Local context is not enough! Towards Query Semantic and Knowledge Guided Multi-Span Medical Question Answering

Abhisek Tiwari [a;*], Aman Bhansali[b], Sriparna Saha[a], Pushpak Bhattacharyya[c], Preeti Verma[a] and Minakshi Dhar[d]

[a]Indian Institute of Technology, Patna
[b]Indian Institute of Technology, Jodhpur
[c]Indian Institute of Technology, Bombay
[d]All India Institute of Medical Sciences, Rishikesh

Abstract. Medical Question Answering (MedQA) is one of the most popular and significant tasks in developing healthcare assistants. When humans extract an answer to a question from a document, they first (a) understand the question itself in detail and (b) utilize relevant knowledge/experiences to determine the answer segments. In multi-span question answering, it becomes increasingly important to comprehend the query accurately and possess relevant knowledge, as the interrelationship among different answer segments is essential for achieving completeness. Motivated by this, we first propose a transformer-based query semantic and knowledge (*QueSemKnow*) guided multi-span question-answering model. The proposed *QueSemKnow* works in a two-phased manner; in the first stage, a multi-task model is proposed to extract query semantics: (i) intent identification and (ii) question type prediction. In the second stage, *QueSemKnow* selects a relevant subset of the knowledge graph as the underlying context/document and extracts answers depending on the semantic information extracted from the first stage and context. We build a multi-task query semantic extraction model for query intent and query type identification to investigate the co-relation among these tasks. Furthermore, we created a semantically aware medical question-answering corpus named *QueSeMSpan MedQA* wherein each question is annotated with its corresponding semantic information. The proposed model outperforms several baselines and existing state-of-the-art models by a large margin on multiple datasets, which firmly demonstrates the effectiveness of the human-inspired multi-span question-answering methodology.

1 Introduction

One of the critical components of the United Nations Sustainable Development Goal (SDG 3) is the development of sustainable healthcare systems for ensuring healthy lives and well-being [31]. Over the past five years, multiple surveys [24, 19] have revealed a concerning shortage of healthcare personnel relative to the expanding population. This shortage presents a significant challenge to healthcare systems, which must increase the number of healthcare workers and optimize their time usage. Moreover, rural regions, where over one-third of the world's population resides, face a pressing issue of inadequate access to medical facilities and doctors [14]. To alleviate doctors' workload and provide timely assistance to patients, AI-based healthcare support, such as medical question-answering, has emerged as a leading-edge research area for both the AI and medical research communities [6, 12, 28].

Question answering is a prominent natural language processing (NLP) problem that has been studied for decades [26]. Despite this, the significance and research potential of question answering has not diminished, given its broad application and its role as a fundamental component in various downstream natural language understanding tasks. Most medical question-answering approaches focus on extracting a consecutive chunk of information from a document to address a query [4, 25]. Nevertheless, in real-life scenarios, answers to queries can be located in a single place or span across multiple paragraphs within a document (Figure 1). The task of multi-span question answering and its associated dataset were introduced by Zhu et al. [33]. To date, only a limited number of studies [33, 17] have been conducted in the field of multi-span question answering, leaving certain crucial aspects unexplored.

In real life, we find an answer to a question from a document in a two-phase manner: We first understand the question and its intent and then identify relevant sentences from documents needed to answer it adequately. The process becomes more critical and effective in case of multiple sentences containing answers, as with the question understated, we effectively identify the segments and necessary spans. Moreover, our medical domain knowledge plays a crucial role in answering medical queries. An experienced individual will be able to extract answers more accurately and efficiently than a 10-year-old. However, none of the existing multiple answer span question answering (MSQA) model [3, 11, 33] has either investigated the efficacy of question semantic or external knowledge in MSQA. Motivated by the research gap and efficacy of two-phased question-answering, we propose a two-phased query semantic and knowledge (*QueSemKnow*) guided transformer-based multi-span medical question-answering model. The motivation is illustrated with an example in Figure 1.

In order to effectively respond and guide health information seek-

* Corresponding Author. Email: abhisek_1921cs16@iitp.ac.in
The dataset and code are available at https://github.com/NLP-RL/QueSemKnow

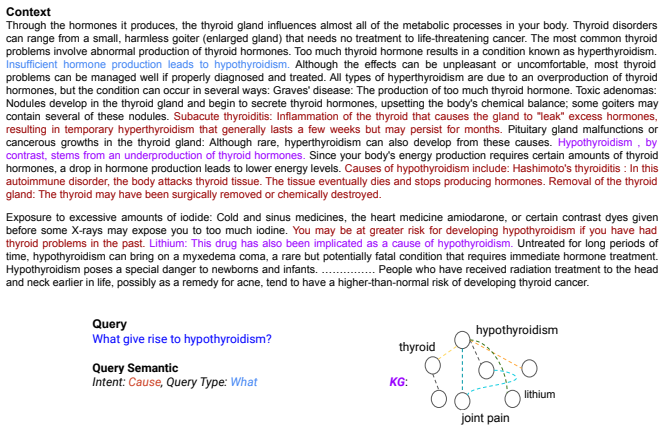

Figure 1: Importance of query semantic and external medical knowledge for multi-span question answering

ers, it is crucial to understand their intent. The intent of a query represents the underlying purpose or intention expressed by the speakers through their query. On the other hand, the query type indicates the specific type of question, such as "what" or "why". We hypothesize that there is a correlation between what is said and how it is said. The type of query can provide valuable insights into the speaker's intent, and vice versa. To address this, we have developed a multi-task model that focuses on identifying query intent and question type. This model is integrated into our proposed *QueSemKnow* framework, enhancing its capabilities for understanding and effectively addressing health information queries.

