relating two medical ontologies which are the Human disease ontology (DO) [5, 6] representing diseases' concepts only and their subsumption relationships, and the Symptoms ontology (SYMP) [7] representing symptoms' concepts only and their subsumption relationships. Human Disease Ontology (DO) is a project created at the Institute for Genome Sciences, at the University of Maryland School of Medicine. This project was developed in 2003 at Northwestern University, it was initiated by the need for an ontology covering diseases' concepts. Disease Ontology is an Open Biomedical Ontologies (OBO) Foundry ontology. The ontology involves a property for each disease concept giving it a unique identifier (DOID) consisting of a prefix DOID followed by a number, and for each disease concept there exists an Internationalized Resource Identifier (IRI) that provides a URI containing the disease information. The core relationship in this ontology is the subsumption relationship between disease concepts. However, DO doesn't state the diseases' symptoms, their causes, or any other information except the diseases' concepts. The ontology also provides the synonyms terms for each disease concept and cross-references to similar concepts in other medical ontologies. The second ontology is the Symptom Ontology (SYMP) which is an ontology of diseases' symptoms, representing symptoms which encompass any perceived changes in function, sensations or appearance reported by a patient indicating a disease. The ontology concepts have also a unique symptom identifier for each symptom, that consists of a prefix SYMP followed by a number. For each symptom there is also an IRI that provides information about the symptom through URI. However, symptoms ontology doesn't relate the symptoms to their diseases. The ontology provides the synonyms terms for each symptom and cross-references for similar concepts with other medical ontologies. Both DO and SYMP are ontologies provided by the OBO Foundry – an organization for building and maintaining ontologies related to the life sciences especially biomedical field.

There is a need to automatically integrate these two ontologies. For this integration to occur, other reliable data sources are needed to be semantically integrated with standardized data sources. Online medical data sources contain medical facts, and health related content created by medical professionals. One of the most popular and reliable online encyclopedias for disease and their conditions and

providing medical facts is the MayoClinic website [8] which provides information of interest to this research about diseases, their symptoms, their causes, risk and prevention factors. MayoClinic website provides comprehensive guides on hundreds of diseases and their conditions. It presents the same information provided by the Centers for Disease Control and Prevention (CDC); however, it is used more by social networks users as it provides content in a more reachable, searchable, and user-friendly navigation options. One of the best features of this site is that its data is always updated. The website provides healthcare information in different languages. MayoClinic ranks third in the most visited health websites ranking analysis of November 2022 [9]. It also ranks fourth in the most popular healthcare websites [10].

The aim of the proposed framework is to automatically construct a knowledge graph by extracting the corresponding domain information from the two ontologies – DO and SYMP - and acquiring relationships between their distinct concepts using information extracted from online medical website and enrich this information with social networks users' experiences.

Knowledge graphs (KGs) are used to describe and organize real-world entities and their interrelations visualized in a graph. A knowledge graph (KG) is considered a dynamically growing semantic network of facts about things. KGs provide additional features than ontologies by representing real world instances and data. Thus, KGs add levels of extra information and real-world experiences enriching the basic concepts extracted from a specific domain of interest [9, 10]. A KG is one form of representing knowledge base for an intelligent healthcare expert advisor system, where the expert system inference engine extracts new rules depending on pattern matching techniques within the KG. The fully automated KGs generation of domain specific from unstructured text is a hot topic research field, particularly in the medical field where there is a need to link diseases with their symptoms for any healthcare application. Due to the complexity brought on by the heterogeneity of medical concepts and resources, reaching a standard knowledge base for disease diagnosis and prediction need medical expertise and professional human intervention. This makes it a challenging task to automate the process of creating healthcare systems when there is no standard automatically generated base for such systems.

Such knowledge base provides a representation for the heterogeneous data resources, having the ability to dynamically grow as more data is provided, and could be visualized. Knowledge graph construction involves data acquisition from different sources whether structured and unstructured resources while extracting relationships that usually requires human professional intervention. Figure 1 illustrates the basic structure of an intelligent expert advisor healthcare and shows how the knowledge base is dependent on an integrated knowledge graph built from different heterogeneous resources.

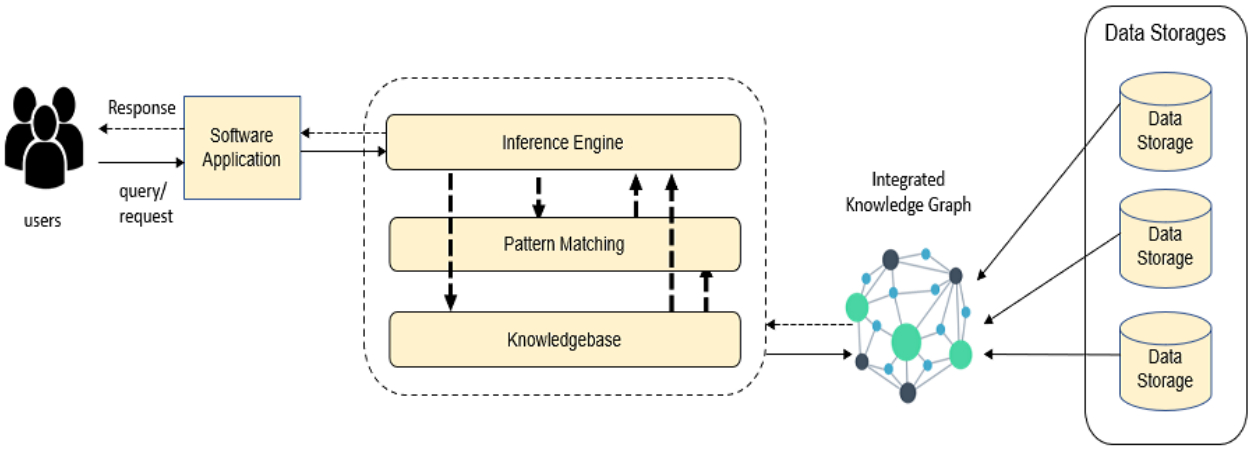


Figure. 1: Structure of an intelligent expert advisor healthcare system.

