applied sciences

MDPI

Review

A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches

Emmanuel Mutabazi [1], Jianjun Ni [1,2,*], Guangyi Tang [1,2] and Weidong Cao [1,2]

[1] College of IOT Engineering, Hohai University, Changzhou 213022, China; mutabemma@hhu.edu.cn (E.M.); tang_gy@hhu.edu.cn (G.T.); cwd2018@hhu.edu.cn (W.C.)
[2] College of Computer and Information, Hohai University, Nanjing 211100, China
* Correspondence: njjhhuc@gmail.com; Tel.: +86-519-85191711

Abstract: The advent of Question Answering Systems (QASs) has been envisaged as a promising solution and an efficient approach for retrieving significant information over the Internet. A considerable amount of research work has focused on open domain QASs based on deep learning techniques due to the availability of data sources. However, the medical domain receives less attention due to the shortage of medical datasets. Although Electronic Health Records (EHRs) are empowering the field of Medical Question-Answering (MQA) by providing medical information to answer user questions, the gap is still large in the medical domain, especially for textual-based sources. Therefore, in this study, the medical textual question-answering systems based on deep learning approaches were reviewed, and recent architectures of MQA systems were thoroughly explored. Furthermore, an in-depth analysis of deep learning approaches used in different MQA system tasks was provided. Finally, the different critical challenges posed by MQA systems were highlighted, and recommendations to effectively address them in forthcoming MQA systems were given out.

Keywords: medical question answering system; textual question; natural language processing; deep neural networks; machine learning

Citation: Mutabazi, E.; Ni, J.; Tang, G.; Cao, W. A Review on Medical Textual Question Answering Systems Based on Deep Learning Approaches. *Appl. Sci.* **2021**, *11*, 5456. https://doi.org/10.3390/app11125456

Academic Editor: Keun Ho Ryu

Received: 10 May 2021
Accepted: 4 June 2021
Published: 11 June 2021

1. Introduction

Progress in the field of Question Answering (QA) is leading the world to new technology heights, especially in the medical field, to assist health workers in finding solutions to medical queries. In the late 1980s, Robert Wilensky developed Unix Consultant at U.C. Berkeley [1]. This system was able to answer questions relating to the Unix operating system. In addition, the LILOG project, a text-understanding system, was developed in the field of tourism information in a German city [2]. The systems developed in the UC and LILOG helped to advance the reasoning and computational linguistics. In 1999, the QA Track of the Text Retrieval Conference (TREC) started the research into QA from the perspective of Information Retrieval (IR) [3]. In this regard, QA systems answer any question by retrieving short text extracts or phrases from lots of documents, including the answer itself. QA is viewed as a combination of information extraction and IR.

Generally, QA can be defined as a computer science discipline in the fields of IR and natural language processing (NLP), which focuses on developing a system that can automatically answer human questions in a natural language [4]. Usually, QA is a computer program, which can obtain answers by querying a structured database of information or knowledge. More commonly, QASs can extract answers from unstructured natural language document collections [5]. Some examples of the document collections used for QASs are as follows: compiled news-wire reports, a local collection of reference texts, internal organization web pages and documents, a subset of World Wide Web (WWW) pages, and a set of Wikipedia pages.

With the significant progress made by academic research, QA can deal with a wide range of question types, including fact, definition, list, hypothetical, How, Why, cross-

lingual, and semantically constrained questions. Based on the QAS' domain coverage, many researchers classify them into closed-domain and open-domain [6,7]. Closed-domain QAS use questions under a definite domain and can exploit domain-specific knowledge that is normally formalized into ontologies. In closed-domain QAS, only a few types of question are allowed, such as giving short answer only to inquiries that are related to Hyderabad city [6]. On the contrary, Open-domain QASs are rely on general ontologies and world knowledge, so they deal with nearly any kind of question. In addition, these systems usually have more data that can be used to extract answers [7]. Based on the modality of data, QAS can use text, images, or both. Multimodal QASs use multiple input modalities from users to answer questions, such as text and images [8]. The recent developments in the fields of biomedicine, electronic publishing, and computing technology have contributed to the rapid growth in the biomedical literature available online to medical practitioners [9,10]. The medical practitioners can assess biomedical literature using the PUBMED database, which consists of more than 32 million citations from life science journals, online books, etc. [11,12].

Question answering is one of the most important and difficult NLP tasks, which focuses on interactions between computers and human language. NLP deals with the problem of how computers are programmed to process and analyze large amounts of natural language data. The result is a computer that can understand the content of a document, including the contextual nuances of the language in the document [13]. The main challenges in NLP for QA include speech recognition, sentiment analysis [14], information extraction [15], text summarization [16], natural-language generation [17], etc.

Knowledge bases and the heterogeneous corpora are among the major sources that QAS use to interpret the natural language question, and return a concise and correct answer to the user. For example, Derici, et al. [18] developed a QA framework that can answer student questions in natural language, by providing answers to a multi-document summary. The system can answer factoid and open-ended questions and provide answers from foreign resources using translation techniques. However, there are still several problems in this field that need solutions, such as how to answer the given health-related questions.

There are different blogs or websites where people can ask questions and obtain answers, but it is still time-consuming to locate the information, which is similar to your own question [19]. For traditional IR systems, health workers are given different possible answers to the questions, then users go through the provided answers to find the correct one. This is time-consuming and delays the instant assistance required to solve the pressing needs of the patients. To address the challenges of accuracy in answering questions and reducing the response time for IR systems, QASs have been introduced to precisely provide timely, correct answers to health workers and their clients [20].

The research methods used in QAS can be categorized into three main categories: Semantic Parsing [21], IR [22] and Neural Networks [23]. A cue-word-based technique that categorizes semantic classes in a patient treatment scenario and analyzes their relationships is proposed in [24]. Besides this, the authors employed an automatic classification process to determine the divergence of the result. In automatic QAS, answering questions is performed in two ways: match and generation, both of which depend on the same QA corpus [25]. Cai, et al. [26] stated that closed-domain QA aiming at domain-specific information is expected to attain a reliable and effective performance in real-world applications. Likewise, Mishra and Jain [27] asserted that closed-domain QA could improve the satisfaction of users by using the specialized information required by domain experts. Specifically, Athenikos and Han [28] stated that the medical domain QA face several challenges, including complex domain-specific terminology, ontological and lexical resources.

Although there has been significant progress in QAS, very few studies have focused on the medical field for various reasons, such as lack of label information for most medical texts, diversity of questions, and models which fail to understand the question/answer text [29]. Deep learning techniques have been found to be very successful in dealing with medical tasks [30]. However, most existing medical systems are not trained on the medical

datasets, which causes poor accuracy, and sometimes errors in the retrieved answers [31]. There is a growing need to develop advanced medical QAS due to the shortage of medical practitioners and the difficulty some people in accessing hospitals [32]. Therefore, this article aims to review the recent deep learning approaches used for QAS in the medical domain. The hierarchical structure of this survey is illustrated in Figure 1.

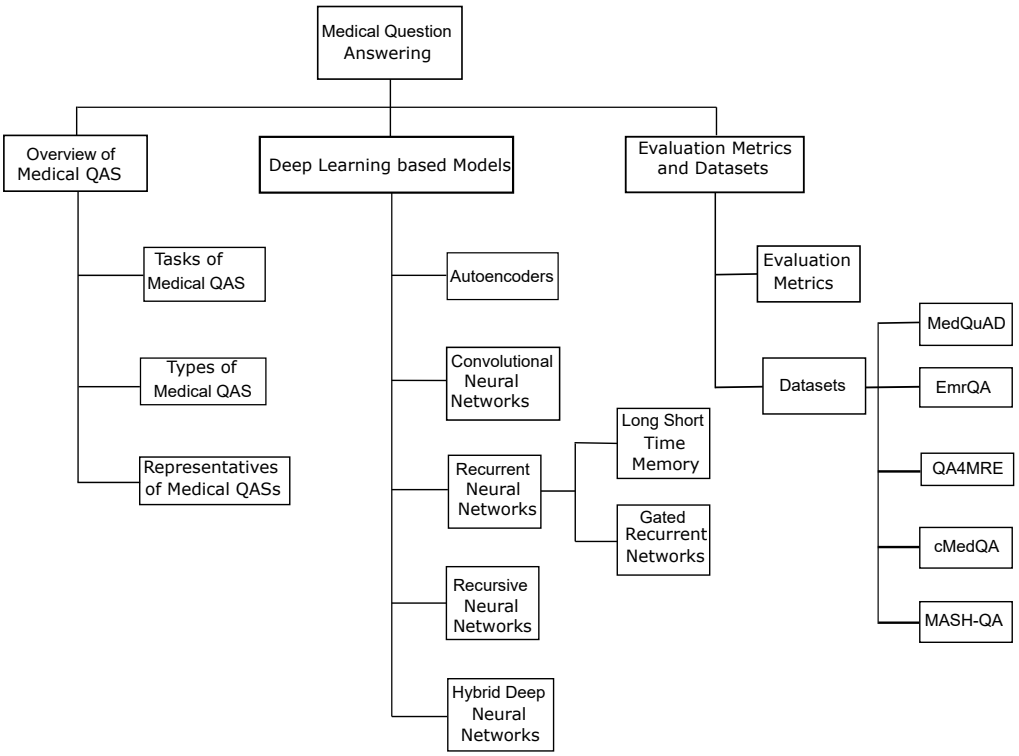

Figure 1. Hierarchical structure of this survey.

The main contributions of this paper are as follows:

(1) Provide a systematic overview of the development of MQA Systems;
(2) Categorizing deep learning techniques applied to various MQA tasks;
(3) Presenting Evaluation Metrics to measure the effectiveness of MQA Systems;
(4) Identifying existing challenges to further research in the MQA field.

