Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering pp. 285–288 ISSN 2300-5963 ACSIS, Vol. 33
DOI: 10.15439/2022R44

A New Approach of Question Answering based on Knowledge Graph in Traditional Medicine

Pham Van Duong[†], Tien-Dat Trinh[†], Hai Van Pham[†], Tran Manh Tuan[‡], Le Hoang Son[§]
Huy-The Vu[¶], Minh-Tien Nguyen[¶], Pham Minh Chuan[¶*]
[†]School of Information and Communication Technology, Hanoi University of Science and Technology, Hanoi, VietNam
[‡]Faculty of Computer Science and Enginering, Thuyloi University, Hanoi, VietNam
[§]VNU Information Technology Institute, Vietnam National University, Hanoi, Vietnam
[¶]Faculty of Information Technology, Hung Yen University of Technology and Education, Hung Yen, Vietnam
{duong.pv4w, tdtrinh11}@gmail.com; haipv@soict.hust.edu.vn; tmtuan@tlu.edu.vn;
sonlh@vnu.edu.vn; {thevh, tiennm, chuanpm}@utehy.edu.vn

***Abstract*—In recent years, it has been great interest for Question Answering (QA) systems applied to many areas placing a high value on the community. The study and development of such QA systems through chatbot tools in medicine raise great needs for clinicians in their daily activities. Chatbots use the knowledge that could be retrieved from a database, but with limited inference capability. In this paper, we propose a new QA system based on Knowledge Graph (knowledge graph) for Traditional Medicine. Data of the knowledge graph is obtained from two sources including those from diagnostic of treatment diagrams and those collected on well-known medical websites through the Internet. The knowledge graph is then formed by combining the entities and relationships using the Named Entity Recognition (NER) model. Diagnosis is made via the node similarity algorithm in the knowledge graph for symptom identification. The effectiveness of the system is demonstrated through theoretical analysis and real-world experimental outcomes.**

***Index Terms*—Knowledge Graph; Traditional Medicine; Node Similarity.**

I. Introduction

Progress in the field of Question Answering (QA) is propelling humanity to innovative technical heights, particularly in the medical field, to benefit health workers in solving medical issues. QA systems can be defined as the task whereby an automated machine (such as a computer) answers arbitrary questions formulated in natural language. Realizing the above benefits, in response to medical concerns, healthcare practitioners can use QA systems to retrieve brief phrases or paragraphs. The advantage of such systems is that they may generate responses and provide clues in seconds. These systems have partly solved difficulties such as consulting and answering patients' concerns about health and medicine. This helps reduce the pressure on the health system quite effectively, especially during the recent Covid-19 epidemic.

Due to the variety of medical textual data sources, the QA structures are just as diverse. Many researchers have done some work for it. For instance, in building a knowledge graph, [1], [2] introduces the problem of entity name recognition to extract entities and relationships in the knowledge graph automatically. [3] proposed to use deep learning focus on CBOW and BiLSTM + CRF for address the above problem. Unlike other studies, we study the construction of a medical knowledge graph in Vietnamese from many data sources by applying deep learning models to the entity extraction problem as well as building information extraction rules with data sources from the diagnostic of disease diagrams. In addition, we apply the latest technology to build our system and propose algorithms for retrieving answers from the medical knowledge graph. To summarize, our contributions are as follows:

- We construct a medical knowledge graph in Vietnamese from two data sources: data from diagnostic of disease diagrams and data collected on medical websites.
- We propose to apply a deep learning model to address the name entity recognition for data collected on medical websites. Therefore, it will reduce the construction time of the knowledge graph as well as increase the efficiency of extracting information from many different sources.
- We build the QA system by using the Rasa framework, which is one of the best frameworks for building and implementing QA systems. Besides, we also improves the search algorithm in the disease diagnosis by combining the node similarity algorithm and cypher query in neo4j to increase the accuracy.