The integration methodology adopted in building the knowledge graph relies on reliable medical encyclopedia website MayoClinic and entity linking with two standardized ontologies – DO and SMP- for linking diseases from the DO with their symptoms in SYMP. The constructed KG is a disease symptom knowledge graph where graph nodes are concepts of diseases, symptoms, diseases' causes, prevention factors and risk factors. Whereas the graph edges represent the relationships between the disease and its related concepts. The entity linked knowledge graph is integrated with data extracted from medical web forums showing the impact of social networks' users experiences on enriching the knowledge graph. This integrated knowledge graph aims to be used as a knowledge base for any intelligent healthcare system that could predict diseases or give alerts whenever any disease ailments are detected [11, 12]. The system could be used by social networks users and medical professionals, taking into consideration user experiences [15]. Entity linking or entity normalization [16] is the task of using the ontology concepts as a dictionary containing a set of entities E and given a text containing a set of entity mentions M, each mention m ϵ M is mapped to its corresponding entity e where e ϵ E.

The outcome from the proposed framework is a constructed knowledge graph G consisting of set of nodes N representing the MayoClinic entities, set of edges representing the relationships R between different entities, it is denoted as G = < N ,R >. The nodes in Graph G represents disease, diseases' symptoms, their causes, disease' risk and prevention factors. The set R represents the relationships between a disease with its symptoms, causes, risk factors and preventions factors. Entity linking algorithm is adopted to perform linking between disease nodes and symptom nodes with their corresponding entities from DO and SYMP ontologies. Integration algorithm with subgraph generated from medical web-forums is also adopted to add the impact of social networks' users' experiences to the knowledge graph generated.

This paper involves five sections. Section I is the introduction; section II discusses related work. The proposed framework for constructing the KG is described in section III, the results is discussed in section IV. Finally, section V presents the conclusion and future work.

2. Literature Reviews

The most recent research focused on general domain knowledge graph construction, such as the work presented in [17] where authors used texts extracted from Wikipedia to extract concepts and relations for constructing an ontology. The work is done using a supervised machine learning technique requiring huge effort for labeling and validation of data manually. In the work presented in [18] a framework for generating ontologies for organizations is proposed. This proposed framework couldn't be adopted for the medical field, as medical field has special terminologies with synonyms. As a result, general domain knowledge graph construction methodologies were proved not to be effective with the medical field as the concepts are not available normally within English corpuses. Some work focused on building knowledge graphs for specific disease such as Alzheimer disease where an ontology generation system presented in [19], in this work domain experts are involved all during the development process, so human professional intervention is involved all through the process. The need for automatic generation

methods for domain independent generation are presented in [20]. They identified biomedical concepts using Linked open Data (LOD) and linked medical knowledge bases, they used linked Unified Medical Language System and applied semantic for concepts enrichment, however their approach cannot be generalized and cannot lead to specifically generate a knowledge graph for disease and symptoms, as it focuses more on biomedical concepts. Of the recent work presented in [21] an approach for constructing ontologies using deep learning techniques is illustrated, however it still suffers from the challenges of the medical field such as data heterogeneity, and the essential need for medical professional intervention. Disease-Symptom Ontology (DS-Ontology) was proposed in [22], it covers the linking between few diseases and their symptoms and entity linking them to DO and SYMP ontologies, thus integrating the two ontologies for few diseases, the approach was manually done, and it covers very limited diseases. Another work was proposed in [23] which was remarkable work involving an Open Information Extraction system based on unsupervised learning. The work didn't depend on a prebuilt dataset; however, it obtained a knowledge graph from a huge amount of text documents about COVID-19, however the study focused only on one disease which COVID-19. In the work of [24] , a computational framework was designed for detecting drug combinations, by extracting drug names from biomedical publications and treatment sections of clinical trial records, a network model is constructed representing the drug names and their associations. The previous work was extended in [25] where an algorithm for constructing a knowledge graph from drug, gene, and disease mentions in the biomedical literature is presented with two querying algorithms for searching the knowledge graph by a single drug or a combination of drugs. Then comes the role of using deep learning techniques for natural language processing (NLP) which has an important role in building ontologies, the work of [26] presented a system using an external domain knowledge for word embeddings enriching using deep learning model for NLP tasks for cancer phenotyping. The system uses Unified Medical Language System concepts and vocabularies for word representations enrichment. An intelligent health diagnosis technique is proposed in [27] where an expert system with an inference engine for answering queries and gathering information from distinct biomedical ontologies, thus automatically generate an ontology, however, this system depends on gathering information based on queries. The system generated an ontology called HDDO which is an ontology for personal health diagnosis, and it basically uses the user's input queries and personal data to identify possible diagnoses. In [28] a human disease symptom network was built based on using large scale medical bibliographic records collected from PubMed, to generate a symptom-based network of human diseases - Human Symptoms Disease Network (HSDN).

The work depended on the stated diseases in PubMed abstracts which doesn't cover all diseases and their symptoms and doesn't cover a lot of common disease and symptoms. A knowledge database of disease-symptom built based on associations generated by an automated method based on information in textual discharge summaries of patients at New York Presbyterian Hospital admitted, the associations were applied on 150 frequent diseases from the hospital records [29], it is a limited dataset. When it comes to considering social networks experiences, there is the work of [30] where disease-symptom knowledge graph (DSKG) is constructed as a cause-effect knowledge graph containing disease-symptom relations as a cause-effect relation type determined from downloaded medical web-board resources.

Based on the related work surveyed, generating a disease-symptom knowledge graph that also takes into consideration diseases synonyms, symptoms synonyms, diseases' causes, risk, and prevention factors is still open research. Integrating distinct concepts and linking such knowledge graph entities with standardized ontologies is still a challenging research field. Such generated knowledge graph could be a standard knowledge base for many healthcare applications, and our proposed framework is one step on the way.