Remark 1. *There are some differences among this survey and the existing related surveys. For example, Sharma, et al. [33] published a survey in 2015, which compares four biomedical systems over features such as corpus, architecture and user interface. The survey in [33] does not give out details of the methods used in these medical QAS. Kodra, et al. [34] gave out a literature review in 2017, which is focused on the general QAS, but not the medical QAS.*

This paper is organized as follows. Section 2 reviews the overview of medical QASs. Deep learning methods used in various phases of medical QASs are reviewed in Section 3. Section 4 provides different medical datasets and evaluation metrics used for measuring the effectiveness of medical QASs. The existing challenges and possible future directions in the field of medical QAS are described in Section 5. Section 6 gives the conclusion.

2. Medical Question Answering Systems

In this section, we provide an overview of the development of MQA systems. The main importance of focusing on restricted domains is that it simplifies the process of finding solutions to specific answer types. It is also possible to incorporate the domain knowledge and various patterns that can be employed to analyze the question and answer. In the next part of this section, the tasks of medical QAS will be analyzed first. Then, the

types of medical QAS will be discussed. At last, the representative medical QASs will be introduced.

2.1. Tasks of Medical QAS

Studies [9,35,36] have shown that the tasks of Medical QAS can be grouped into three modules: Question Processing, Document Processing, and Answer Processing, as summarized in Figure 2.

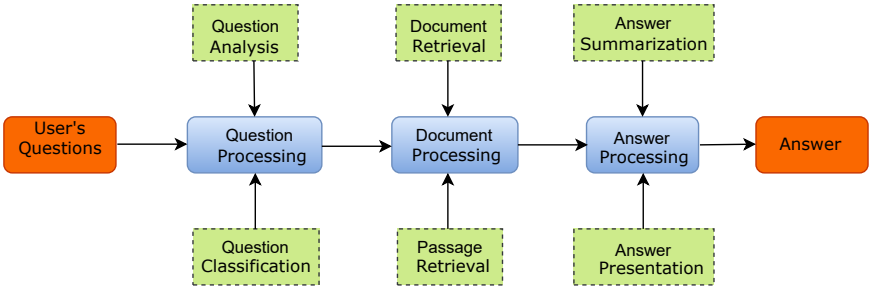

Figure 2. General architecture of QAS.

2.1.1. Question Processing

The point of focus in this module is to identify the question word. QAS accepts user input (a question in natural language) to evaluate and classify it. The evaluation is carried out to discover the type of question or what the question is, focusing on avoiding uncertainties in the answer [35]. Question-processing converts a question into a search query. Then, stop words (words that are filtered out before or after processing natural language data) and words with a particular part of speech are removed. By using deep learning technology, there is the possibility of creating a vector that can convey the exact meaning of a sentence. Classifying a question is an important step in a QAS' process.

Dina, et al. [37] highlighted that the main procedure is to translate the semantic relations expressed in questions into a machine-readable representation to deeply and efficiently analyze natural language questions. Ray, et al. [38] presented two main approaches to question classification, namely, manual and automatic methods, which is the first time that handmade rules have been used to identify the expected answer types. Although these rules can provide accurate results, they have some drawbacks such as being time-consuming, monotonic and not scalable.

Gupta, et al. [39] classified the question type as How, Where, What, Who, Why questions, etc. This type of definition facilitates better answer detection. In contrast, automatic classification can be extended to new question types with acceptable accuracy [40]. Question processing can be divided into two main procedures [41]: the structure of the user's question needs to be analyzed first, then the question needs to be transformed into a significant question formula, which is compatible with QA's domain.

Questions can also be described by the type of answer we expect to obtain. The question types include general questions with Yes/No answers, factoid questions, definition questions, list questions and complex questions [42]. General questions with Yes/No answers are the ones whose expected answer is one of two choices, one that affirms the question, and another that denies the question [43]. Factoid questions are determined by a question asking about a simple fact and receiving a concise answer in return [44]. Typically, a factoid question starts with a Wh-interrogated word, such as When, What, Where, Which, and Who [42], for example, "Which is the most common disease attributed to malfunction of cilia?" Definition questions get in return a short passage [36]: "What is diabetes?" A list question needs a set of entities that fulfils the given criteria [44]: "What are the symptoms of skin cancer?" Complex questions deal with information in a given context, whereby the answer is a combination of retrieved passages, for example: "What is the mechanism of action of a vaccine?" To implement this combination, different algorithms are used, such as Round-Robin, Normalized Raw-Scoring and Logistic Regression [45].

Questions asked by users may be in any form. It is the role of the system to deal with all types of answerable question [40]. We cannot imagine any format for the question, the question word may be placed at any place in the question, and it cannot cause any problem. The variation in question words is the main issue on which researchers need to focus.

2.1.2. Document Processing

In this module, the primary task is to select a group of related documents and extract a group of passages based on the focus of the question or text understanding through NLP [46]. The answer extraction sources are obtained by generating neural models or datasets. Thus, the retrieved data will be sorted according to their relevance to the question [36]. For each group of documents under the same domain, a document is selected and then split into multiple sentences using a sentence tokenizer and stored in an array. Each element of the array is extracted and split into words using a word tokenizer and a lemmatizer [40]. The pattern matching method can be used to find the rank for each sentence to compare the words in the question and the sentence in the document. The sentence that contains more words that are similar to the question is selected as the candidate answer. The chosen sentence is named a highly classified sentence.

2.1.3. Answer Processing

In this module, extraction techniques are applied to the results of the document processing module to show the answer [47]. This is the most difficult task on a QAS. Although the answer to the question has to be simple, it requires the merging of information from different sources, summarization, contradiction or dealing with uncertainty. When dealing with answer processing, each question word is expected to have a given label as its answer key. Question words and their corresponding expected answer labels are analyzed so that the answer keys can be found from the labelled corpus [40]. Machine learning and NLP methods implanted with probabilistic, algebraic, and neural network models have been used by different researchers to solve various answer-processing issues [34].

2.2. Types of Medical QAS

There are different classification standards for QASs. For example, the authors in [27] identified eight key criteria that researchers usually follow to classify QASs, namely, application domains, types of data, types of questions, characteristics of data sources, types of techniques used to retrieve answers, types of analyses performed on questions and source documents, and forms of answers. The authors in [34] divided the QASs into different categories based on five criteria, namely system domain, question type, system type, information source, and information source type.

In this section, we categorize the medical QAS by the paradigm each one implements. According to [48], there are four types of MQAS: Based on Knowledge Base (KB), Based on IR, Based on NLP, and Based on Hybrid paradigm:

(1) KB-based MQA: Usage of structured data source, rather than unstructured text. For example, Ontologies are formal representations of a set of concepts and their relationships in a given domain;
(2) IR-based MQA: Retrieving answers using search engines and, when the recovered passage appears, filters and ranking are applied to them;
(3) NLP-based MQA: The use of linguistic insights and machine learning procedures to extract answers from the retrieved snippet;
(4) Hybrid-based MQA: A hybrid paradigm is the combination of all three types (IR MQA, KB MQA, and NLP MQA). It uses modern search engines and enriching community contributing knowledge on the web. An example of this paradigm is IBM Watson.

2.3. Representative Medical QASs

Several authors proposed different approaches to medical QASs. For example, Sarrouti, et al. [49] developed an end-to-end biomedical QAS named SemBioNLQA, which

consists of question classification, document retrieval, passage retrieval and answer extraction modules. The SemBioNLQA system takes input questions in a natural language format, then generates short and precise answers, as well as summarizing the results. Hou, et al. [50] presented a biomedical QAS that provides answers to multiple-choice questions while reading a given document. In their study, Question Answering for Machine Reading Evaluation (QA4MRE) was used as a dataset with a focus on Alzheimer's disease. Several other medical QASs have been discussed in [33]. Some of the main MQA systems will be introduced in detail as follows, based on what was discussed in [33] and other MQA systems that they did not mention.

(1) AskHermes

AskHermes is an online MQA system designed to help healthcare providers find answers to health-related questions in order to take care of their patients. The goal of this system is to achieve a robust semantic analysis of complex clinical questions, resulting in question-focused summaries as answers [51]. Queries in natural language are allowed without the need for many representations. Furthermore, this system can answer complex questions with the help of structured domain-specific ontologies. The authors in [33] investigated the AskHermes tasks and categorized them into five modules: data sources and preprocessing, question analysis, document processing, passage retrieval, and summarization and answer presentation, as shown in Figure 3. In the data sources and preprocessing modules, the medical literature and articles are collected and indexed. The collected data are then preprocessed to retain the semantic content. In the question analysis module, questions are classified into several topics for making easy information retrieval. The binary classifier (yes, no) is used to avoid a question being assigned to multiple topics. In the document retrieval module, a designed probabilistic BM25 model has been analytically adjusted for retrieval. The final module of summarization and answer presentation is divided into two sub-sections: the first is topical clustering, ranking, and hierarchical answer presentation; the second is redundancy removal based on the longest common substring.

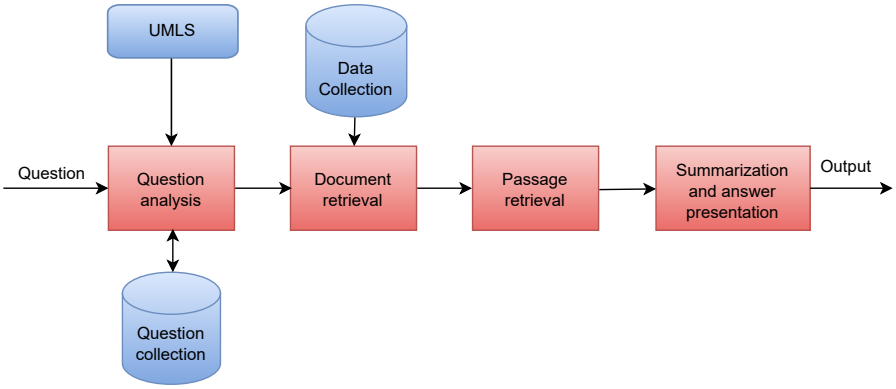

Figure 3. AskHermes's system architecture [51].