Research Questions In this paper, we aim to investigate the following three research questions related to multi-span medical question answering: **(i)** Does the incorporation of query semantics affect the efficiency of extracting multi-span answers from documents? **(ii)** Does external medical knowledge provide the background and foundation to understand and extract answers from multiple paragraphs in a document? **(iii)** Is there a correlation between the question's intent and its type recognition?

In the last few years, tremendous efforts have been made by both research and industry communities to automate various medical operations. Nevertheless, the outcomes of these efforts are limited primarily due to the lack of medical datasets [20]. There is not a single multi-span question answering data where question semantic information has been tagged. Motivated by the unavailability of query semantic associated multiple answer span QA corpus, we make an attempt to develop a large scale *Que*ry *Se*mantic information aware *M*ulti-*Span Med*ical *Q*uestion *A*nswering (*QueSeMSpan MedQA*) dataset. The corpus includes 34K question-answer pairs spread among 12 different types of queries and 11 different medical intents, such as prevention and suggestion.

The key contributions of the work are as follows:

- We propose a query semantic and knowledge-guided multi-span question-answering framework that first extracts semantic information from a question, then extracts a relevant subset from the medical knowledge graph as per the underlying context (document), and finally identifies relevant sentences accordingly.
- We propose a multi-task intent and query type identification model that exploits the interrelation between the intent and query type of a question.
- We curate large-scale semantic information annotated medical multi-span question answering corpus, which contains intent and question type for each context-question pair.
- The proposed model surpasses the existing state-of-the-art models over multiple datasets across different evaluation metrics.

Social Impact The current work makes the very first attempt towards building Query Semantic and Knowledge (*QueSemKnow*) guided Medical Multi-span question-answering model with the following objectives: (a) Individuals searching for health information online can access comprehensive medical knowledge, enabling them to effectively plan their healthcare management and make optimal use of healthcare resources in a timely manner. (b) With the proposed MedQA framework, web pages can be ranked and clustered effectively based on a structured knowledge graph that is generated from the knowledge and queries of the concerned users. (c) The developed medical corpus could be utilized for building several types of telemedicine tools such as: (i) user-intent focused medical context summarization, (ii) automatic frequently asked question (FAQ) recommendation, and (iii) healthcare consultancy and service recommendation.

2 Related Works

The work is mainly related to the following three research areas: Medical question answering, Knowledge infused question answering, and Multi-span question answering. The relevant works related to these areas are summarized in subsequent sections.

Medical Question Answering The following tasks have received the most attention in recent years: question comprehension [1], question entailment [4], and answer extraction from a concerned context [2]. Wang et al. [29] proposed a novel answer extraction method that first extracts all relevant sentences from context, and then selects a set of sentences to frame an appropriate answer. In [22], the proposed model first extracts medical entities from sentences of the EHR (electronic health record) document and then selects an appropriate question from an existing template for the extracted entity.

Knowledge Infused Question Answering In [10], the authors have infused disease-symptom knowledge in the BERT model and showed that the proposed disease-BERT significantly outperforms BERT and BioBERT [16] models. In real life, we seamlessly utilize background knowledge, such as well-known principles and facts [15], to extract relevant answers to a question from a concerned document. The same has been found in Nararatwong et al. [21], which showed a significant impact of infusing medical entity information into their question-answering model.

Multi-Span Question Answering Zhu et al. [33], first introduced the problem and developed a multiple-answer span healthcare question answering (MASH-QA) system. The paper also proposed a transformer-based classification model that classifies each context sentence as either relevant or non-relevant. The work [17] created a multi-span question-answering dataset that consists of Google queries and corresponding annotated answers from Wikipedia and other websites.

With our extensive literature survey, we did not find any work investigating neither the role of query semantics nor knowledge infusion for multi-span question answering. To the best of our knowledge, the paper is the first attempt towards building query semantic and knowledge-guided multi-span question answering framework.

3 Dataset

We first extensively reviewed existing benchmark medical question-answering corpora and the observations are summarized in Table

1. Motivated by the efficacy of the human-inspired two-phased question-answering practice, we first take an attempt to develop *Que*ry *Se*mantic information aware *M*ulti-*Span Med*ical *Q*uestion *A*nswering (*QueSeMSpan MedQA*) corpus with the help of the benchmarked MASHQA dataset [33] and clinician-provided guidelines.

Table 1: Statistics of the existing medical datasets for medical question answering task

Dataset	#QA	Context	Intent	QA Type	Multi-Span	QuesSem
HealthQA [34]	8K	✗	✗	Ranking	✗	✗
MedQuaD [4]	47K	✗	✗	Ranking	✗	✗
Medication [5]	690	✗	✓	Ranking	✗	✗
MASH-QA [33]	35K	✓	✗	Extractive	✓	✓
QueSeMSpan (ours)	34.8K	✓	✓	Extractive	✓	✓

3.1 QueSeMSpan MedQA

Motivated by the adequacy and credibility of the corpus, we decided to consider the MASH-QA as a reference corpus. It contains common healthcare queries posted on a popular health website WebMD. The answers to these queries have been marked by healthcare experts and in relevant medical documents. With the detailed analysis of queries, we found that speakers have concerns ranging from primarily 15 intents (Figure 2), such as diagnosis and prevention. We considered only eleven intents and discarded the other four intents (comparison, confirmation, medication, and case study) as we did not get enough samples in the dataset. We observed 12 different query types (Figure 3) that are used in the question-answering corpus. In the curation process, firstly, the clinician developed a sample dataset with 200 query context pairs annotated with speaker intent and question type. Afterward, we employed three biology graduate students to annotate a subset of 5000 query context pairs of the MASH-QA corpus based on the clinical author's detailed guidelines and curated samples. The kappa coefficient (k) was calculated to verify annotation agreement and it was 0.73, indicating substantial uniformity. The dataset statistics are reported in Table 2.