3. The proposed framework for Entity Linked knowledge graph construction

The proposed framework aims to automatically construct a knowledge graph as a base for any intelligent expert advisor healthcare system. The framework depends on three reliable resources for medical concepts. Another fourth resource for extracting medical facts and their relationships is used. The framework also considers the user experiences available in healthcare forums which enriches the knowledge graph with personal user experiences. The three reliable resources for medical concepts are the Human Disease ontology, the Symptoms ontology and **UMLS** Meta thesaurus [31]. **DO and SYMP ontologies not interlinked till now and integrating them is still a challenging task. The third reliable resource is the UMLS** Meta thesaurus which is a large biomedical thesaurus that is organized by concept or meaning, it links **synonymous names** from over 200 different source vocabularies. The Meta thesaurus also identifies useful relationships between concepts preserving the meanings, the concept names, and the relationships from each vocabulary. The proposed framework makes use of the UMLS concepts for concept mapping. The framework relies in getting the information about the disease, their symptoms, diseases' causes, their risk factors, and their prevention factors from a fourth resource, a reliable medical online encyclopedia which is Mayo-clinic website. MayoClinic website is a reliable scientific encyclopedia for diseases created by nonprofit American academic medical center focusing on integrated healthcare, education, and research. The framework takes also advantages of enriching the knowledge graph with user experiences extracted from web-forums including md-talks forums [32], e-health forums [33]and webmd message boards [34].

The proposed framework for constructing the entity linked knowledge graph is composed of four phases – Phase one is a MayoClinic knowledge graph generator where data is gathered about diseases, symptoms, causes, prevention, and risk factors from the MayoClinic website and generates a knowledge graph based on the data gathered. Phase two is concept extraction from standardized medical ontologies where concepts are extracted from the standardized disease and symptom ontologies are processed, and these concepts are cross mapped with the UMLS meta-thesaurus. The third phase is generating web-based graphs and linking them to standardized ontologies. The fourth phase is the Entity Linking and Integration phase where entity linking algorithm is adopted to link nodes from MayoClinic-based knowledge graph to the entities from the standardized ontologies. The knowledge graph is then integrated with the web-based graphs. The integrated knowledge graph by the end is a disease symptom knowledge graph with interlinked nodes to standardized ontologies and enriched with the social networks' user . The framework for generating the integrated entity linked knowledge graph is shown in figure 2.

3.1 Mayo-clinic knowledge graph generator

Initially, the MayoClinic web pages for diseases and conditions are fed to the crawler as seed URLs. The crawler starts by crawling the pages of all diseases from the diseases and conditions on the website crawling from pages of diseases starting with letter A till diseases starting with letter Z. The parser parses the relevant information using regular expressions. The relevant information of interest in the page are the title of the disease, the list of its symptoms, the disease causes, prevention factors and risk factors. Text pre-processing techniques are applied during the parsing stage where all parsed information is preprocessed by removing punctation, brackets, apostrophe-s and s-apostrophe. This step is checked and compared with the web site information until an accepted accuracy level of parsed data is reached. The pre-processing step is followed by data filtration step where each disease, each symptom, each cause, each risk factor, and each prevention factor are given a unique number identifying it so that each item is represented once in the generated MayoClinic-based knowledge graph. The data filtered step is essential for creating and extracting relationships, based on core Resource Description Framework RDF triples where four types of triples are considered for each disease node - “has_symptom”, “caused_by”, “prevented_by”, “has_risk”. Figure 3 shows the basic RDF for the knowledge graph with the four types of triples.
The output of this stage is an online-based knowledge graph composed of triples based on the content parsed and processed from MayoClinic encyclopedia. The graph resulted in 9387 nodes of 5 labels – ‘Disease’, ‘Symptom’, ‘Cause’, ‘PreventionFactor’ and ‘RiskFactor’. The graph involves 11539 relationships of 4 distinct types – ‘has_symptom’. ‘caused_by’, ‘prevented_by’, and ‘has_risk’. Figure 4 shows a subgraph of the website-based generated knowledge graph focusing on disease ‘lung cancer’ with all its relationships with other nodes representing lung cancer symptoms, causes, risk factors and prevention factors as stated in the MayoClinic website.

3.2 Concept Extraction from Ontologies

The second phase starts by extracting concepts from the DO and SYMP ontologies. Each concept in the ontology has its own properties involving the concept name and the synonyms of the same concept, the concept description, concept unique identifier (CUI) , the internationalized resource identifier (IRI) along with the cross-references to UMLS meta thesaurus concepts. Text pre-processing is then applied on the concepts’ names and their synonyms where the list of synonyms for each disease or symptom is given the same unique identifier as their original concept. The diseases and symptoms’ names, and synonyms are preprocessed by removing punctation, removing brackets, s-apostrophe and apostrophe-s. Each of the disease and symptom concepts are multi-term expressions. The concepts extracted with all the relevant information are stored in a dictionary.