(2) MedQA

MedQA is an MQA system proposed to learn answering questions in clinical medicine using knowledge in a large-scale document collection [52]. The primary purpose of MedQA is to answer real-world questions with large-scale reading comprehension, which read individual documents and integrate information across several documents in a timely manner. MedQA's architecture has the following modules: question classification, query generalization, document retrieval, answer extraction, and text summarization. Question classification categorizes the posed question into a question type. After identifying the question type, the query generalization module evaluates the question to extract noun phrases as query terms. The document retrieval module uses query terms to retrieve documents from the locally indexed MEDLINE collection or the Web documents [53]. After retrieving the same document, which contains the answer,the answer extraction module

detects sentences that provide answers to questions. Finally, in the text summarization module, redundant sentences are removed, then the summary is presented to the user.

(3) HONQA

HONQA is a medical question-answering system designed by the Health On the Net Foundation (HON) in two different versions: English and French [33]. This system uses two corpora to answer health-related questions. These health-related questions were extracted from health experts' discussions on the internet and in the forums of health professionals. The user can choose the field of research in which he/she wants to ask a question, such as websites accredited by HON or all the websites, irrespective of accreditation. Figure 4 demonstrates a detailed process, which the HONQA system follows to answer questions asked by users.

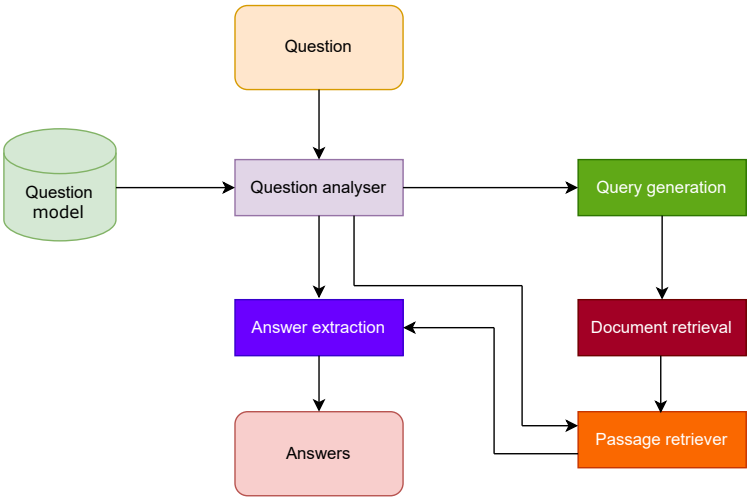

Figure 4. The architecture of HONQA [33].

(4) EAGLi

EAGLi is the biomedical question-answering and information-retrieval interface of the EAGL project [19]. This system uses MEDLINE's abstracts as an information source. EAGLi can answer definition questions with the help of Medical Subject Headings. However, EAGLi is very slow, and cannot support high-level traffic. The system answer questions by displaying a list of possible answers alongside the confidence level of answers. EAGLi interface is easy and clear, and provides the possibility of either using the PubMed search or a special search engine while the user asks a question.

(5) MiPACQ

The Multi-source Integrated Platform for Answering Clinical Questions (MiPACQ) is a QA pipeline incorporating a diversity of IR and NLP systems into an extensible QAS [52]. It uses a human-annotated evaluation dataset based on the Medpedia health and medical encyclopedia. As illustrated in Figure 5, the system can receive questions from clinicians using a web-based interface, and then semantic annotation is used to process questions. The IR system is then applied to retrieve candidate answer paragraphs; after the re-ranking system re-orders the paragraph, the results are finally presented to the user [33].

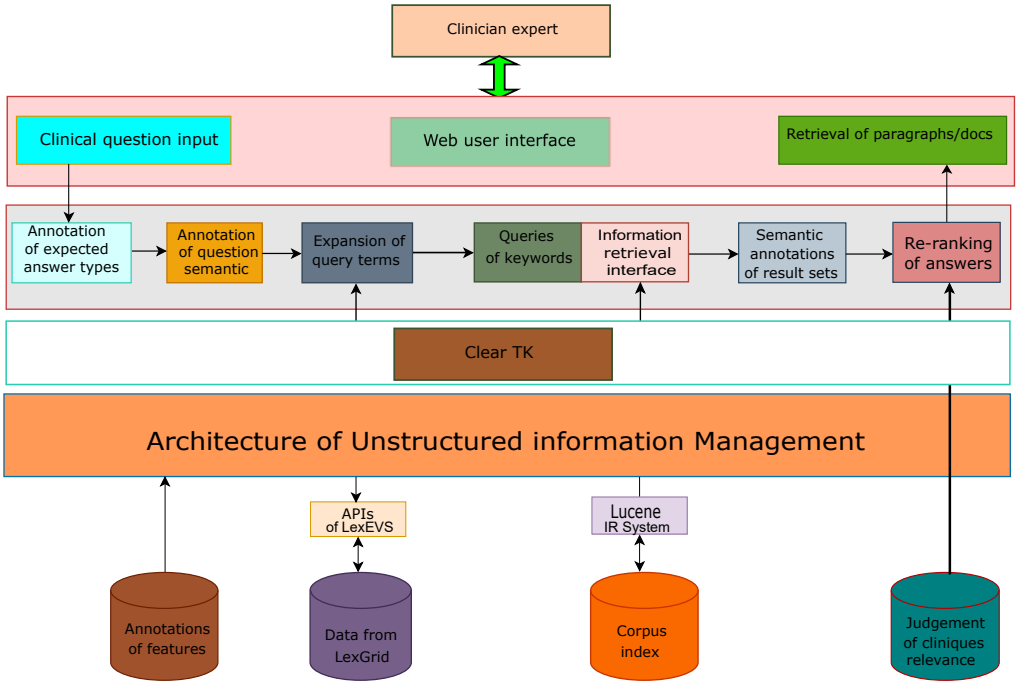

Figure 5. The architecture of MiPACQ, where ClearTK is a framework used to develop NLP and machine learning components [52].

(6) CHiQA

The Consumer Health Question Answering System (CHiQA) is an application system, which can find the health-related questions for the consumers. CHiQA used a hybrid answer retrieval strategy, which combines a free text search with a structured search based on the focus and type of information. Therefore, this system can obtain good results [37]. The CHiQA system is made of a back end and a responsive web interface. The back end comprises a preprocessing module, a question-understanding module, two complementary answer-retrieval modules, and an answer-generation module. Figure 6 exhibits a generalized architecture of CHiQA and gives more details of different QA tasks such as query formulation, query classification, answer ranking, and answer extraction.

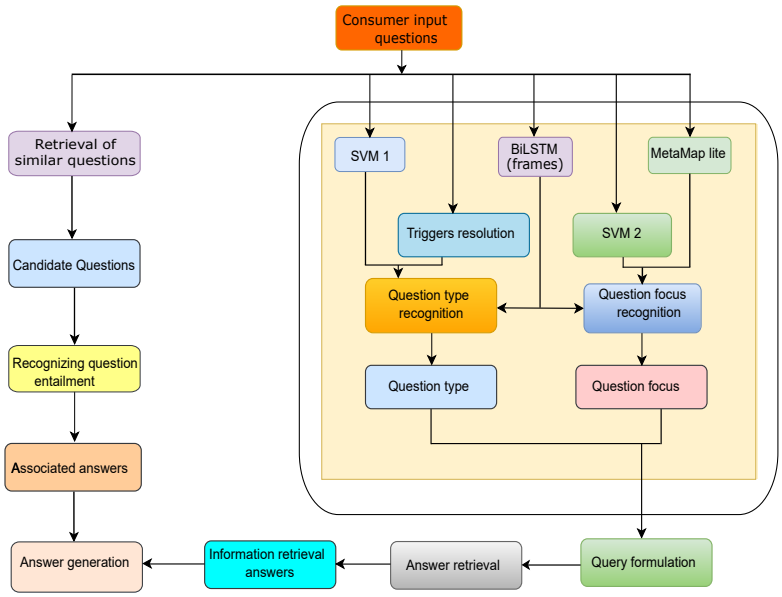

Figure 6. CHiQA's system architecture [37].

(7) CLINIQA

The CLINIcal Question Answering system (CLINIQA) is an automatic clinical question answering system, developed to answer the questions medical practitioners face in their daily work [54]. The system is made of four major sections: question classification, query formulation, answer extraction, and answer ranking (see Figure 7). CLINIQA system analyses semantically medical documents and questions with the help of the Unified Medical Language System (UMLS). Besides this, the system makes use of the machine learning algorithms to identify the question focus, classify documents, and select the answer. Once a clinical question is given to the system, CLINIQA retrieves highly reliable answers from existing medical research documents. The PUBMED abstracts are also used to extract and locate the most relevant answers.