II. Proposed Method

A. Knowledge Graph Construction

1) Data Acquisition: Data acquisition is the first step to establishing a knowledge graph. In this study, the data source of the knowledge graph is divided into two types: data from diagnostic of disease diagrams and data collected on medical websites such as vinmec.com[1], nhathuoclongchau[2], etc. As for the data source from the diagnostic of disease diagram, which is a highly reputable data source, this study will apply manual conversion of the data source from images to text.

*Corresponding Author.

[1]https://www.vinmec.com/vi/benh/
[2]https://nhathuoclongchau.com/benh

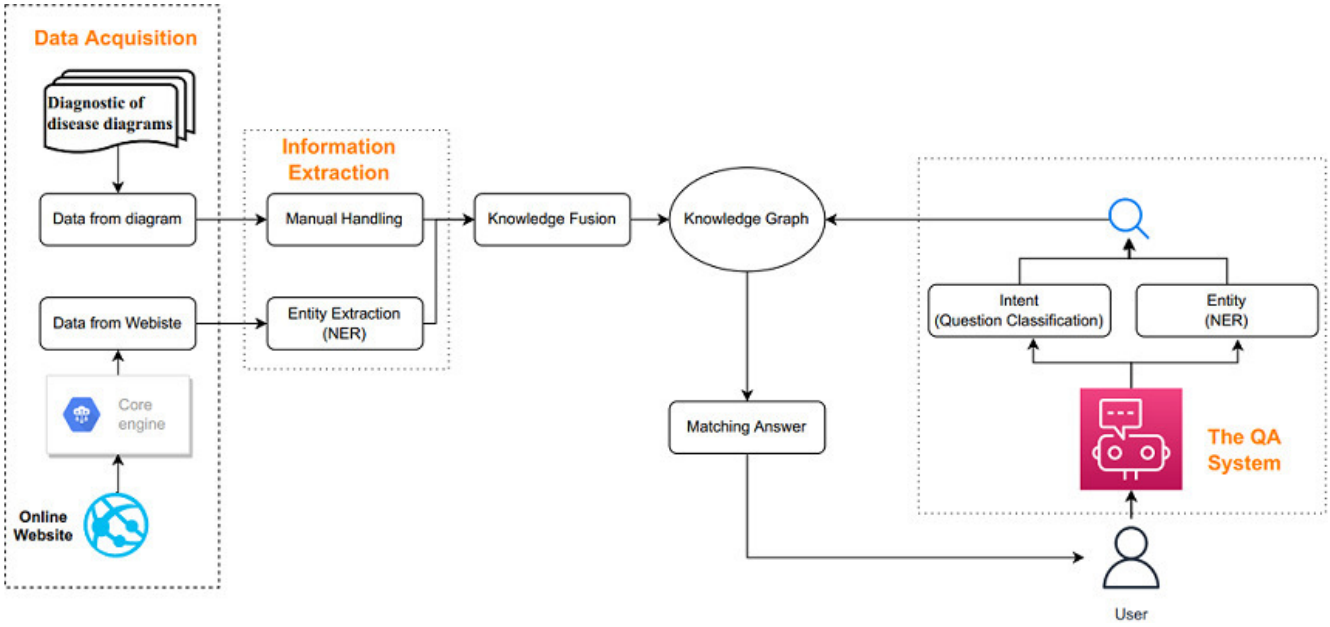

Fig. 1. Architecture of QA system based on Knowledge Graph

2) Information Extraction: Information extraction is the second step in establishing a knowledge graph. We extract structured information including Entity Extraction about symptoms and disease names while Relation Extraction and Attribute Extraction have been processed in the first step with the core engine.