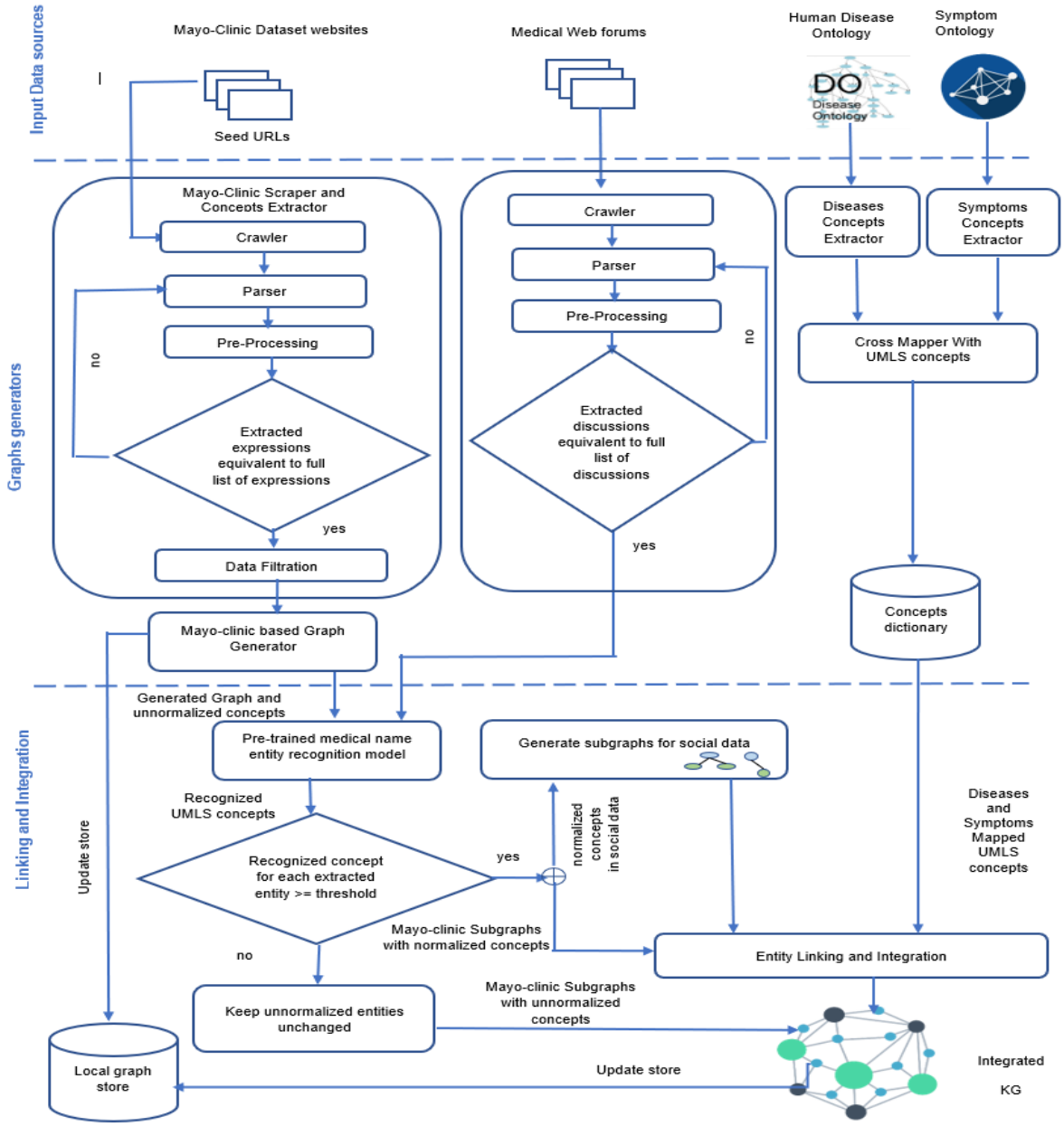


Figure. 2: Framework for generating the integrated entity linked knowledge graph.

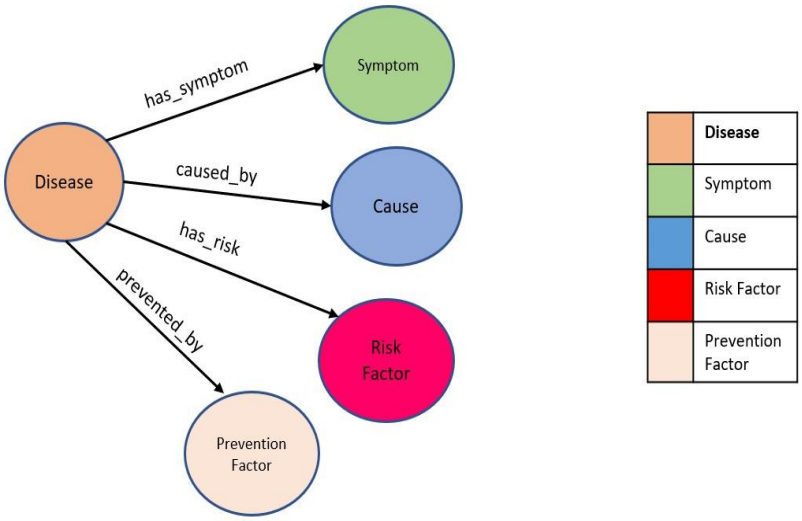


Figure. 3: The basic RDF triples showing four relationships.

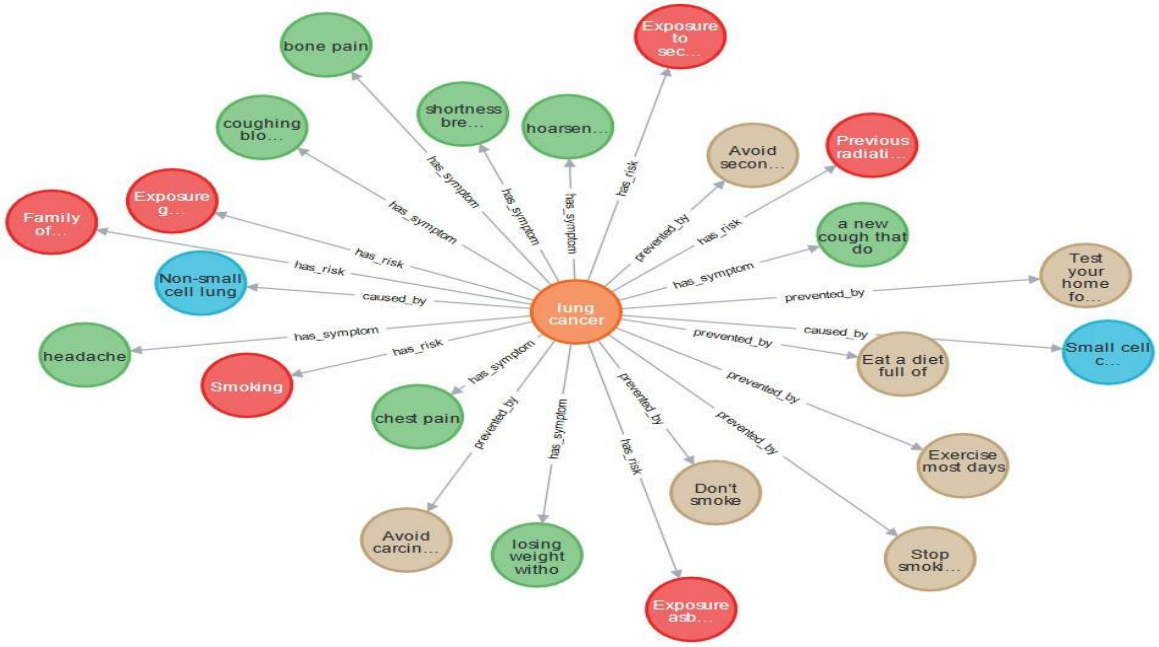


Figure. 4: A subgraph showing 'lung cancer' disease node with all its relationships with other nodes.