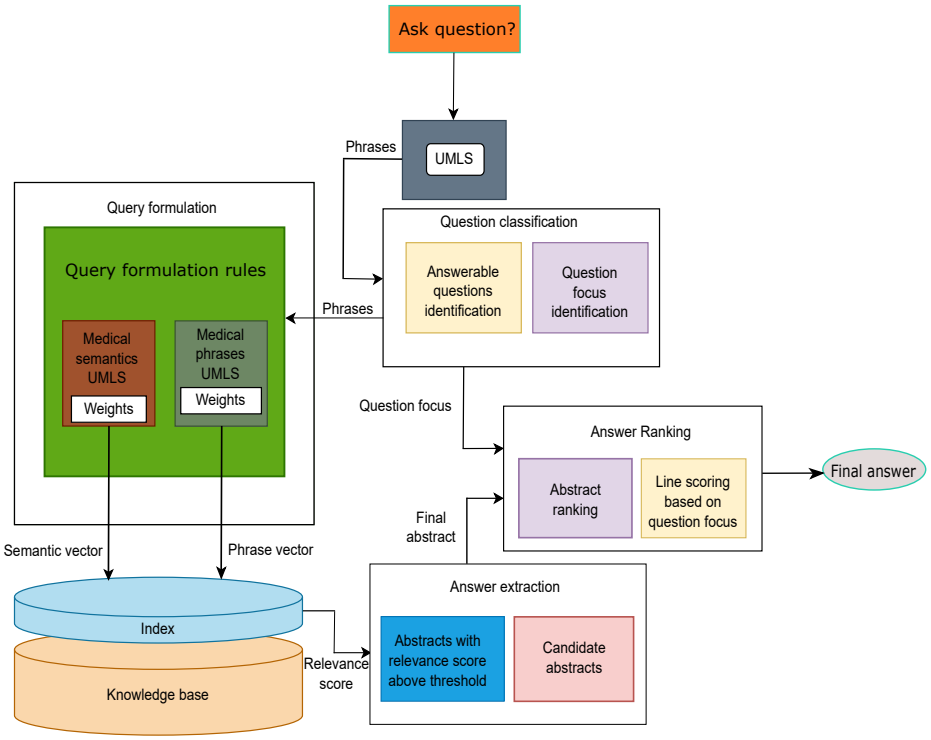

Figure 7. CLINIQA's system architecture [54].

A comparison of the representative medical QASs introduced above is listed in Table 1.

In addition, other biomedical QAS have been studied, which are all publicly accessible online. For example, Zhu, et al. [55] created a biomedical QAS based on Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT). The system has the following features: (1) the semantic network is used as a knowledge base to answer the clinical questions in natural language form, (2) a multi-layer nested structure of question templates is designed to map a template into the different semantic relationships, (3) a template description logic system is designed to define the question templates and tag template elements, (4) a textual entailment algorithm with semantics is proposed to match the question templates by considering both the accuracy and flexibility of the system.

Table 1. Comparison of different Medical QASs.

	Web Address	Corpus	Answers	Language	Interface Complexity	Key Features	System Response
AskHermes	http://www.AskHERMES.org, accessed on 15 December 2020	MEDLINE abstracts, eMedicine, clinical guidelines, PubMed Central, and Wikipedia	Multiple sentence passages	English	Simple	Answers are presented in three ways: answers clustered by terms, simple ranked answer list, and answers clustered by content	Fast
MedQA	http://www.askhermes.org/MedQA/, accessed on 15 December 2020	Web documents, (using Google: Definition), MEDLINE abstracts,	Paragraph-level answers	English	Very simple	Removes the redundant sentences and condenses the sentences into a coherent summary.	Fast
HONQA	http://services.hon.ch/cgi-bin/QA10/qa.pl, accessed on 15 December 2020	HON Certified websites	Sentence	English/French	Very simple	Use of certified health websites which allow for information to be geared towards people with varying levels of health literacy.	Slow
EAGLi	http://www.eag.unige.ch/EAGLi, accessed on 15 December 2020	Medline abstracts	Multi-phrase passages and a list of single entities	English	Complex but many tooltips	Returns a list of ranked terms to answer factual questions	Slow
MiPACQ	No Data	Online medical encyclopedia (Medpedia), and clinical questions collected by the National Library of Medicine	Ranked paragraphs	English	No Data	The annotation pipeline incorporates multiple rule-based and ML-based systems that build on each other to produce question and answer annotations.	No Data
CHiQA	http://www.chiqa.nlm.nih.gov, accessed on 15 December 2020	MedlinePlus website	Paragraph-level answers	English	Simple	Consumer health questions annotated with named entities, question topic, question triggers, and question frames.	Fast
CLINIQA	No Data	UMLS, PUBMED abstracts	The results contain the abstracts with the sentences with maximum scores highlighted	English	Simple	The architecture is robust and applicable to any disease and knowledge source	Fast

Recently, some expert QASs that use natural language have been developed. For example, Wolfram Alpha, an expert QAS that uses natural language have been developed [56]. This is an online computational knowledge engine that answers factual queries by computing the answer from outside source data. According to [37], traditional MQA approaches consist of rule-based algorithms and statistical methods with handcrafted feature sets. One of our major observations is that there is a gap in terms of performance between open domain and medical tasks, which proposes that larger medical datasets are needed to boost deep-learning-based approaches to address the linguistic complexity of consumer health questions and the issue of finding complete and accurate answers [57]. Thus, the deep-learning-based MQA will be focused and reviewed in this paper as follows.

3. Deep Learning Based MQA Approaches

Deep-learning-based models have been widely applied in various research domains, such as natural language processing [17], computer vision [58], etc. For MQA tasks, deep neural network methods perform better than traditional methods. By using deep neural network models, it is possible to process a very large amount of data in a very short time. Thus, complex medical or clinical questions can be answered with a very high accuracy comparing to traditional methods. This section gives a detailed explanation of the state-of-the-art deep neural-network-based MQA models according to the leading neural networks they adopt. We also classify papers according to the methods used, as shown in Table 2.

Table 2. classification of papers based on approaches used.

Approaches	References	Remarks
Autoencoders based	Dai, et al. (2019) [25]	An inception convolution autoencoder model is proposed to address various issues, including high dimensionality, sparseness, noise, and non-professional expression.
	Yan, et al. (2020) [59]	The proposed deep ranking recursive autoencoders architecture is used for ranking question-candidate snippet answer pairs (Q-S) in order to get the most relevant candidate answers for biomedical questions.
CNN based	Zhang, et al. (2017) [60]	A multiscale convolutional neural network framework is proposed to deal with Chinese medical question answer matching task at the character level.
	He, et al. (2019) [61]	A CNN-based semantic clustered representation (CSCR) is proposed to merge the word segments and bring forth a new representation that is compatible with deep matching models.
RNN based	Luo, et al. (2017) [62]	An LSTM model was proposed to classify relations from clinical notes. They only used word embedding without manual feature engineering, but still achieved good results.
	Asma, et al. (2019) [63]	This entailment approach was proposed to identify entailment between two questions: premise and hypothesis.
Recursive NN based	Iyyer, et al. (2014) [64]	This recursive neural network based approach proved that, paragraph-level representations results in good prediction comparing to individual sentence's level representation.
Hybrid DNN based	Cai, et al. (2017) [65]	A CNN-LSTM attention model is proposed to predict user intents, and an unsupervised clustering method is applied to mine user intent taxonomy
	Zhang, et al. (2019) [66]	A hybrid model of CNN and GRU is employed to solve the problem of complex relationship between questions and answers, in order to enhance Chinese medical question answer selection.

3.1. Autoencoders

Autoencoders are artificial neural networks employed to learn effective data encoding in unsupervised problems, which aim to learn the representation (encoding) of a set of data by training the network to ignore noises in a signal. In addition to the reduction side, a reconstruction side is also learnt, in which the autoencoder tries to generate a representation from the reduced encoding that is more close to its original input. The main process of

the autoencoder is shown in Figure 8. The final goal of the autoencoder is to combine a sequence of word vectors into a single vector, which has a fixed size and magnitude. At each step, it encodes two adjacent vectors, which satisfy a specific criterion as vectors.

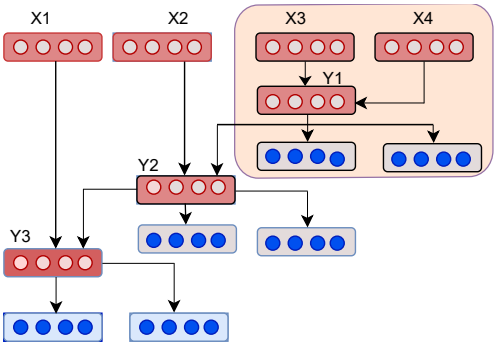

Figure 8. Deep ranking recursive autoencoders architecture, where $Y_1 = f\left(W^{(1)}[X_3; X_4] + b\right)$; $Y_2 = f\left(W^{(1)}[X_2; Y_1] + b\right)$; $Y_3 = f\left(W^{(1)}[X_1; Y_2] + b\right)$.

In the field of medical QASs, a convolutional autoencoder model for Chinese healthcare question clustering (ICAHC) was developed in [25], to solve the problem of sparsity, nonprofessional expression, high dimensionality, and noise. In this model, a set of kernels with different sizes was selected to explore both the diversity and quality in the clustering ensemble firstly, by using convolutional autoencoder networks. These kernels can capture diverse representations. Second, four ensemble operators are designed to merge representations based on whether they are independent. The output of these operators are input into the encoder. Finally, the features are mapped from the encoder into a lower-dimensional space.

Recently, Yan, et al. [59] studied the issue of matching and ranking in MQA by proposing a recursive autoencoder architecture. As the authors in [59] stated, the newly proposed method can obtain the most relevant candidate answers extracted from the upcoming relevant documents. The main concept of this method is to convert the problem of ranking candidate answers to several instantaneous binary classification problems, which will be introduced as follows.

The authors in [59] choose a binary tree, which can constantly encode more properly than the parse tree. In addition, the binary tree can encode vectors together. The tree structure can be denoted by several triplets $p \rightarrow c_1c_2$, where p is the parent node and c_1c_2 are the children.

As shown in Equation (1), with the same neural networks, the parent representations p can be computed from the children c_1, c_2 with

$$p = f\left(W^{(1)}[c_1 : c_2] + b^{(1)}\right) \tag{1}$$

where the concatenation of the two children is multiplied by a matrix of parameters $W^{(1)} \in \mathbb{R}^{n \times 2n}$. After adding a bias term b, the *tanh* is applied as an activation function.

Equation (2) shows how a reconstruction layer is usually designed to validate the combination process by reconstructing the children with

$$\left[c_1' : c_2'\right] = W^{(2)}p + b^{(2)} \tag{2}$$

Then, through comparisons between the reconstructed and the original children vectors, the reconstruction errors can be computed by their Euclidean distance, as shown in Equation (3)

$$E_{rec}([c_1; c_2]) = \frac{1}{2}\left\|[c_1; c_2] - \left[c_1'; c_2'\right]\right\|^2 \tag{3}$$

Now that the full tree is obtained by the triplets and recursive combinations, and the reconstruction error of each nonterminal node is available.