Table 2: *QueSeMSpan* dataset statistics

Entries	Value
# of question context pairs	34,808
# of context	5574
# of question annotated with intent and query type	5041
Avg. context length (in words)	696
Avg. answer length (in words)	67
# of intents	11
# of query types	12

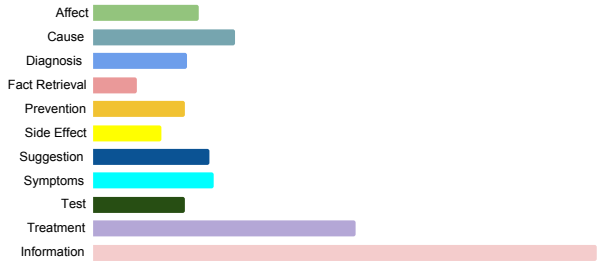

Figure 2: Intent type distribution in *QueSeMSpan MedQA* dataset

3.2 Qualitative Aspects

Question answering is a subjective task, where an individual has to understand both question and context and respond accordingly. When we provide the same question and context to a small child and an experienced individual, it is most likely that the experienced individual will perform significantly better. The reason behind this superiority is the broader background knowledge of the experienced individual. Thus, query semantics and medical knowledge become crucial in answering a medical query efficiently.

Role of Intent The sole aim of medical question-answering forums is to provide medical assistance to end-users. Therefore, it is desirable to comprehend user intent and expectations in addition to offering pertinent information. It can be effectively used to filter relevant context for a query. For example, a user's goal might be to receive a suggestion for a medical condition. We can use it to locate the relevant span from the relevant text and determine whether the span is sufficient or if additional information is required. Some common intents' details with examples are provided in Table 3.

Table 3: Common intent types, their frequency, and one example corresponding to each of them

Intent	%age	Query
information	59.30	Is there anything that I need to consider other than glycemic index when making dietary choices for diabetes?
treatment	14.07	What are some treatments for neutropenia? Will my doctor prescribe chemotherapy alone or with other treatments?
suggestion	6.57	What is the best way for people with HIV/AIDS to prevent the flu?
cause	6.29	What causes delusional disorder?
symptoms	5.89	How can one be certain of having gastritis?

Importance of Query Type The types of questions provide a key hint regarding whether a statement from a document could be a viable response. For instance, if a query begins with "when", the answer will most likely contain sentences from the concerned document, which contain a date, day, or duration. The distribution of query types are shown in Figure 3.

Figure 3: Query type distribution in *QueSeMSpan MedQA* dataset

Ethical Consideration We have strictly followed the guidelines established for legal, ethical, and regulatory standards in medical research during the *QueSeMSpan* curation process. Therefore, we have not added or removed any medical entities from a context or query of MASHQA dataset. Also, the curated dataset does not reveal users' identities. Moreover, the annotation guidelines are provided by a clinical author and thoroughly checked by the clinician. Furthermore, we have also obtained approval from our institute's healthcare committee and institutional ethical review board (ERB) to employ the dataset and carry out the research.

4 Methodology

The proposed model architecture for MSQA is illustrated in Figure 4. It extracts an answer to a question from a document in a two-step knowledge-guided process: (i) (a) *Query Semantic Extraction* (i) (b)*Knowledge Infusion* (ii) *Sentence Relevance Identification*. In the first stage, the semantic extractor extracts the speaker's intent

and query type from the user's query and retrieves additional relevant medical knowledge as per the given context. Using the processed query and additional medical knowledge (global context), it analyzes each sentence of the referenced context (local context) and marks its relevance accordingly in the later stage. Each of the stages is described in the subsequent sections.

4.1 Query Semantic Extraction

What we say in our conversation is tied to how we say it. Thus, we hypothesize that there is a strong correlation between the user's intent and query-type identification tasks. So, we jointly optimize these two tasks and build an RNN-based multi-task framework for query intent and its type recognition. We first pass the query to a BiLSTM network which generates a representation vector. The vector is fed to a feed-forward neural network ($ffnn$) that processes the vector and passes the obtained representation to two different FFNNs responsible for intent classification and query recognition, respectively. The model processes a query ($X : x_1, x_2, x_3, ...x_n$) as follows:

$$h_b^i = f(W_{hh}^b \cdot h_b^{i-1} + W_{hx}^b \cdot x_i)$$
$$h_f^i = f(W_{hh}^f \cdot h_f^{i-1} + W_{hx}^f \cdot x_i)$$
$$H = [h_b, h_f]$$
$$i, q = ffnn(H)$$

where x_i is i^{th} word of the query and f represents the activation function. Here, W, h_b^i and h_f^i denote learnable parameters (W^b: backward, and W^f: forward), the hidden representations of x_i with backward sequence and forward sequence, respectively. The terms i, and q represent constant matrix, intent, and query type, respectively. The loss function is a combination of intent loss (IL) & query-type recognition loss (QL): L= 0.7*IL + 0.3*QL.

4.2 Knowledge Distillation and Infusion

Question answering is a typical task where commonsense (in addition to domain knowledge) also plays a significant role. General commonsense concepts are crucial for understanding sentences and the underlying information in a semantic manner, thereby preventing the issue of fixed word matching. On the other hand, medical concepts play a significant role in comprehending medical information accurately. Thus, we select ConceptNet [27] for knowledge graph construction, which is one of the largest knowledge graphs (8 million nodes and 21 million edges) that contains concepts of various domains, such as news and healthcare.