3.3 Generating entity linked Web-forums Graphs

The third phase focuses on scraping and extracting the conversations from medical social networks. The conversations are extracted from three resources- mdtalks, webmd cancer message boards and e-health forums. The total number of records collected from web forums are 15408 records. The records

collected are preprocessed. The next step is to apply Disease and Symptom entity recognition to identify disease and symptoms mentions within each record. There are pretrained models using deep learning techniques such as Bidirectional Encoder Representations from Transformers (BERT) [33, 34] which is **a transformer-based machine learning technique for natural language processing** and bidirectional long short-term memory (BILSTM) networks with a Conditional Random Field (CRF) layer [37].

Nowadays, these are the best models for NLP systems that provides reliable annotation and mapping of the text containing medical terms to UMLS concepts, which is a comprehensive resource of medically relevant concepts and relationships. The pretrained model adopted here is based on BERT [36], it annotates the text using the concept unique identifier (CUI) of UMLS giving a similarity score. For each record scraped, the text is taken as input to the Pre-trained NLP model and similarity score is calculated for diseases or symptoms mentioned within the record – negation is not considered. Similarity to UMLS concepts is calculated as shown in equation 2. Given two diseases x and y with their vectors d_x and d_y, n is the number of terms in a multiterm concept and i is term iterator, the cosine similarity is:

$$\cos(d_x, d_y) = \frac{\sum_{i=1}^{n} d_{x,i} d_{y,i}}{\sqrt{\sum_{i=1}^{n} d_{x,i}^2} \sqrt{\sum_{i=1}^{n} d_{y,i}^2}} \quad (1)$$

The cosine similarity ranges from 0 (no shared terms) to 1 (identical concepts). The annotated concepts based on UMLS are cross mapped with concepts of diseases and symptoms available in the dictionary generated during phase two. The only concepts considered are those for diseases and symptoms. For each record, a small graph is generated representing the disease mentioned in the record as a node and its symptoms mentioned as separate nodes connected to their disease by an edge with relationship "has_symptom". Each node has a property stating the unique identifier of the linked entity from the standardized ontologies. Each node has also an extra property score stating the similarity between the node entity and the mapped entity from the standard ontology. The entities linked are the ones with threshold of cosine similarity greater than or equal to 0.7. The relationship $R_{i,j}$ between disease i and symptom j has a weight property $w_{i,j}$ indicating the certainty of semantic relatedness to standardized ontology and the strength of association of the disease to symptom based on the social media dataset. The weight is calculated based on equation (2). The output of this phase is a group of subgraphs with nodes linked to the standardized ontologies concepts.

$$w_{i,j} = P(d_i|s_j) * \frac{score_i + score_j}{2} = \frac{P(d_i \cap s_j)}{P(s_j)} * \frac{score_i + score_j}{2}$$
$$= \frac{Total\ occurences\ of\ d_i\ with\ s_j}{Total\ mentions\ of\ the\ s_j} * \frac{score_i + score_j}{2} \quad (2)$$

3.4 Entity linking and Integration

During phase four, the entities of diseases and symptoms within the MayoClinic-based knowledge graph generated in phase one, are input to the pretrained NLP model for similarity score calculation. A threshold of 0.7 is adopted to indicate the relatedness of the node entity to the standardized ontologies. A property of unique identifier is added to nodes stating the concept unique identifier to the mapped concept of standard ontology. Each node has also an extra property score stating the similarity between the node entity and the mapped entity from the standard ontology. The relationship $R_{i,j}$ between disease i and symptom j has a weight property $w_{i,j}$ indicating the certainty of semantic relatedness to standardized ontology. The weight is calculated based on equation (3).

$$w_{i,j} = \frac{score_i + score_j}{2} \quad (3)$$

The phase of integration aims to integrate MayoClinic generated knowledge graph and the subgraphs generated from the web forums records. Graph Alignment methodology is applied, where the objective of the graph aligner is to align two graphs G from MayoClinic and H subgraph generated from a web forum record. The alignment considers a set of pairs (x,y) , where x is a node in G and y is a node in H. Graph Alignment methodology adopted works on property unique identifier of nodes from both graphs. Matching considered is an exact phrase matcher for the unique identifier property from both graphs. If the node x from graph G is similar to node y from graph H, both nodes are merged into one node z inheriting all its edges from both graphs. The merged node will have a degree equal to the sum of degrees of node x and node y as shown in the following equation:

$$deg(z) = deg(x) + deg(y)$$

The algorithm is iterative, it is repeated for each subgraph H with graph G. The integrated knowledge graph not only represents the medical facts extracted from the encyclopedia. The graph is enriched with additional symptoms based on user experiences from medical web forums. The linked graph is then pruned where for disease i and symptom j, there exists two edges with two weights, the least important edge is pruned, and the most significant edge is kept based on the weights value. The algorithm of integration as follows:

Algorithm 1 : Integration Algorithm

```
Input: Graph G, set_of_subgraphs
Output: Graph  G modified
for each subgraph_H in set_of_subgraphs:
    disease_found = search (disease_node) in G
   if disease_found is true:
      for each symptom in symptoms nodes:
         symptom_found = search (symptom_node) in G
         if symptom_found is true:
            if (Edge(disease_node, symptom_node)) exists:
               weight = maximum (wG,wH)
                weight (Edge(disease_node, symptom_node)) = weight
           else:
              new_edge = Create Edge (disease_node,symptom_node)
              weight(new_edge)  = wH
         else:
            Add symptom node to Graph G
```

```
            new_edge = Create Edge (disease_node,symptom_node)
              weight(new_edge)  = wH

else if disease_found is false:
      for each symptom in symptoms nodes:
       symptom_found = search (symptom_node) in G
      if  symptom_found is true:
          Add disease node to Graph G
            new_edge = Create Edge (disease_node,symptom_node)
              weight(new_edge)  = wH
     else:
          Add disease node to Graph G
            new_edge = Create Edge (disease_node,symptom_node)
              weight(new_edge)  = wH
```

4. Results

The integrated knowledge graph resulted in 24615 nodes with 29165 relationships. The graph has 594 nodes linked with disease concepts from standardized disease ontology and 588 nodes linked with standardized symptom ontology. The threshold adopted for entity linking is 0.7, after several techniques are applied - exact disease and symptom name phrase matcher, pretrained medical name entity recognition (NER) model with threshold 0.8 and with threshold 0.7. Table 1 shows the number of nodes interlinked after each technique applied on the MayoClinic graph. The social data from the datasets gathered has affected the weights of 23 diseases-symptoms relationships. According to the web-forums dataset used, no extra diseases' nodes are added.

Table 1: Number of linked nodes after each technique applied for MayoClinic graph

	Disease nodes	Symptom nodes
Phrase Matcher	410	108
PreTrained NER with Threshold >=0.8	580	588
PreTrained NER with Threshold>=0.7	594	588

Figure 5 shows the improvement of percentage of diseases linked with the standardized disease ontology using different methodologies for MayoClinic entities, mdtalks, e-health forums, webmd entities. Figure 6 shows the percentage of symptoms linked with the standardized symptom ontology using different methodologies for the mentioned resources.

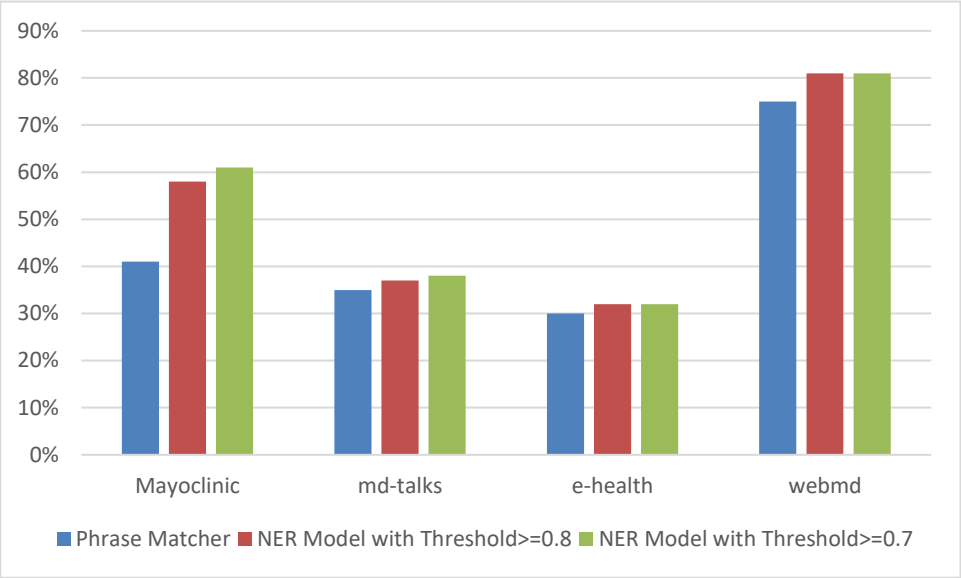


Figure. 5: Chart showing improvement of percentage of diseases linked with disease ontology using different methodologies for different resources.

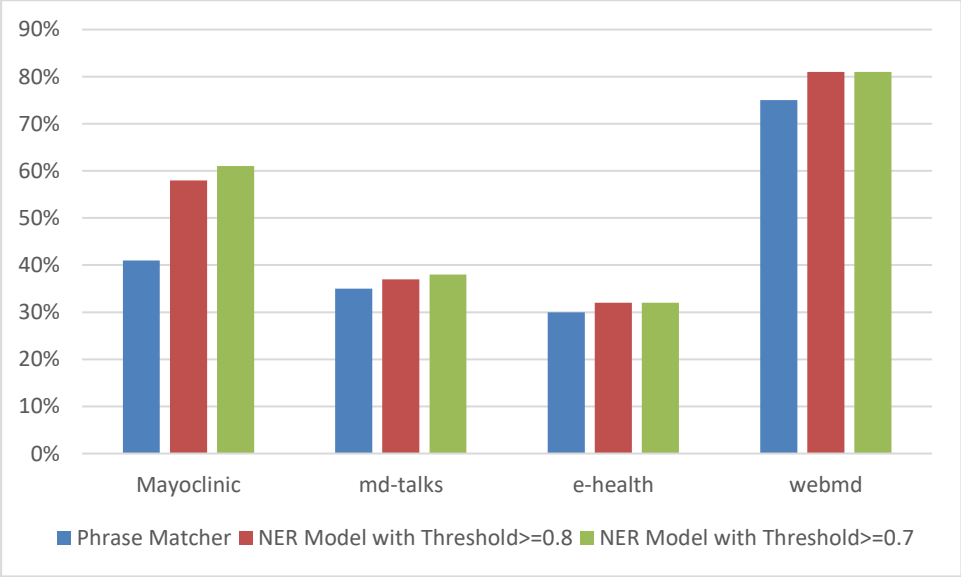


Figure. 6: Chart showing improvement of percentage of symptoms linked with symptom ontology using different methodologies for different resources.