In this work, we manually built a dataset based on collected data on medical websites with 2 labels (disease, symptom) containing 2,104 sentences and 26,106 labels. After collecting information about symptoms of diseases on medical websites, we preprocess the data such as removing HTML code in data, normalizing Vietnamese Unicode, removing unnecessary characters, etc. To ensure the accuracy and assurance of the construction data source, we calculate the percentage of correctly labeled sentences and correctly labeled sentences per 400 sentences by 3 independent annotators. The corresponding result is 90.5% and 96%. In addition, we calculate the Cohen's Kappa[4] measure and the result is 0.81. That proves, the dataset has quite high confidence and can serve the models well

3) Knowledge Fusion: The third step in constructing a knowledge graph is knowledge fusion. Due to the complexity of information extraction sources, the absence of hierarchy and logicality in knowledge relationships, redundancy of non-homologous knowledge, unequal knowledge expression, and other issues, this procedure requires a number of processes such as Entity Disambiguation and Entity Linking. The core application objects in this system are the atlas triple's entities, attributes, and relationships. We begin by defining the entities and their relationships, then use the similarity matching algorithm to find similar entities collectively, and then generate any attributes that exist in that entity. The entities and attributes are combined and fed into the knowledge graph through neo4j [5] database (this is a graph database used by many researchers) to store data about nodes and relationships. Nodes have 4 labels including typical_symptom, symptom, disease, and diagram corresponding to which relationships include HAS_TYPICAL_SYMPTOM, HAS_TYPICAL_SYMPTOM, HAS_SYMPTOM, YES, NO, HAS, HAS_NO. For the disease node, the knowledge graph provides additional attributes including an overview, diagnostic_measures, disease_cause, preventive_measures, treatment_measures. The constructed medical data knowledge graph includes 13,311 nodes and 16,638 relationships. The detailed description of the graph is shown in Table I and Table II. As we can see in Table I and II, the number of symptom nodes is the largest with 10,520 node and the "HAS_SYMPTOM" relationship is the largest with 15,247 relationships. Because the knowledge graph is mainly built to aid in the diagnosis of disease, the number of symptom nodes and relationships should be the majority.

TABLE I
ENTITY TYPES IN THE MEDICAL KNOWLEDGE GRAPH

Entity Type	Number of Entities
typical_symptom	62
symptom	10,520
disease	2,666
diagram	63
Total	13,311

B. QA System based on Knowledge Graph

The content of this section will present the technologies and algorithms that we use to build a QA system from the knowledge graph built from the above section.

TABLE II
RELATIONSHIP TYPES IN THE MEDICAL KNOWLEDGE GRAPH

Relationship Type	Number of Relationships
HAS_TYPICAL_SYMPTOM	63
HAS_FIRST_SYMPTOM	124
HAS_SYMPTOM	15,247
YES	122
NO	422
HAS	521
HAS_NO	139
Total	16,638

1) Rasa Framework: In this paper, we use the Rasa Framework in building Q&A scenarios by initializing user intents in domain.yml file and setting bot responses for each intent type in actions file. actions.py Rasa[6] is an open source machine learning framework for automated text and voice-based conversations. In addition, we use rasa-x to deploy the system. It is a tool for Conversation-Driven Development (CDD), the process of listening to your users and using those insights to improve your AI assistant.

2) Question Classification: For the initial requirements of building a chatbot and based on the actual types of questions that users can ask with a medical chatbot system, the article divides the types of questions from users into 6 types of questions include: asking about disease diagnosis, getting an overview of the disease, asking about the causes of the disease, asking about the treatment measures, asking about the disease diagnosis and prevention measures. Taking advantage of the rasa framework's support for user intent determination, the article customizes some changes in rasa's nlu[7] intent detector with Vietnamese characteristics. We initialize possible user intents, then for each intent we provide 10 to 20 sample questions. Because the question and answer system is specific to Vietnamese, so we customize in the config file of rasa the nlp model in the pipeline to ConveRTTokenizer and ConveRTFeaturize, this is the language model that supports Vietnamese

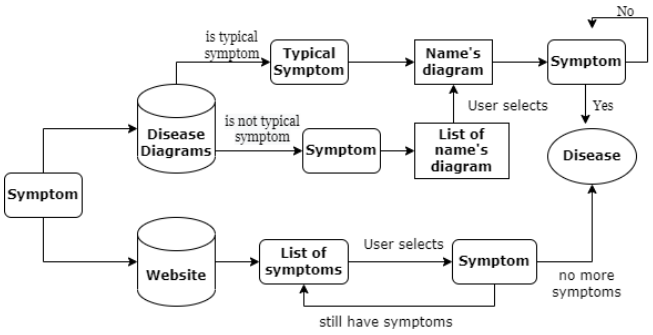

Fig. 2. Diagram about Query Processing for finding the best answer for the question of disease diagnosis