Knowledge Graph Construction and Distillation While knowledge is crucial, focusing on relevant knowledge is more significant while solving a task. Thus, infusing the entire ConceptNet knowledge with the proposed question-answer setup would be ineffective and may even deteriorate the performance because a large chunk of it would be irrelevant in every query-answer extraction. We propose to distill the external knowledge based on context and inject a subset of the knowledge graph dynamically depending on the context. In the proposed framework, we first extract essential words (keywords) from a context using an unsupervised statistical-based keyword extractor called YAKE [7]. The extracted keywords are passed to the ConceptNet, which identifies relevant concepts associated with them as described in the Algorithm 1.

4.3 Multi-Span Answer Extraction

When we respond to a question, the response may arise from a variety of contexts with varied time phases. In the process of answer framing, we consider three aspects (a) query understating (b) context, and (c) global context (additional relevant knowledge). The proposed method aims to incorporate all three aspects effectively, as shown in Figure 4. Pre-trained language models have demonstrated superiority for different general language understanding tasks (GLU) in recent years, owing to massive data pre-training. XLNet [32] has been ascertained extremely effective in encoding long documents. Thus, we first pass query, query semantics, context, and knowledge triplets to the XLNet network for encoding the query and its respective context. We segregate different segments (query, query semantic, context, and knowledge triplets) with a special token ([SEP]).

Algorithm 1 Context relevant Knowledge Distillation

Input Context ($C : s_1, s_2, s_3, ...s_n$) where s_i represents i^{th} sentence of the context (C) having n sentences
Output Context relevant Knowledge Graph (KG_C)
Initialization n_k: threshold for number of keywords from a single context, n_r: threshold for number of concepts for a keyword

KG_C = []
K[$1, 2, ...n_k$] = YAKE(C, n_k) $\Rightarrow K$: list of keywords
for entity in K **do**
  KGT_{entity} =[] $\Rightarrow KGT_{entity}$: KG triplet corresponding to keyword (entity)
  for j in range(0, n_r) **do**
    <r_j, h_j, t_j> = ConceptNet(entity, KGT_{entity})
    KGT_{entity} = KGT_{entity} + [r_j, h_j, t_j]
  end for
  KG_C = KG_C + KGT_{entity}
end for
return KG_C

Self-attention Layer In the multi-span question-answer setting, context (C) usually contained multiple sentences with a varying number of words. Thus, in order to have a fixed length sentence vector, we applied self-attention over the encoded words as shown in Figure 4, which is calculated as follows.

$$h_{ij} = w_s tanh(W_s \cdot C_{ij}) \quad (1)$$
$$att_{ij} = Softmax_j(h_{ij}) \quad (2)$$
$$S_i^{'} = \sum_{j=1}^{j=k} att_{ij} \cdot C_{ij} \quad (3)$$

where w_s and W_s are learnable parameters. Here, C_{ij} denotes the encoded representation of j^{th} word of i^{th} sentence of the context. The obtained $S_i^{'}$ indicates the attended hidden representation for the i^{th} sentence of the context(C_i).

Inter-Sentence Self-attention The number of sentences selected to answer a query is far smaller than the total number of sentences. Hence, the traditional method for attention weight calculation using softmax is most likely to suffer from skewness. Thus we calculated sparsified inter-self attention α-entmax [23] over sentences (si_sa) as follows:

$$si\_sa_{ij} = w_s \cdot tanh(W_s \cdot S_{ij}^{'}) \quad (4)$$
$$\beta_{ij} = f_s(si\_sa_{ij}) \quad (5)$$
$$S_i^{''} = \sum_{j=1}^{j=k} \beta_{ij} \cdot S_{ij}^{'} \quad (6)$$

where f_s represents a sparse attention function. In contrast to softmax, it focuses on only relevant sentences (with user query) for at-