For evaluation, an expert advisor system was built. The system was built on the linked knowledge graph considering it as its knowledge base. The test used the database of disease-symptom associations [29] where 150 frequent diseases with their symptoms from a hospital are represented and mentioned using

UMLS concepts. The database has 1865 records stating diseases and their symptoms from patients' records. The symptoms for each disease from the records were entered as an input to the system, generating cypher queries executed on the knowledge graph. The system would then infer possible diseases based on a list of symptoms selected and rank possible diseases returned as a result of the query based on the disease having the maximum count of the selected symptoms. Figure 7 shows a screenshot of the intelligent advisor expert system showing results cypher query generated. The system also directs the user to the IRI for the disease for more information and guidance.

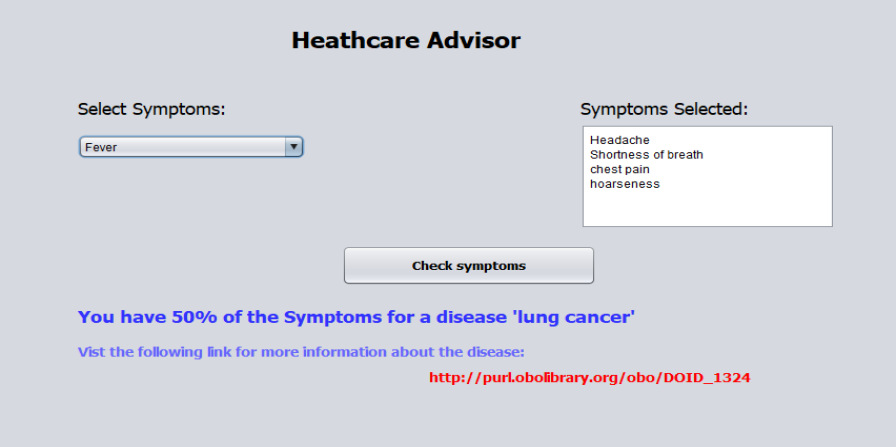


Figure. 10: a screenshot of the advisor system.

5. Conclusion and Future Work

The developed framework generated a disease-symptoms knowledge graph based on medical facts from reliable online medical encyclopedia and links the graph nodes to standardized ontologies – Human Disease Ontology and Symptom Ontology. Linking entities to these ontologies provides a standard platform for building an intelligent expert advisor healthcare system that would be used by both professional staff and normal users for disease prediction. The graph takes into consideration the impact of the symptoms experienced by social networks' users, which gives an additional level of abstraction and support to the relationships between diseases and their symptoms and could add new medical reliable facts based on user experiences. The knowledge graph also provides insights for causes, prevention factors, or risk factors of diseases.

The intelligent expert system is presented to the user through an interface internally based on semantic queries, where the user would be able to choose from the symptoms linked to the standard SYMP ontology with their unique identifiers. The system would infer possible diseases based on a list of symptoms selected and rank possible diseases based on the disease having the maximum count of the selected symptoms. Analyzing the knowledge graph would be useful for research, as measuring the density of the graph helps in extracting the most common diseases and the most shared symptom among distinct diseases. The methodologies and algorithms adopted within the proposed framework provides an automatic way to build and enrich the knowledge graph with more nodes and relationships representing different levels of abstraction whenever other datasets are taken into consideration.

Analyzing the weights between diseases and their symptoms would bring an insight of the common and highest symptom indicator for each disease. There is still work needed in the field of training models for medical entity name recognition to be able to identify accurate entity mentions, especially for the symptoms. This work needs datasets annotated by professional users, which is not available covering all diseases and symptoms nowadays. This would be a promising field of research and it will have its impact on the percentage of linked nodes to standardized ontologies.

References

[1] A. Bohr and K. Memarzadeh, The rise of artificial intelligence in healthcare applications, In Artificial Intelligence in healthcare, Academic Press, vol. January, pp. 25-60, 2020.
[2] C. Science, C. Science, and S. Medicine, "Artificial intelligence, machine learning and health systems," vol. 8, no. 2, pp. 1–8, 2018, doi: 10.7189/jogh.08.020303.
[3] A. Shaban-Nejad, M. Michalowski, and D. L. Buckeridge, "Health intelligence: how artificial intelligence transforms population and personalized health," npj Digit. Med., vol. 1, no. 1, 2018, doi: 10.1038/s41746-018-0058-9.
[4] H. Pan et al., "Biomedical ontologies and their development, management, and applications in and beyond China," J. Bio-X Res., vol. 2, no. 4, pp. 178–184, 2019, doi: 10.1097/jbr.0000000000000051.
[5] "Disease Ontology Project," Wikipedia. https://en.wikipedia.org/wiki/Disease_Ontology. [Accessed December 2022]
[6] L. M. Schriml et al., "Human Disease Ontology 2018 update: Classification, content and workflow expansion," Nucleic Acids Res., vol. 47, no. D1, pp. D955–D962, 2019, doi: 10.1093/nar/gky1032.
[7] L. M. Schriml and M. Swen, "Symptom Ontology." https://obofoundry.org/ontology/symp.html.
[8] "MayoClinic diseases and conditions." https://www.mayoclinic.org/diseases-conditions. [Accessed December 2022]
[9] "Top 15 Most Popular Health Websites." https://healthcareconsumernavigatorcenter.com/consumer-information-navigator/top-15-popular-health-websites/.[Accessed December 2022]
[10] "Health Websites Ranking." https://www.similarweb.com/top-websites/category/health/ [Accessed December 2022]
[11] A. Rossanez, J. C. dos Reis, R. da S. Torres, and H. de Ribaupierre, "KGen: a knowledge graph generator from biomedical scientific literature," BMC Med. Inform. Decis. Mak., vol. 20, no. Suppl 4, pp. 1–24, 2020, doi: 10.1186/s12911-020-01341-5.
[12] J. Tan, Q. Qiu, W. Guo, and T. Li, "Research on the construction of a knowledge graph and knowledge reasoning model in the field of urban traffic," Sustain., vol. 13, no. 6, 2021, doi: 10.3390/su13063191.
[13] D. Rizk, H. Hosny, S. ElHorbety, and A.-B. Salem, "SMART Hospital Management Systems Based on Internet of Things: Challenges, Intelligent Solutions and Functional Requirements," Int. J. Intell. Comput. Inf. Sci., vol. 0, no. 0, pp. 1–13, 2021, doi: 10.21608/ijicis.2021.82144.1107.
[14] N. E. Maghawry and S. Ghoniemy, "A proposed internet of everything framework for disease prediction," Int. J. online Biomed. Eng., vol. 15, no. 4, pp. 20–27, 2019, doi: 10.3991/ijoe.v15i04.9834.