3.2. Convolutional Neural Networks Based Models

The use of convolutional neural networks (CNNs) are inspired by the processes in the visual cortex of animals [60]. Each neuron that constitutes the visual cortex is covered with a receptive field (i.e., region under the filter) of the image. These receptive fields overlap with the whole image to allow its complete visualization. Typically, CNNs have three main types of layer: an input layer, a feature extraction layer and a classification layer. CNN models are extremely effective in feature representation, including object recognition [67], sentiment analysis [14], and question answering [34].

A CNN architecture typically consists of two main processes: convolution and pooling. The convolution process is responsible for extracting features from the input content using sliding filters. On the other hand, the pooling process selects the maximum or average value of the features extracted from the former process (i.e., convolution) to reduce the feature map size.

In the QA system, the question and answer are represented by character embedding sequences, which can be denoted by $c_1, \ldots, c_{l_c}$, and $c_i \in \mathbb{R}^{d_c}$, where d_c. is the dimensionality of the character vectors. Each sentence is normalized to obtain a fixed-length sequence. After embedding, each question and answer can be represented by matrix $Q_e \in \mathbb{R}^{l_c \times d_c}$ and $A_e \in \mathbb{R}^{l_c \times d_c}$.

Given a sequence, $Z = [z_1, z_2, \ldots, z_{l_1} - s + 1]$, where s is the size of the feature map, and $z_i = [c_i, c_{i+1}, \ldots, c_{i+s-1}] \in \mathbb{R}^{s \times d_c}$ is the concatenation of continuous s character vectors of the sentence.

The convolutional operation can be defined as shown in Equation (4)

$$O_j = f\left(W_j \circ [z_1, z_2, \ldots, z_{l_c - s + 1}] + b\right) \tag{4}$$

where $O_j \in \mathbb{R}^{l_c - s + 1}$ is the output of the convolutional layer, $W_j \in \mathbb{R}^{s \times d}$ and b are the parameters to be trained, $W \circ Z$ indicates the element-wise multiplication of W with each element in Z, and $f(\cdot)$ is the activation function.

Through the convolutional layer, $Q_E \in \mathbb{R}^{l_c \times d_c}$ and $A_E \in \mathbb{R}^{l_c \times d_c}$ can be converted into $Q_O \in \mathbb{R}^{(l_c - s + 1) \times d_o}$ and $A_O \in \mathbb{R}^{(l_c - s + 1) \times d_o}$, where d_o is the number of filter maps.

Then, a pooling layer is used after the convolutional layer. The pooling layer chooses the max or average value of the features extracted from the former layer, which can reduce the representation. Equation (5) explains how the max-pooling operation is performed

$$p = [\max O_1, \max O_2, \ldots, \max O_{d_o}] \tag{5}$$

where $\max O_i$ is the max value of O_i and $p \in \mathbb{R}^{d_o}$ is the output of the pooling layer.

Equation (6) shows how to measure the similarity between the questions and the answers.

$$\mathrm{Sim}(q, a) = \mathrm{Cos}(q, a) = \frac{\|q \cdot a\|}{\|q\| \cdot \|a\|'} \tag{6}$$

where $\| \cdot \|$ is the length of the vector, $q \in \mathbb{R}^{d_o}$ and $a \in \mathbb{R}^{d_o}$ are the output of the max-pooling layer, used to represent the question and answer, respectively.

In [60], Zhang, et al. proposed the multi-scale convolutional neural network's (multi-CNNs) architecture in [60]. This end-to-end character-level architecture was employed to deal with the Chinese MQA matching task, as shown in Figure 9. The authors in [60] introduced an architecture for extracting contextual information from either question or answer sentences over various scales. Both questions and answers are all limited to the Chinese language. The reason for their choice of character embedding over word embedding is to avoid the segmentation of Chinese word in text preprocessing.

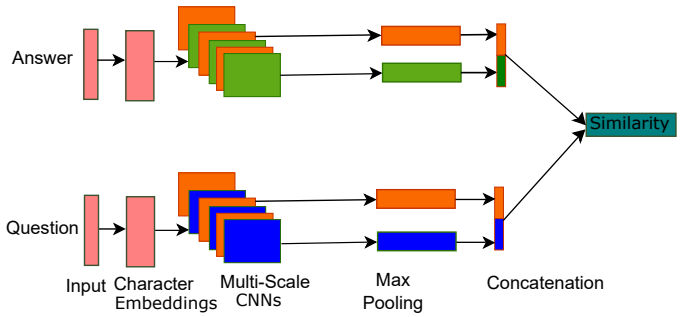

Figure 9. Architecture of multi-scale convolutional neural networks [60].

In the system based on multiCNNs architecture, convolutional operation is performed over different fixed-length regions, to extract a different number of adjacent character embeddings. A concatenation of several vectors is used to represent the question and the answer from the pooling layer. Similar to Equation (4), the output of the convolution for the i-th CNN $O_j^{s_i} \in \mathbb{R}^{l_c - s_i + 1}$ is as follows

$$O_j^{s_i} = f(w_j^{s_i} \circ [z_1, z_2, \ldots, z_{l_c - s_i + 1}] + b^{s_i}) \tag{7}$$

where s_i is the i-th CNN filter's map size.

Therefore, the output of the layer becomes $O^{s_i} = \left[O_1^{s_i}, O_2^{s_i}, \ldots, O_{d_o}^{s_i}\right]$. Similarly, after the pooling layer, the output vector from the i-th CNN is shown in Equation (8)

$$p^{s_i} = \left[\max O_1^{s_i}, \max O_2^{s_i}, \ldots, \max O_{d_o}^{s_i}\right] \tag{8}$$

Unlike single CNN, as shown in Equation (9), the output vectors from different-scale CNNs are concatenated as the final representation of the question or the answer.

$$p^{s_i} = [p^{s_1}, p^{s_2}, \ldots, p^{s_t}] \tag{9}$$

After that, the similarity measurement was calculated in a similar way to Equation (6).

3.3. Recurrent Neural-Network-Based Models

The recurrent neural network (RNN) is one of the underlying network architectures for building other deep learning architectures. The main difference between an RNN and a typical multi-layer network is that an RNN may not have fully feed-forward connections; instead, connections are fed back to previous layers (or to the same layer). This feedback allows RNNs to store previous inputs and model statuses on time.

RNNs are sequential architectures, good at modeling units in sequence. Typical RNN can be interpreted as a standard neural network for sequence data $\left(x_1, \ldots, x_p\right)$ that updates the hidden state vector h_s as shown in Equation (10)

$$h_s = \text{sigmoid}\left(W x_s \bigcup h_{s-1}\right) \tag{10}$$

An entailment approach to identify entailment between two questions, premise (PQt) and hypothesis (HQt), was proposed in [63]. As presented in Figure 10, the model treats the stacked sentence representations as inputs and the last layer as a Softmax classifier. The sentence embedding model combines the words in RNN embeddings. The word embeddings are first initialized using pre-trained GloVe vectors, to generate vector representations of words [68]. In previous experiments using RQE data, this tuning provided the best performance.

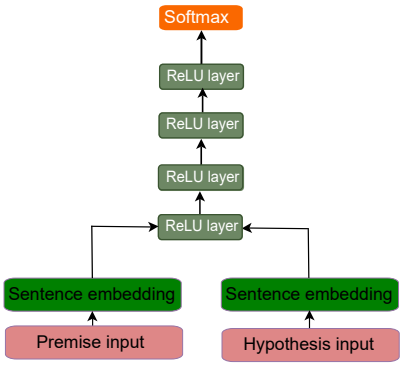

Figure 10. RNN architecture for sentence embedding [63].

3.3.1. Long Short Time Memory (LSTM)

LSTM is an artificial RNN architecture used in the field of deep learning [12]. Unlike normal feedforward neural networks, LSTM has feedback connections that enable it not only to process single datapoints (such as images) but also entire sequences of data (such as speech or video).

An LSTM model is proposed in [62] to classify relations from clinical notes. The model was tested on the i2b2/VA relation classification challenge dataset, and the results showed that, with only word embedding feature and no manual feature engineering, they achieved a micro-averaged f-measure of 0.661 to classify medical problem-treatment relations, 0.683 for medical problem–medical problem relations, and 0.800 for medical problem–test relations.

Moreover, a multiple positional sentence representation with Long–Short-Term Memory (MV-LSTM) and MatchPyramid were proposed in [61], to generate two hidden states to reflect the meaning of the whole sentence from the two directions for each word, as demonstrated in Figure 11, where MV-LSTM is a basic matching model that has a steady performance, and MatchPyramid is a model usually used for text matching.

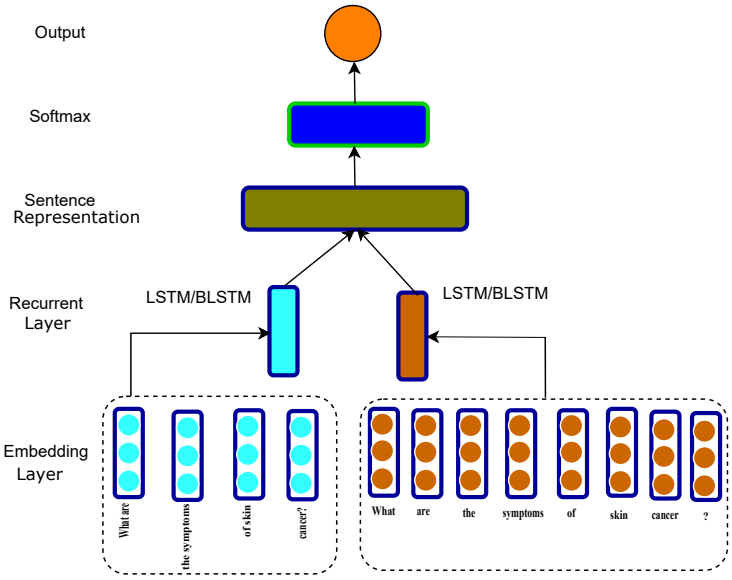

Figure 11. Architecture of Long Short Time Memory.