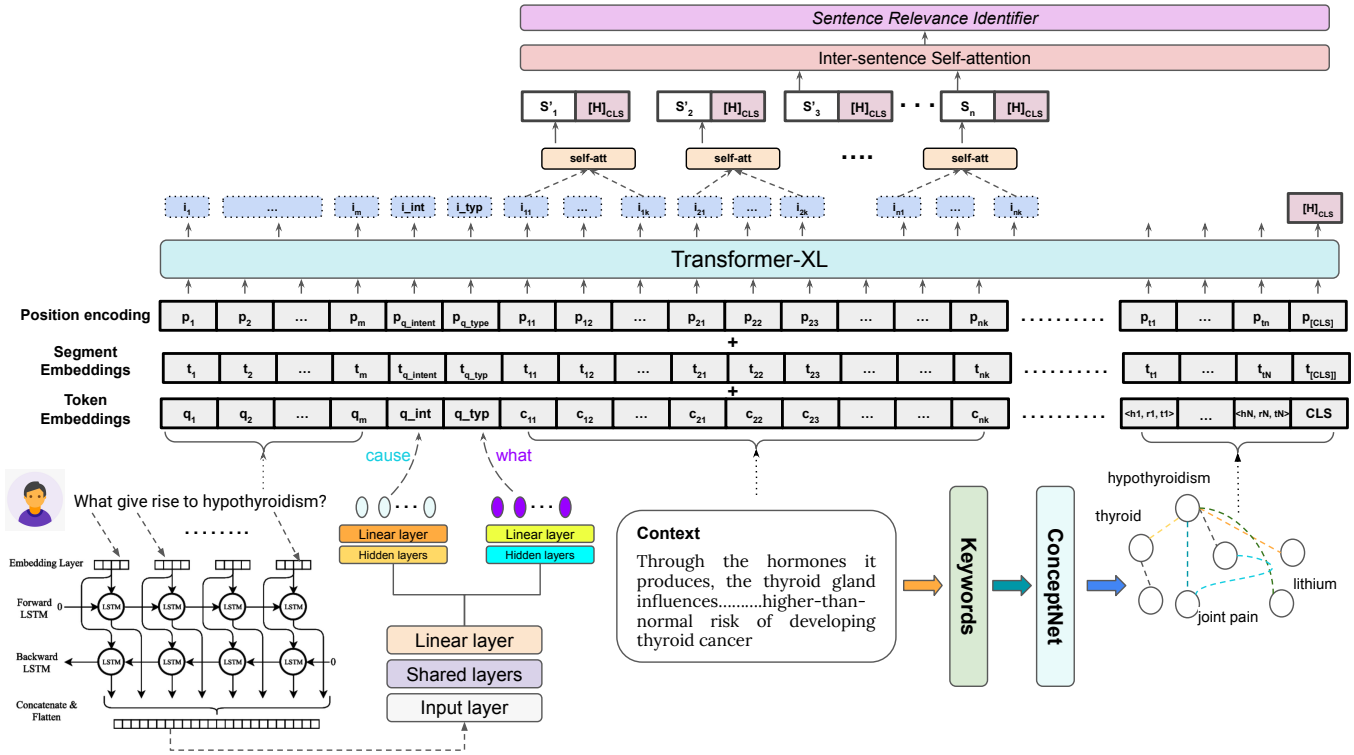

Figure 4: Architecture of the proposed query semantic and knowledge guided multi-span question-answering model. The model first extracts query semantics, distills relevant knowledge, selects a sub-graph, and finally extracts an answer from the provided context as per the query, its semantics, and the infused additional knowledge

tention weight normalization and nullifies other sentences' impact. The vectors, w_s and W_s are learnable parameters. Here, β_{ij} denotes the attention weight of i^{th} sentence with respect to the j^{th} sentence. The sparse attention function (f_s) is measured as follows:

$$f_s = \ ReLU[(\alpha - 1) \cdot a - \tau]^{\frac{1}{\alpha - 1}} \tag{7}$$

where α is constant and τ is the threshold computed using the bisection approach [23].

Sentence Relevance Identifier The obtained attended sentence vectors are passed to a feedforward neural network (FFNN), followed by softmax (Equation 8). The final layer predicts whether the sentence should be considered as an answer or not.

$$y_i = softmax(W_o \cdot S_i^{''} + b_o) \tag{8}$$

The terms W_o and b_o signify weight and bias vectors, respectively. The model employs binary cross-entropy loss to backpropagate the discrepancy between the actual data and the model's predictions. It is calculated as follows:

$$loss = -\sum_{j=1}^{j=N}\sum_{i=1}^{i=n}[y_i^{(j)} log(\hat{y}_i^{(j)}) + (1 - y_i^{(j)}) log(1 - \hat{y}_i^{(j)})] \tag{9}$$

where $y_i^{(j)}$ and $\hat{y}_i^{(j)}$ indicate predicted probability and true probability for i^{th} sentence of the j^{th} sample being considered as answer, i.e., relevance = 1. Here, N and n are no. of samples and no. of sentences in the respective sample, respectively.

5 Experimental Setup

The proposed *QueSemKnow* model was trained for 25 epochs on an RTX 2080 Ti GPU, which took around 7 hours. It has been trained, validated, and evaluated with 80%, 10%, and 10% samples of the curated *QueSeMSpan* and MASH-QA conseq [33] (answer belongs to a single span) datasets, respectively. We utilized a pre-trained version of XLNet (24 layers) as the backbone of our proposed QA framework and allowed only the top 12 layers to be trainable to leverage the pre-training. In our approach, we set the maximum length for query encoding and sentence encoding of the context as 64 and 32 tokens, respectively. Considering that XLNet can encode a maximum length of 512 tokens, we divided larger contexts into multiple segments, each consisting of 13 sentences and query encoding. We selected different hyperparameters empirically, which are as follows: batch size (4), learning rate, α (0.00002), number of keywords: 10, number of relations for each keyword: 20, and optimizer (Adam).

Baselines: In order to understand the efficacy of the proposed model, we compared the obtained performance with existing state-of-the-art models. The baselines are as follows:

- **BERT [8]** BERT is a pre-trained language model that utilizes transformer-based encoder architecture, which is trained on masked language modelling and next sentence prediction tasks.
- **RoBERTa [18]** RoBERTa is an extension of the BERT language model [8], trained on a larger set of training data with the sole objective of masked token prediction.
- **XLNet [32]** XLNet is pre-trained using the unsupervised learning task of predicting masked tokens, which uses a permutation-based language modeling objective that considers all possible permutations of the input sequence.
- **TANDA [9]** It is a BERT-based question-answering state-of-the-art model trained for (query, sentence) relevance identification.
- **MultiCo [33]**: It is the current state-of-art model for multiple answer span question-answering tasks, which is built on a pre-trained transformer-based framework for identifying the relevance

of different sentences of a context for a query.

6 Results and Discussion

We employed both sentence level and answer level evaluation metrics, namely accuracy, F1-Score, and exact match (EM) to evaluate and compare the performances of the proposed multi-span question-answering model. Accuracy and F1-Score are computed by comparing the proportion of predicted sentences that match the gold-level answers for the relevant sentences. EM determines the percentage of answers where the predicted label matches the true label for each sentence in the answer. It enables us to assess whether the model can accurately predict the entire answer. We have also measured the performances of different models in terms of context relevance prediction (CRP). CRP is the ratio of the total number of relevance matches (both relevant: 1, and non-relevant: 0) to the total number of sentences in the concerned context. Note all the reported results (Tables 6. 7, and 8) are statistically significant as validated using Welch's t-test [30] at 5% significance level. Based on the experiments, we report the following answers with evidence to our investigated research questions (RQs).