[15] ahmed samir, T. Gharib, and S. Rady, “The Identification of the Top Positive Influential Users of the Social Networks to Help in the Control of Covid-19 Spread,” Int. J. Intell. Comput. Inf. Sci., vol. 0, no. 0, pp. 1–12, 2022, doi: 10.21608/ijicis.2022.105691.1139.
[16] H. Cho, W. Choi, and H. Lee, “A method for named entity normalization in biomedical articles: Application to diseases and plants,” BMC Bioinformatics, vol. 18, no. 1, pp. 1–12, 2017, doi: 10.1186/s12859-017-1857-8.
[17] J. X. Huang, K. S. Lee, K. S. Choi, and Y. K. Kim, “Extract reliable relations from wikipedia texts for practical ontology construction,” Comput. y Sist., vol. 20, no. 3, pp. 467–476, 2016, doi: 10.13053/CyS-20-3-2454.
[18] S. Elnagar, V. Yoon, and M. A. Thomas, “An automatic ontology generation framework with an organizational perspective,” Proc. Annu. Hawaii Int. Conf. Syst. Sci., vol. 2020-Janua, pp. 4860–4869, 2020, doi: 10.24251/hicss.2020.597.
[19] D. E. Cahyani and I. Wasito, “Automatic Ontology Construction Using Text Corpora and Ontology Design Patterns (ODPs) in Alzheimer’s Disease,” J. Ilmu Komput. dan Inf., vol. 10, no. 2, p. 59, 2017, doi: 10.21609/jiki.v10i2.374.
[20] M. Alobaidi, K. M. Malik, and S. Sabra, “Linked open data-based framework for automatic biomedical ontology generation,” BMC Bioinformatics, vol. 19, no. 1, p. 319, 2018, doi: 10.1186/s12859-018-2339-3.
[21] R. Navarro-Almanza, R. Juárez-Ramírez, G. Licea, and J. R. Castro, Automated Ontology Extraction from Unstructured Texts using Deep Learning, vol. 862. Springer International Publishing, 2020.
[22] L. Mhadhbi and J. Akaichi, “DS-ontology: A disease-symptom ontology for general diagnosis enhancement,” ACM Int. Conf. Proceeding Ser., vol. Part F1282, pp. 99–102, 2017, doi: 10.1145/3077584.3077586.
[23] T. Kim, Y. Yun, and N. Kim, “Deep learning-based knowledge graph generation for covid-19,” Sustain., vol. 13, no. 4, pp. 1–20, 2021, doi: 10.3390/su13042276.
[24] A. A. Hamed, T. E. Fandy, K. L. Tkaczuk, K. Verspoor, and B. S. Lee, “COVID-19 Drug Repurposing: A Network-Based Framework for Exploring Biomedical Literature and Clinical Trials for Possible Treatments,” Pharmaceutics, vol. 14, no. 3, 2022, doi: 10.3390/pharmaceutics14030567.
[25] A. A. Hamed, M. Rey, and M. Rey, “Mining Literature-Based Knowledge Graph for Predicting Combination Therapeutics : A COVID-19 Use Case,” no. August, 2022, doi: 10.20944/preprints202208.0305.v1.
[26] M. Alawad et al., “Integration of Domain Knowledge using Medical Knowledge Graph Deep Learning for Cancer Phenotyping,” 2021, [Online]. Available: http://arxiv.org/abs/2101.01337.
[27] G. W. Kim and D. H. Lee, “Intelligent Health Diagnosis Technique Exploiting Automatic Ontology Generation and Web-Based Personal Health Record Services,” IEEE Access, vol. 7, pp. 9419–9444, 2019, doi: 10.1109/ACCESS.2019.2891710.
[28] X. Zhou, J. Menche, A. L. Barabási, and A. Sharma, “Human symptoms-disease network,” Nat. Commun., vol. 5, no. May, 2014, doi: 10.1038/ncomms5212.

[29] “Disease-Symptom knowledge Database.” https://people.dbmi.columbia.edu/~friedma/Projects/DiseaseSymptomKB/index.html. [Accessed December 2022]
[30] C. Pechsiri and R. Piriyakul, “applied sciences Construction of Disease - Symptom Knowledge Graph from Web - Board Documents,” 2022.
[31] “UMLS Metathesaurus.” https://www.nlm.nih.gov/research/umls/knowledge_sources/metathesaurus/index.html. [Accessed December 2021]
[32] “MDTalks.” https://www.mdtalks.com/. [Accessed December 2020]
[33] “e-Health forums.” https://ehealthforum.com/. [Accessed December 2020]
[34] “WebMD messageboards.” http://messageboards.webmd.com/.[Accessed December 2020]
[35] Y. He, J. Chen, D. Antonyrajah, and I. Horrocks, “BERTMap: A BERT-based Ontology Alignment System.” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 5, pp. 5684-5691, 2022.
[36] M. Neumann, D. King, I. Beltagy, and W. Ammar, “ScispaCy: Fast and robust models for biomedical natural language processing,” BioNLP 2019 - SIGBioMed Work. Biomed. Nat. Lang. Process. Proc. 18th BioNLP Work. Shar. Task, pp. 319–327, 2019, doi: 10.18653/v1/w19-5034.
[37] K. Xu, Z. Yang, P. Kang, Q. Wang, and W. Liu, “Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition,” Comput. Biol. Med., vol. 108, no. April, pp. 122–132, 2019, doi: 10.1016/j.compbiomed.2019.04.002.