3) Query Processing for finding the best answer: Fig. 2 describes diagram about Query Processing for finding the best answer for the question of disease diagnosis. Our work uses two sources of data, because data from diagnostic of disease diagrams has a higher reputation so that when the patient provides symptoms, it will be prioritized to search symptoms in the diagram. The system will first check for that symptom in the typical symptoms of the disease diagrams, if it exists, it will use that disease diagram. Ask the user for the symptoms in turn until the name of the diagnosis is found. If that symptom is not included in any typical symptom, the system will check all disease diagrams with that symptom and give that list to the user. Then, the system will be performed as above when the disease diagram is determined. When the symptoms do not belong to the diagram in the knowledge graph, the system will switch to searching for symptoms collected from the website. In a real situation, when the patient provides symptoms, the doctor will ask and give other symptoms to determine the patient's likelihood of being sick. Based on that experiment, the study proposes to combine with the node similarity algorithm [8] to find the symptoms that frequently occur simultaneously with the symptoms provided by the patient.

4) Matching Answer: When determining the question type from the user and the entities and attributes searched from the above step, we use a mechanism to match the answers corresponding to the question to follow the originally defined template and return it to the user.

III. EXPERIMENTS AND RESULTS

In this section, we will evaluate the performance of deep learning models for the NER problem and the medical knowledge graph in Vietnamese as well as demo the QA system based on the knowledge graph.

1) Evaluation of NER Model: In order to find the best performance of the deep learning model with the NER problem, we survey the most used and most effective models in recent times to put into practice the construction of knowledge graphs. The five chosen models include ACE (Automated Concatenation of Embeddings) [9], BERT [10], PhoBERT model [11], XLM-R-large [12].

Table III describes the F1 score of five NER models. For the ACE, XLM modelsachieving the highest efficiency on the 2 data sets ShARe/CLEF, NCBI-disease. However, on the dataset of the proposed paper, the model's performance is worse than the Fine-tuning PhoBERT model. Because the dataset is in Vietnamese, being different from the above datasets in English and the Fine-tuning PhoBERT model achieves state-of-the-art for Vietnamese. From the above results, we decide to use the Fine-tuning PhoBERT model for the NER problem in building the knowledge graph and the QA system.

2) Evaluation of Knowledge Graph Experiment: To measure the trust of a knowledge graph from the data source on the website, we calculate its confidence through the link prediction in the graph. The main idea of knowledge graph evaluation is by hiding some of the relationships between entities in the graph, and then use association prediction problem-solving

TABLE III
F1 SCORE OF NER MODELS

Model	F1 Score
ACE - Embedding: XLM-R	65.7%
ACE - Embedding: Glove	70.3%
Fine-tuning PhoBERT	76.3%
Fine-tuning BERT	66.4%
Fine-tuning XLM-R-large	66.7%

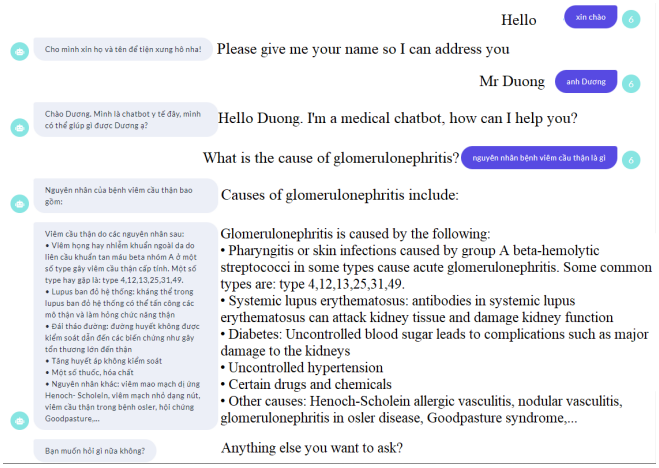

Fig. 3. Sample dialogue about disease investigation

models to determine if the relationship exists or not. Table IV describes the results of the models. The results demonstrated that the knowledge graph is excellent, with nodes in close proximity to one another.