RQ 1: Is there any correlation between the intent identification of a question and its question type recognition? We first experimented with different neural networks and RNN-based models for query intent identification and query type recognition. The obtained results are reported in Table 4. The performances of the proposed multi-tasking intent-query identification model are reported in Table 5, demonstrating a significant improvement (intent: 2.9 ↑ and qtype: 0.8 ↑) over uni-task models (Table 4). The model performed superior in both metrics (Accuracy and F1-score) across both tasks. It firmly shows the importance of optimizing the tasks simultaneously, establishing a strong correlation between intent identification and query type recognition tasks.

Table 4: Performances of different models for intent identification (left) and query type recognition (right)

Model	Accuracy(%)	F1-Score
FFNN	75.4	0.746
LSTM	81.4	0.782
BiLSTM	86.2	0.848

Model	Accuracy(%)	F1-Score
FFNN	94.8	0.943
LSTM	96.3	0.889
BiLSTM	98.1	0.968

Table 5: Performance of the proposed multitasking intent-query identification model

Task	Accuracy (%)	F1-Score
Intent idn	**89.1 (2.9 ↑)**	**0.882 (0.034 ↑)**
Query type idn	**98.9 (0.8 ↑)**	**0.984 (0.016 ↑)**

RQ 2: Is there any impact of query semantics on extracting multi-span answers from a document efficiently? Table 6 shows the obtained performance of query semantic guided query understating model and its comparison with existing models. We have also experimented with one more medical question-answering corpus, MASH-QA conseq, and the obtained findings are summarized in Table 7. The findings (improvements over state-of-the-art models on both datasets) established the crucial importance of query semantics utilization in multiple answer span QA.

RQ 3: Does external medical knowledge provide the background and foundation to understand medical documents comprehensively and extract multi-span answers effectively? With the motivation of comprehensively understanding medical documents, we infused external medical knowledge using a structured knowledge graph with relevant medical entities and the corresponding relations among them. The obtained performances of the knowledge-infused model are reported in Table 6 and Table 7. As a result of additional knowledge about medical concepts (medical entities and their relationships), improvements of 2.71% and 4.38% were achieved in finding the relevance of different sentences that span multiple paragraphs. Thus, yes, the infusion of external medical knowledge played a critical role in extracting multi-span answers to a medical query.

Table 6: Performance of Query semantic guided Question Answering Models on QueSeMSpan dataset

Model	F1-Score	EM1	CRP
BERT [8]	25.21	8.89	/
RoBERTA [18]	28.65	9.40	/
XLNet [32]	29.19	9.09	/
TANDA [9]	25.44	8.95	/
MultiCo [33]	50.81	17.80	93.22
QueSem (ours)	**55.29 (4.48 ↑)**	**21.15 (3.35 ↑)**	**94.15 (0.93 ↑)**
QueKnow (ours)	**53.10 (2.71 ↑)**	**19.55 (1.75 ↑)**	**93.72 (0.50 ↑)**
***QueSemKnow* (ours)**	**55.81 (5.00 ↑)**	**21.33 (3.53 ↑)**	**94.45 (1.23 ↑)**

The Combined Impact of Query Semantic and Knowledge Infusion We also experimented with a model that infuses both queries' semantic and external medical knowledge (*QueSemKnow*), the obtained results are reported in Table 6. The model outperformed the models that leverage either of query semantic or knowledge. Thus, with the robust and in-depth evaluation of every sentence of context using query, query semantics, and external medical concept reduces the likelihood of the sentence not being considered despite its high relevance.

Table 7: Performance of Query semantic guided Question Answering Models on MASH-QA conseq dataset

Model	F1-Score	EM	CRP
BERT [8]	27.93	3.95	/
XLNet [32]	56.46	22.78	/
SpanBERT [13]	30.61	5.62	/
MultiCo [33]	59.38	26.40	95.65
QueSem (ours)	**60.03 (0.65 ↑)**	**26.72 (0.032 ↑)**	**95.68 (0.03 ↑)**
QueKnow (ours)	**63.76 (4.38 ↑)**	**27.71 (1.31 ↑)**	**96.13 (0.48 ↑)**
***QueSemKnow* (ours)**	**61.88 (2.50 ↑)**	**27.65 (1.25 ↑)**	**95.99 (0.34 ↑)**

Ablation Study In order to understand the impact and contribution of different components of the proposed model, we performed an ablation study (Table 8). It demonstrates the following: (a) Query type understating is more influential than intent for answer extraction because query type implicitly conveys user intent, (b) the incorporation of external medical knowledge with multi-span question answering substantially improves the model capability to understand the relevance of document's content for a given query.

Table 8: Impact of different components of the proposed *QueSemKnow* model

Model	F1-Score	EM	CRP
QueSemKnow w/o Intent, Query and KG	50.81	17.80	93.22
QueSemKnow with only Intent	51.89	18.56	93.25
QueSemKnow with only Query type	52.62	19.21	93.66
QueSemKnow with only KG	53.10	19.55	93.72

Human Evaluation To rule out the possibility of under-informative assessment performed by automatic metrics, we chose four top-performing models in automatic evaluation for subsequent human review. We conducted the human evaluation of 100 randomly selected test samples of the *QueSeMSpan* dataset. The evaluators were provided with predicted answers and respective queries without revealing actual labels. In the evaluation, three evaluators (two medical

Query
What medicines do doctors use to treat delusional disorder?