A SeaReader model that uses LSTM networks to model the context representation of text was proposed in [29]. This reading comprehension model uses the Attention to model information flow between questions and documents, and across several documents. Information from various documents is merged to make the last prediction. The model in [29] will be introduced in detail as follows:

The matching matrix is computed as the dot-product of the context embeddings of the question and every document, as shown in Equation (11)

$$M_n(i,j) = S(i) \odot D_n(j) \tag{11}$$

where S denotes the statement and D denotes the document.

Then, in the path of question-centric, the column-wise attention is performed on the matching matrix as shown in Equation (12)

$$\alpha_n(i,j) = \text{softmax}(M_n(i,1), \ldots, M_n(i,L_D))(j) \tag{12}$$

Each word $S(i)$ in question-answer gets a summarization read $R_n^Q(i)$ of related information in the document

$$R_n^Q(i) = \sum_{j=1}^{L_D} \alpha_n(i,j) D_n(j) \tag{13}$$

Moreover, in the document-centric path, the row-wise attention is performed to read related information in the question. Finally, the cross-document attention is performed on attention reads of all the documents.

3.3.2. Gated Recurrent Unit (GRU)

GRU, introduced in 2014, is another variant of RNN, similar to LSTM [69]. GRU aims to solve the problem of vanishing gradient that comes with a classical RNN. GRU uses two gates (update gate and reset gate) to solve the vanishing gradient problem of a classical RNN. The reset gate governs the combination of new input and previous computations. This gate is used to decide how much of the past information to forget. The update gate determines what information is retained from past computations. GRU is defined as a simplified LSTM model, which is computationally more efficient compared to both LSTM and normal RNNs.

He, et al. [70] used a bidirectional GRU model to encode the n-gram feature representations, which contains a forward GRU and a backward GRU. Equations (14)–(17) give more details about the model in [70].

$$\mathbf{r}_j = \sigma\left(W_r \cdot C_j + U_r \cdot \mathbf{h}_{j-1} + \mathbf{b}_r\right) \tag{14}$$

$$\mathbf{z}_j = \sigma\left(W_z \cdot C_j + U_z \cdot \mathbf{h}_{j-1} + \mathbf{b}_z\right) \tag{15}$$

$$\tilde{\mathbf{h}}_j = \tanh\left(W_h \cdot C_j + \mathbf{r}_j \odot \left(U_h \cdot \mathbf{h}_{j-1}\right) + \mathbf{b}_h\right) \tag{16}$$

$$\mathbf{h}_j = \left(1 - \mathbf{z}_j\right) \odot \mathbf{h}_{j-1} + \mathbf{z}_j \odot \tilde{\mathbf{h}}_j \tag{17}$$

where σ is the sigmoid function, $\odot$ stands for the element-wise multiplication, C_j is the current n-gram feature representation, $\mathbf{h}_{j-1}$ and $\mathbf{h}_j$ are the previous and the candidate hidden state, respectively, and $\mathbf{h}_j \in \mathbb{R}^{d^h}$ is the current hidden state.

The final j-th hidden state can be obtained by concatenating the j-th forward and backward hidden state: $\mathbf{h}_j = \left[\overrightarrow{\mathbf{h}}_j^{\mathrm{T}}, \overleftarrow{\mathbf{h}}_j^{\mathrm{T}}\right]$, which contains the dependencies of the preceding and following n-gram features.

In addition, Stroh, et al. [71] discussed the application of deep learning models to the QA task. In their project, RNN-based baselines were described with more attention to the state-of-the-art, end-to-end memory networks which are fast to train and provide better results on various QA tasks. As shown in Figure 12, the GRU model is trained using Keras. The model proposed in [71] generates separate representations for the query and each sentence of the passage using a GRU cell. They combine the representation of the

query with the representation of each sentence by adding the two vectors. Afterwards, the combined vector is projected to a dense layer $D \in \mathbb{R}^V$. Finally, the output of the model is obtained by taking a Softball over the dense layer D.

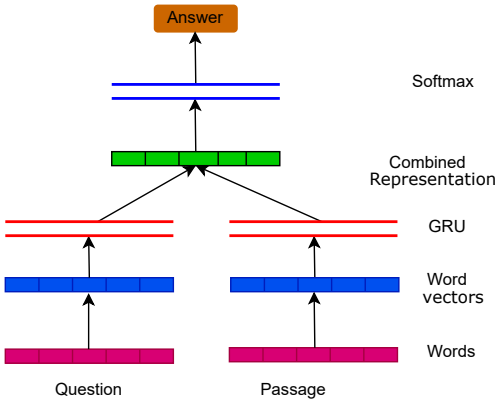

Figure 12. Architecture of gated recurrent unit.

3.4. Recursive Neural-Network-Based Models

A recursive neural network is a DNN category invented by recursively applying the same set of weights to structured inputs, generating structured or scalar predictions by crossing a given structure over variable-sized input structures in a topological order. Recursive neural networks perform the same as RNNs in handling variable length inputs. The primary difference is that RNN can model the hierarchical structures in the training dataset.

A recursive neural network architecture consists of a shared weight matrix and a binary tree structure, which allows the recursive network to learn sequences of word variations (see Figure 13). This network uses a variant of back propagation called back propagation through structure (BPTS). However, a recurrent neural tensor network (RNTN) computes a supervised target at each tree's node. The tensor (a matrix of more than two dimensions) part means that it computes gradients in a slightly different way, taking more information at each node into account by using the tensor to exploit information from another dimension.

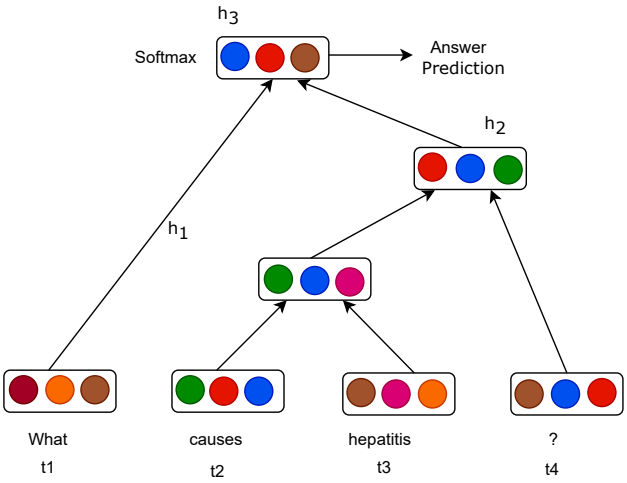

Figure 13. Architecture of recursive neural networks, where $h_1 = f\left(w\begin{bmatrix} t_2 \\ t_3 \end{bmatrix}\right)$; $h_2 = f\left(w\begin{bmatrix} h_1 \\ t_4 \end{bmatrix}\right)$; $h_3 = f\left(w\begin{bmatrix} t_1 \\ h_2 \end{bmatrix}\right)$.

Iyyer, et al. [64] proposed a dependency-tree recursive neural network (DT-RNN) model that can compute distributed representations for the individual sentences within quiz bowl questions. They extended their method to join predictions across sentences to create a question-answering neural network with trans-sentential averaging (QAN-

TAS). They claimed that, once sentence-level representations are combined, they return paragraph-level representations which give a good prediction compared to individual sentences. Their model is described as follows.

They start by relating each word w in their vocabulary with a vector representation $x_w \in \mathbb{R}^d$. Then, they store vectors as the columns of a $d \times V$ dimensional word embedding matrix W_e, where V is the size of the vocabulary. Their model considers dependency parse trees of question sentences.

Each node n in the parse tree for a particular sentence is associated with a word w, a word vector x_w and a hidden vector $h_n \in \mathbb{R}^d$ of the same dimension as the word vectors. For internal nodes, this vector is a phrase-level representation, while at leaf nodes it is the word vector x_w mapped into the hidden space.

Unlike in constituency trees, where all words reside at the leaf level, internal nodes of dependency trees are associated with words. Hence, the DT-RNN combines the current node's word vector with its children's hidden vectors to form h_n. This procedure continues in a recursive way up to the root, which presents the whole sentence.

They relate a separate $d \times d$ matrix W_r with each dependency relation r in their dataset, and these matrices are learnt during training. They include an extra $d \times d$ matrix, W_v to integrate the word vector x_w at a node into the node vector h_n. For example, the hidden representation h_{helots} is

$$h_{helots} = f(W_v \cdot x_{helots} + b), \tag{18}$$

where f stands for a non-linear activation function such as $tanh$ and b is a bias term.

When all leaves are completed, they proceed to interior nodes with already processed children. Continuing from "helots" to its parent, "called", they compute

$$h_{called} = f\left(W_{DOBJ} \cdot h_{helots} + W_v \cdot x_{called} + b\right) \tag{19}$$

They repeat this process up to the root, which is

$$h_{depended} = f\left(W_{NSUBJ} \cdot h_{economy} + W_{PREP} \cdot h_{on} + W_v \cdot x_{depended} + b\right) \tag{20}$$

The composition equation for any node n with children $K(n)$ and word vector x_w is

$$h_n = f(W_v \cdot x_w + b + \sum_{k \in K(n)} W_{R(n,k)} \cdot h_k) \tag{21}$$

where $R(n,k)$ is the dependency relation between node n and child node k.

3.5. Hybrid Deep Neural Networks Based Models

Different neural network models can be combined to make a very effective model. In this section, we discussed various hybrid deep models used by many researchers to improve the MQA task.