TABLE IV
AUC_ROC SCORE OF MACHINE LEARNING MODEL FOR LINK PREDICTION IN KNOWLEDGE GRAPH

Model	AUC_ROC score
LR Model	74.46%
Random Forest Model	92.89%
LightGBM Model	98.26%

3) Demo System: In this paper, we have built a medical knowledge graph in Vietnamese from two data sources including data from diagnostic disease diagrams and data collected on medical websites. The data on the medical website is automatically collected as well as the extraction of information from this data is done by a deep learning model with the problem of name entity recognition to put into the knowledge graph. After building a knowledge graph, we built the QA system (Fig. 3) based on the rasa framework and applied it for finding the answer to each question. The Rasa framework supports the intent recognition of input questions, so the paper also uses this utility for identifying the type of questions that the user provides.

Fig 3 shows a short description of a sample dialogue about disease investigation.

IV. CONCLUSION AND FUTURE WORK

In this paper, we present the QA system based on the knowledge graph for the medical domain in Vietnamese. In this method, we create a knowledge graph that comprises medical entities, attributes of the entities, and relationships between the entities from two data sources (diagnostic of disease diagrams and an online website). The notion and entities of a given query are then recognized. An inference method is used to process the idea and entities in the inquiry. Finally, we apply the node similarity algorithm for finding symptoms that can best diagnose the disease. The outcomes of real-world experiments show that our method is effective. In the future, we will expand the dataset on more sources and consider increasing the knowledge graph's construction and retrieval speed. In addition, we will use multi-loop dialogue and complex intellectual reasoning to find the most suitable answer. This is a special feature of the system that is different from similar Q&A systems.

ACKNOWLEDGMENT

This research is funded by the Ministry of Education and Training under grant number B2022-SKH-01.

REFERENCES

[1] Z. Jiang, C. Chi, and Y. Zhan, "Research on medical question answering system based on knowledge graph," *IEEE Access*, vol. 9, pp. 21 094–21 101, 2021.
[2] D. N. Tien and H. P. Van, "Graph neural network combined knowledge graph for recommendation system," in *International Conference on Computational Data and Social Networks*. Springer, 2020, pp. 59–70.
[3] Y. Xie, "A tcm question and answer system based on medical records knowledge graph," in *2020 International Conference on Computing and Data Science (CDS)*. IEEE, 2020, pp. 373–376.
[4] M. Wirtz and M. Kutschmann, "Analyse der beurteilerübereinstimmung für kategoriale daten mittels cohens kappa und alternativer maße," *Die Rehabilitation*, vol. 46, no. 06, pp. 370–377, 2007.
[5] J. J. Miller, "Graph database applications and concepts with neo4j," in *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, vol. 2324, no. 36, 2013.
[6] R. K. Sharma and M. Joshi, "An analytical study and review of open source chatbot framework, rasa," *International Journal of Engineering Research and*, vol. 9, no. 06, 2020.
[7] A. Jiao, "An intelligent chatbot system based on entity extraction using rasa nlu and neural network," in *Journal of Physics: Conference Series*, vol. 1487, no. 1. IOP Publishing, 2020, p. 012014.
[8] W. Lu, J. C. M. Janssen, E. E. Milios, N. Japkowicz, and Y. Zhang, "Node similarity in the citation graph," *Knowledge and Information Systems*, vol. 11, pp. 105–129, 2006.
[9] X. Wang, Y. Jiang, N. Bach, T. Wang, Z. Huang, F. Huang, and K. Tu, "Automated concatenation of embeddings for structured prediction," *arXiv preprint arXiv:2010.05006*, 2020.
[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
[11] D. Q. Nguyen and A. T. Nguyen, "Phobert: Pre-trained language models for vietnamese," *arXiv preprint arXiv:2003.00744*, 2020.
[12] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," *arXiv preprint arXiv:1911.02116*, 2019.