Context
Delusional disorder, previously called paranoid disorder, is a type of serious mental illness called a psychotic disorder. People who have it can't tell what's real from what is imagined. Delusions are the main symptom of delusional disorder….. [52] The primary medications used to attempt to treat delusional disorder are called antipsychotics. **[53]** Drugs used include: Conventional antipsychotics: Also called neuroleptics, these have been used to treat mental disorders since the mid-1950s. **[54]** They work by blocking dopamine receptors in the brain. **[55]** Dopamine is a neurotransmitter believed to be involved in the development of delusions. **[56]** Conventional antipsychotics include Chlorpromazine(Thorazine) Fluphenazine (Prolixin) Haloperidol (Haldol) Loxapine (Oxilapine) Perphenazine (Trilafon), Thioridazine (Mellaril), Thiothixene (Navane) ---. **[58]** Serotonin is another neurotransmitter believed to be involved in delusional disorder. **[59]** These drugs include: Aripiprazole (Abilify) Aripiprazole Lauroxil (Aristada) Asenapine (Saphris) Brexpiprazole (Rexulti) Cariprazine (Vraylar) Clozapine (Clozaril) Iloperidone (Fanapt) .--- disorder. **[60]** Tranquilizers might be used if the person has a very high level of anxiety or problems sleeping. [61] Antidepressants might be used to treat depression, which often happens in people with delusional disorder Psychotherapy can also be helpful, along with medications, as a way to help people better manage and cope with the stresses related to their delusional beliefs and its impact on their lives. [62] Psychotherapies that may be helpful in delusional disorder include: Individual psychotherapy can help the person recognize and correct the thinking that has become distorted. …… There's no known way to prevent delusional disorder. But early diagnosis and treatment can help lessen the disruption to the person's life, family, and friendships.

Gold [52, 61, 62]
MultCo [59]
QueSem [52,53,54,55,56,57,58,59,60,61]
QueKnow [52,53,54,55,56,57,58,59,60]
QueSemKnow [52,53,59, 60,61,62]

Figure 5: Performance of different model for a common test case. The numbers in the third braces represent sentence sequence number of the context. The proposed model adequately identifies the relevant sentences from multiple segments, along with some additional sentences, while the other models predict some contiguous segments that miss essential sentences

experts and one researcher other than the authors) were employed to assess 35, 35, and 30 question-answer pairs to assess the appropriateness of the answers provided by these models. Each question-answer sample is evaluated on a scale of 0 to 5 based on adequacy (A), fluency (F), relevance (R), completeness (C), and multi-span identifiability (MSI). The obtained scores, an average of the tests, are reported in Table 9.

Table 9: Human evaluation of different multi-span question answering models

Model	A	F	R	C	MSI	Avg.
MultiCo [33]	2.44	2.16	1.78	1.18	2.14	1.94
QueSem	2.58	2.36	1.96	1.24	2.32	2.09
QueKnow	2.74	2.34	1.65	1.39	2.48	2.12
QueSemKnow	**2.86**	**2.60**	**2.04**	**1.88**	**2.82**	**2.44**

7 Case Study and Analysis

We have analyzed the different models' performances on some common test cases to comprehend their strengths and weaknesses. One such case study is illustrated in Figure 5. The comprehensive analyses of the performances of different models lead to the following key observations: **(i)** In the multi-task framework, the NLU model's efficacy improved more for intent detection than query type identification. The behavior is primarily due to the fact that query type effectively narrows down the possible space for speaker intent, whereas, given an intent (diagnosis), a query can be framed using various types of information. **(ii)** The case study reveals that the baseline model, MultiCo, lacks a deep comprehension of the query and context, resulting in it marking fewer sentences as relevant. On the other hand, the proposed *QueSemKnow* model leverages query semantics and external knowledge to obtain a wider and relevant global context, enabling it to select all relevant sentences to provide a comprehensive answer. **(iii)** The primary cause of the proposed model's failure is that it not only identifies pertinent sentences but also includes some unrelated sentences that are connected and sequential with the selected ones, as shown in Figure 5.

Limitations Despite demonstrating effectiveness and superiority over existing state-of-the-art models, it is important to acknowledge that the proposed query semantic and knowledge guided multi-span question answering model has certain limitations. If a medical query with new concept emerges, the effectiveness of the proposed model may not be as high as it is for the represented concepts already present in the knowledge graph. However, it can still provide useful information for emerging medical problems by identifying related medical concepts and relationships that may be relevant to the problem. For instance, for COVID-19, the knowledge graph may not have specific information on the virus itself, but it may have information on related concepts such as respiratory diseases, viral infections, and immune system responses, which can be helpful in understanding patient query and concern more effectively. As a result, it would still aid in more accurately identifying pertinent segments of medical documents. To address this limitation, an effective approach would be to augment the knowledge graph, provided that the relevant knowledge is accessible.

8 Conclusion

In this work, we make an effort to advance the efficacy of an AI-assisted medical question answering framework. We proposed a two-phased query semantic and knowledge guided medical question answering model that first extracts query semantic and relevant external medical knowledge in the first phase and identifies the relevance of sentences (of a relevant document) based on query semantic and infused knowledge in the later phase. We also developed a multi-task framework for identifying query intent and type, which exploits the interrelationship between tasks to recognize query semantics more effectively. The proposed model outperforms the existing state-of-the-art model across all evaluation metrics on multiple datasets, which firmly demonstrates the efficacy of infusing query semantic and external knowledge in multiple answer span question answering. When a response contains multiple sentences, the coherence among the sentences is crucial. In the future, we would like to build an abstractive response generation model for multiple answer spans healthcare question answering.

9 Acknowledgement

Abhisek Tiwari expresses sincere appreciation for receiving the Prime Minister Research Fellowship (PMRF) Award from the Government of India, which provided support for conducting this research. Dr. Sriparna Saha extends heartfelt gratitude for the Young Faculty Research Fellowship (YFRF) Award, supported by the Visvesvaraya Ph.D. Scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, and implemented by Digital India Corporation (formerly Media Lab Asia), which has been indispensable in facilitating this research.

References

[1] Fatima Alshehri and Ghulam Muhammad, 'A comprehensive survey of the internet of things (iot) and ai-based smart healthcare', *IEEE Access*, **9**, 3660–3678, (2020).