A hybrid method was proposed by Zhang, et al., where they addressed the problem of complex relationship between questions and answers, with the aim of enhancing the Chinese medical question–answer selection [65]. They combined single and hybrid models with CNN and GRU to benefit from the merits of different neural network architectures.

Another hybrid method named Template-Representation-Based Convolutional Recurrent Neural Network (T-CRNN) was proposed by Reddy, et al., to select an answer in the Complex Question Answering (CQA) framework. Firstly, they replaced the entity from the input question with the templates. A divide and conquer approach was employed to decompose the question, based on the replaced entities. Then the CNN, RNN and scoring is used to determine the correct answer [72].

Similarly, Duan, et al. [73] proposed two types of question generation approach. The first one is a retrieval-based method that uses CNN; the second one is a generation-based method that uses RNN. In addition, they also show how the generated questions can be used to improve existing question answering systems.

In [66], a CNN-LSTM attention model is proposed to predict user intent, and an unsupervised clustering method is applied to mine user intent taxonomy (see Figure 14). The CNN-LSTM attention model has a CNN encoder and a Bi-LSTM attention encoder. The two encoders can capture both global semantic expression and local phrase-level information from an original medical text query, which helps the intent prediction. Their model is described as follows.

For CNN-Encoder, they implemented a convolutional layer with one dimensional convolution operation. After this, they applied a max-over-time pooling operation with the feature maps built by CNN filters.

For a feature map $c = [c_1, c_2, \ldots, c_n], c \in \mathbb{R}$, $\hat{c} = \max\{c_i\}$ is the maximum feature, which is selected to represent the intensity of this particular filter in a query.

For Bi-LSTM Attention Encoder, an LSTM model process a vector sequence input $X = [x_1, x_2, \ldots, x_n]$ from beginning to end and calculates a hidden state for each time step as

$$h_i = LSTM(h_{i-1}, x_i) \tag{22}$$

A Bi-LSTM model has two LSTMs (LSTM, LSTM) reading the same input sequence with different directions. They concatenate two hidden state $\left[\overrightarrow{h_i}, \overleftarrow{h_i}\right]$ as the final hidden state output h_i of input x_i.

Moreover, they applied the attention mechanism to extract important words to the intent of the query q to a sequence of hidden state $H = [h_1, h_2, \ldots, h_n]$. In this way, a query could be encoded into a vector v, and each attention scalar α_i can demonstrate the attention degree for the i-th word in query q.

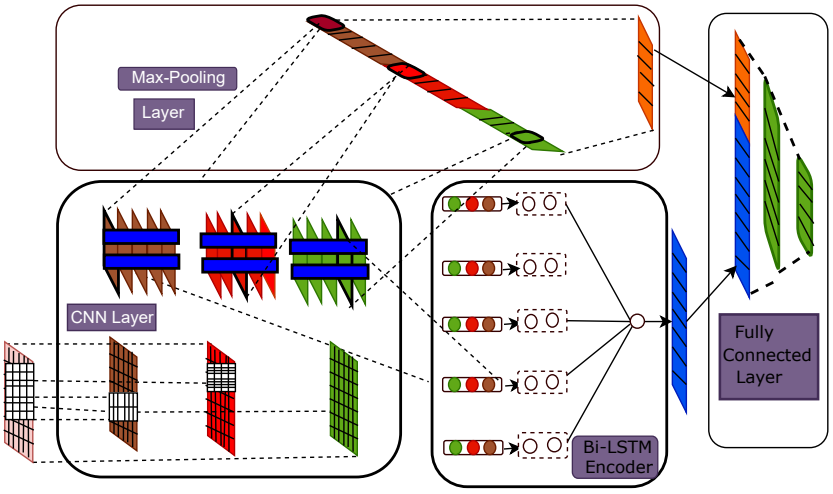

Figure 14. Hybrid architecture of CNN and BiLSTM.

4. Evaluation Metrics and Datasets

In this section, we discuss the evaluation metrics and datasets used for QA in the medical field. Evaluation metrics intend to objectively measure the effectiveness of a given method. In this case, well-defined and effective evaluation metrics are crucial to MQA research.

4.1. Evaluation Metrics

Evaluation is one of the essential dimensions in QA, as it assesses and compares the answers to measure the performances of QASs [74]. Much effort has been put into addressing the problem of the performance evaluation of IR systems [40]. Based on these efforts, there are two approaches to the performance evaluation of QASs: system-centered and user-centered.

Evaluation methods play an important role in the QA system. With the rapid development of QA methods, reliable evaluation metrics are needed to compare these imple-

mentations [34]. The metrics often used in QA are *F*1 and accuracy [38]. For any given datapoint to be assessed, there must be two categories: a segment that is correctly selected (true positive) or not correctly selected (false negative), and a segment that is not correctly selected (false positive) or incorrectly not selected (true negative). Equation (23) shows how *Accuracy* is calculated

$$Accurary = \frac{TP + TN}{TP + FP + FN + TN} \tag{23}$$

where TP, FP, TN, and FN are representing true positive, false positive, true negative, and false negative, respectively.

It can be observed for the QA systems metric in evaluation, that the system gives the correct answer when a fact question is asked, and vice versa. Therefore, the system can be found to have a high true negative rate. For example, the system may have high computational accuracy, but it is not meaningful. To solve this problem, the *F*-measure is adopted, which is based on two metrics: Precision and Recall. The first is the percentage of selected answers among correct answers, while the second is the reverse measure. That means that Recall is the percentage of correct answers selected. When using Precision and Recall, there is no longer a real fact that the rate of negative answers is high.

Table 3 summarizes the concepts of the three evaluation metrics: *Accuracy*, *Precision*, and *Recall*.

$$Precision = \frac{TP}{TP + FP}(A1) \tag{24}$$

$$Recall = \frac{TP}{TP + FN}(A2) \tag{25}$$

Table 3. Accuracy, Precision and Recall.

Actual \ Predicted	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Equations (24) and (25) give a fundamental understanding of the researchers' trade-off to conduct in each measure in looking for the best metrics to evaluate their systems. Most of the fact QA systems should use Equation (25) as a measured metric, since it does not matter how high the false positive rates are; the result will be good if there are high true positive rates. However, for list or definition QA systems, Equation (24) is assumed to function better. To balance this trade-off, the *F*-measure is presented.

$$F_1 = \frac{2(Precision \cdot Recall)}{Precision + Recall} \tag{26}$$

Equation (26) implements a weighted approach to assessment of the Precision and Recall trade-off. Some metrics can be used to evaluate QA systems, such as Mean Average Precision (MAP) presented in Equation (27), and Mean Reciprocal Rank (MRR) shown in Equation (29) used to calculate the answer relevance. The two are mainly used in IR paradigms.

$$MAP = \frac{\sum_{Q}^{q=1} AveP(q)}{Q} \tag{27}$$

where Q stands for queries and the average precision ($AveP$) can be expressed as

$$AveP = \frac{\sum_{N}^{k=1} P(k) \times rel(q)}{|\{relevant\}|} \tag{28}$$

$$MRR = \frac{1}{n}\sum_{i=1}^{N} RR_{(q_i)} \quad (29)$$

Evaluation metrics can also be categorized based on the types of question. For factoid questions, they consider the highest probability answer as the exact answer. Three evaluation metrics used for factoid questions are: Strict Accuracy, Lenient Accuracy, and MRR. For the list questions, they set a threshold, then all the predictions above that threshold are considered as the list of answers to the question. The three evaluation metrics used for list questions are *Precision*, *Recall*, and *F*1 score. For yes/no questions, the first CLS (which can be used as a sentence representation) from the output layer is used, combined with a fully connected layer along with dropout to get the logit values. The positive logit values represent a 'yes', while negative represents a 'no'. For each question, all the logit values for all question-context are added together and if the final value obtained is positive, then it is classified as a "yes" otherwise as "no". The evaluation metrics used for yes/no questions are: Accuracy, *F*1 score, *F*1 yes, and *F*1 no scores.

Figure 15 demonstrates how Accuracy, Precision, Recall, and *F*1 are employed to evaluate the different methods used in the question answering systems, where KNN is the *K*-Nearest Neighbor method, GaussianNB is Gaussian Naive Bayes method, RF is Random Forest method, SVM is Support Vector Machine method, PPN is Perceptron method, and Dia-AID is the method proposed in [75].

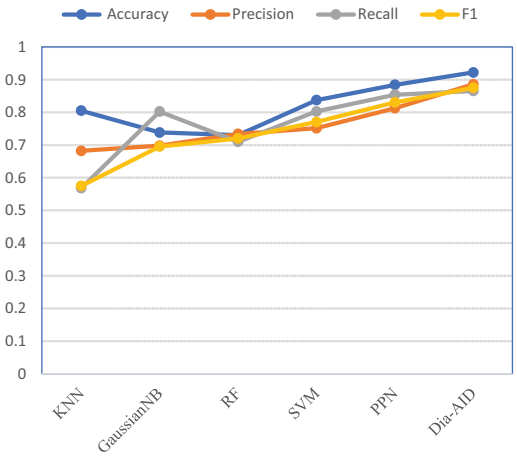

Figure 15. The evaluation for different classification methods in [75].

4.2. Datasets

In this subsection, we discuss the common datasets used for QA task in the medical domain. Our review excludes datasets used in the open domain QA, where any kind of question can find an answer. We only focus on the publicly available medical datasets used by health workers to find relevant answers that assist them in their profession.

(1) MedQuAD: Medical Question Answering Dataset

This dataset was created from 12 National Institutes of Health (NIH) websites, such as cancer.gov, niddk.nih.gov, MedlinePlus Health Topics, GARD, etc. It contains 47,457 medical question–answer pairs, and covers 37 question types, such as Treatment, Diagnosis, and Side Effects, associated with diseases, drugs, and other medical entities [63]. The experiments of [63] were conducted on this dataset, and the experimental results show that the inter-annotator agreement through *F*1 score achieved 88.5% agreement on the four categories and a 94.3% agreement when the categories are reduced to two.