[2] Sofia J Athenikos and Hyoil Han, 'Biomedical question answering: A survey', *Computer methods and programs in biomedicine*, **99**(1), 1–24, (2010).

[3] Seongsu Bae, Daeyoung Kim, Jiho Kim, and Edward Choi, 'Question answering for complex electronic health records database using unified encoder-decoder architecture', in *Machine Learning for Health*, pp. 13–25. PMLR, (2021).

[4] Asma Ben Abacha and Dina Demner-Fushman, 'A question-entailment approach to question answering', *BMC bioinformatics*, **20**(1), 1–23, (2019).

[5] Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan, and Dina Demner-Fushman, 'Bridging the gap between consumers' medication questions and trusted answers', in *MEDINFO 2019*, (2019).

[6] Maria M Bujnowska-Fedak, Joanna Waligóra, and Agnieszka Mastalerz-Migas, 'The internet as a source of health information and services', in *Advancements and Innovations in Health Sciences*, 1–16, Springer, (2019).

[7] Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt, 'Yake! keyword extraction from single documents using multiple local features', *Information Sciences*, **509**, 257–289, (2020).

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding', *arXiv preprint arXiv:1810.04805*, (2018).

[9] Siddhant Garg, Thuy Vu, and Alessandro Moschitti, 'Tanda: Transfer and adapt pre-trained transformer models for answer sentence selection', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7780–7788, (2020).

[10] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee, 'Infusing disease knowledge into bert for health question answering, medical inference and disease name recognition', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4604–4614, (2020).

[11] Quentin Heinrich, Gautier Viaud, and Wacim Belblidia, 'Fquad2. 0: French question answering and knowing that you know nothing', *arXiv preprint arXiv:2109.13209*, (2021).

[12] Raghav Jain, Anubhav Jangra, Sriparna Saha, and Adam Jatowt, 'A survey on medical document summarization', *arXiv preprint arXiv:2212.01669*, (2022).

[13] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy, 'Spanbert: Improving pre-training by representing and predicting spans', *Transactions of the Association for Computational Linguistics*, **8**, 64–77, (2020).

[14] MM Kamruzzaman, 'New opportunities, challenges, and applications of edge-ai for connected healthcare in smart cities', in *2021 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6. IEEE, (2021).

[15] Ugur Kursuncu, Manas Gaur, and Amit Sheth, 'Knowledge infused learning (k-il): Towards deep incorporation of knowledge in deep learning', *arXiv preprint arXiv:1912.00512*, (2019).

[16] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang, 'Biobert: a pre-trained biomedical language representation model for biomedical text mining', *Bioinformatics*, **36**(4), 1234–1240, (2020).

[17] Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin, 'Multispanqa: A dataset for multi-span question answering', in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1250–1260, (2022).

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 'Roberta: A robustly optimized bert pretraining approach', *arXiv preprint arXiv:1907.11692*, (2019).

[19] Jacek Lorkowski and Agnieszka Jugowicz, 'Shortage of physicians: a critical review', *Medical Research and Innovation*, 57–62, (2020).

[20] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley, 'Deep learning for healthcare: review, opportunities and challenges', *Briefings in bioinformatics*, **19**(6), 1236–1246, (2018).

[21] Rungsiman Nararatwong, Natthawut Kertkeidkachorn, and Ryutaro Ichise, 'Kiqa: Knowledge-infused question answering model for financial table-text data', in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 53–61, (2022).

[22] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng, 'emrqa: A large corpus for question answering on electronic medical records', *arXiv preprint arXiv:1809.00732*, (2018).

[23] Ben Peters, Vlad Niculae, and André FT Martins, 'Sparse sequence-to-sequence models', in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519, (2019).

[24] Richard M Scheffler and Daniel R Arnold, 'Projecting shortages and surpluses of doctors and nurses in the oecd: what looms ahead', *Health Economics, Policy and Law*, **14**(2), 274–290, (2019).

[25] Sheng Shen, Yaliang Li, Nan Du, Xian Wu, Yusheng Xie, Shen Ge, Tao Yang, Kai Wang, Xingzheng Liang, and Wei Fan, 'On the generation of medical question-answer pairs', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8822–8829, (2020).

[26] Marco Antonio Calijorne Soares and Fernando Silva Parreiras, 'A literature review on question answering techniques, paradigms and systems', *Journal of King Saud University-Computer and Information Sciences*, **32**(6), 635–646, (2020).

[27] Robyn Speer, Joshua Chin, and Catherine Havasi, 'Conceptnet 5.5: An open multilingual graph of general knowledge', in *Thirty-first AAAI conference on artificial intelligence*, (2017).

[28] Abhisek Tiwari, Manisimha Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, and Sarbajeet Tiwari, 'Dr. can see: towards a multi-modal disease diagnosis virtual assistant', in *Proceedings of the 31st ACM international conference on information & knowledge management*, pp. 1935–1944, (2022).

[29] Zhen Wang, Jiachen Liu, Xinyan Xiao, Yajuan Lyu, and Tian Wu, 'Joint training of candidate extraction and answer selection for reading comprehension', in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1724, (2018).

[30] Bernard L Welch, 'The generalization ofstudent's' problem when several different population variances are involved', *Biometrika*, **34**(1/2), 28–35, (1947).

[31] Gavin Yamey, Rima Shretta, and Fred Newton Binka. The 2030 sustainable development goal for health, 2014.

[32] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, 'Xlnet: Generalized autoregressive pretraining for language understanding', *Advances in neural information processing systems*, **32**, (2019).

[33] Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K Reddy, 'Question answering with long multiple-span answers', in *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3840–3849, (2020).

[34] Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy, 'A hierarchical attention retrieval model for healthcare question answering', p. 2472–2482, New York, NY, USA, (2019). Association for Computing Machinery.