(2) EmrQA: Electronic Medical Records for Question Answering

EmrQA is a dataset that focuses on the medical domain [76]. It is worth mentioning that the EmrQA dataset is leveraged by utilizing the existing expert annotations on clinical notes for different NLP tasks in the community-shared i2b2 dataset. The EmrQA dataset

has one million question-logical forms and 400,000+ question–answer evidence pairs. The model proposed in [76] was evaluated on the dataset EmrQA. The experimental results show that the proposed model can achieve 60.6% for $F1$ and 59.2% for EM (Exact Match).

(3) QA4MRE: Question Answering for Machine Reading Evaluation

This dataset contains three topics: Climate change, Aids, and Music & Society [13,77]. Each topic includes four reading tests. Each reading test consists of one single document, with 10 questions and a set of five choices per question. In total, there are 16 test documents (four documents for every three topics), 160 questions (10 questions for each document), with 800 choices (5 for each question). Test documents and questions were made available in English, German, Italian, Romanian, and Spanish. These materials were the same in all languages, created using parallel translations. The evaluation performed on the QA4MRE dataset is presented in [77], and the experimental result showed that 53% of the questions were answered with the correct candidate.

(4) cMedQA: Chinese Medical Question and Answers

This is the dataset for Chinese community MQA [60]. The cMedQA dataset facilitates choosing QA pairs from some real-world online health (http://www.xywy.com/, accessed on 15 December 2020) and wellness communities, such as DingXiangYuan and XunYi-WenYao. The dataset consists of 101,743 QA pairs, where the sentences were split into individual characters. The vocabulary has a total of 4979 tokens. In [60], different models such as SingleCNN, biLSTM, and Multi-CNN were compared on the cMedQA dataset. The results show that these models can achieve 64.05%, 63.20%, and 64.75% accuracy, respectively.

(5) MASH-QA: Multiple Answer Spans Healthcare Question Answering

MASH-QA is a large-scale dataset for QA, with many answers coming from multiple spans within a long document. The dataset consists of over 35,000 QA pairs and is based on questions and knowledge articles from the consumer health domain, where the questions are generally non-factoid in nature and cannot be answered using just a few words [78]. The experimental results in [78] show that using models of DrQA Reader, BiDAF, BERT, SpanBERT, XLnet and MultiCo, on the MASH-QA dataset, the $F1$ are 18.92%, 23.19%, 27.93%, 30.61%, 56.46%, and 64.94% and EM are 1.82%, 2.42%, 3.95%, 5.62%, 22.78%, and 29.49% respectively.

Table 4 gives the details of different QA datasets used in the healthcare domain.

Table 4. Comparison of different healthcare QA datasets.

Dataset	#QA	QA Type	Source
MedQUAD	47K	Ranking	Health articles
EmrQA	400K	Extractive	Medical records
QA4MRE	600	Ranking	Medical records
cMedQA	101,7K	Extractive	Health articles
MASH-QA	35K	Extractive	Health articles

5. Existing Challenges and Future Directions

Although MQA has made considerable progress, there are still many uncertainties and limitations in the existing research. For example, the binary relationships still cannot represent all questions when the answer extraction is obtained by the comparison between the annotations of knowledge bases and user question. In addition, the acquisition of scientific knowledge is also a limitation of the system, because only experts in the field are able to add knowledge and increase system coverage [68,79].

During the review of MQASs based on deep learning approaches, we realized that researchers still face several challenges, and we suggested key directions that they should focus on in the future. Some of the following challenges remain either unsolved or answered to some extent.

(1) Retrieval of relevant and reliable answers

The existing medical search engines do not provide relevant answers on time. There is still a delay in patients obtaining the answer they need [80]. Researchers should build models that can provide relevant answers in few seconds or automatically. Document summarization still take a long time to generate the summary of the documents based on the question asked by the user. In addition, there is also a problem in the MQA system's ability to provide precise summaries from the original document. To reduce the response time of summary generation, more MQA corpus should be built for frequently asked questions that do not have answers [81].

(2) Lack of large medical datasets

The key issue with medical datasets, especially for clinical paraphrasing, consist of either short passages or web page title texts, both of which are not suitable to build a paraphrase generator for QA [82]. In addition, there is a lack of annotated data, and ambiguity in the clinical text which hinders the development of medical datasets. Therefore, there is a need to develop appropriate medical datasets to improve QA in the medical field [83].

(3) Development of medical recommendation systems

There is a need for medical recommendation systems that can provide treatment recommendations according to the description of the symptoms given by users. Specifically, there is an increasing demand for Q&A systems to successfully and efficiently assist diabetes patients [36,84,85]. The current diabetes management applications only provide general information management and search, but ignore a vital counselling service, which is critical for dealing with the health condition of patients living with diabetes. Therefore, by developing a QAS that can recommend the patients the types of medicines, they can take or give advice based on the symptoms provided by the patients.

(4) Development of collaborative medical question answering systems

In collaborative QAS (also known as community QAS), such as Wiki Answers and Yahoo Answers, answers are provided by users questions asked by other users, and the best answer is chosen manually by the questioner or by all participants through voting [74]. In the medical field, these systems allow physicians to interact with patients by effectively answering their questions.

(5) Diversity of medical questions

To answer questions in the medical field requires an in-depth understanding of the field. MQA passages from textbooks often do not directly answer questions, especially for case problems. One must discern relevant information scattered in passages and determine the relevance of each piece of text [86].

6. Conclusions

Medical QA has made significant progress in recent years due to the use of deep learning techniques in this area. Automatic QA has been possible in many medical question–answering systems, and the availability of corpus data in the medical domain is increasing over time. In this paper, we provided an extensive review of the prominent works on deep-learning-based medical textual QA. The study started with an overview of QAS and provided a brief outline of the tasks, types, and the representative of medical QAS. Next, we highlighted recent deep learning approaches and their various architectures, utilized in MQA tasks. Moreover, we discussed the existing medical textual QA datasets and evaluation metrics used to measure the performance of the medical QAS. Finally, we summarized recent QA challenges in the medical domain and recommended some promising future research directions. Our contributions in this work are gathering the literature of the recent works on medical QAS, summarizing the application of deep learning approaches in the medical domain and providing the relevant information to potential researchers who want to choose MQA as their research field.

Author Contributions: Funding acquisition, J.N.; Project administration, J.N. and W.C.; Writing—original draft, E.M.; Writing—review and editing, E.M. and G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China (61873086).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the authors of all the references.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wilensky, R.; Chin, D.N.; Luria, M.; Martin, J.; Mayfield, J.; Wu, D. The Berkeley UNIX consultant project. In *Intelligent Help Systems for UNIX*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 49–94. [CrossRef]
2. Ludwig, T.; Walter, B.; Ley, M.; Maier, A.; Gehlen, E. LILOG-DB: Database Support for Knowledge-Based Systems. In *Datenbanksysteme in Büro, Technik und Wissenschaft*; Springer: Berlin/Heidelberg, Germany, 1989; pp. 176–195. [CrossRef]
3. Olvera-Lobo, M.D.; Gutiérrez-Artacho, J. Question answering track evaluation in TREC, CLEF and NTCIR. In *New Contributions in Information Systems and Technologies*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 13–22. [CrossRef]
4. Lopez, V.; Uren, V.; Motta, E.; Pasin, M. AquaLog: An ontology-driven question answering system for organizational semantic intranets. *J. Web Semant.* **2007**, *5*, 72–105. [CrossRef]
5. Barskar, R.; Ahmed, G.F.; Barskar, N. An approach for extracting exact answers to Question Answering (QA) system for english sentences. *Procedia Eng.* **2012**, *30*, 1187–1194. [CrossRef]
6. Badugu, S.; Manivannan, R. A study on different closed domain question answering approaches. *Int. J. Speech Technol.* **2020**, *23*, 315–325. [CrossRef]
7. Mittal, S.; Mittal, A. Versatile question answering systems: Seeing in synthesis. *Int. J. Intell. Inf. Database Syst.* **2011**, *5*, 119–142. [CrossRef]
8. Yu, Z.; Yu, J.; Xiang, C.; Fan, J.; Tao, D. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 5947–5959. [CrossRef] [PubMed]
9. Balikas, G.; Krithara, A.; Partalas, I.; Paliouras, G. BioASQ: A challenge on large-scale biomedical semantic indexing and question answering. In *International Workshop on Multimodal Retrieval in the Medical Domain*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 26–39. [CrossRef]
10. Wasim, M.; Asim, M.N.; Khan, M.U.G.; Mahmood, W. Multi-label biomedical question classification for lexical answer type prediction. *J. Biomed. Inform.* **2019**, *93*, 103143. [CrossRef] [PubMed]
11. Yoo, I.; Mosa, A.S.M. Analysis of PubMed user sessions using a full-day PubMed query log: A comparison of experienced and nonexperienced PubMed users. *JMIR Med. Inform.* **2015**, *3*, e25. [CrossRef] [PubMed]
12. Al Fayez, R.Q.; Joy, M. Using Linked Data for Integrating Educational Medical Web Databases Based on BioMedical Ontologies. *Comput. J.* **2017**, *60*, 369–388. [CrossRef]
13. Hamon, T.; Grabar, N.; Mougin, F. Natural Language Question Analysis for Querying Biomedical Linked Data. In Proceedings of the 1st Natural Language Interfaces for Web of Data Workshop (NLIWoD)—ISWC, Riva del Garda, Italy, 19 October 2014.
14. Habimana, O.; Li, Y.; Li, R.; Gu, X.; Yan, W. Attentive convolutional gated recurrent network: A contextual model to sentiment analysis. *Int. J. Mach. Learn. Cybern.* **2020**, *11*, 2637–2651. [CrossRef]
15. Piskorski, J.; Yangarber, R. Information extraction: Past, present and future. In *Multi-Source, Multilingual Information Extraction and Summarization*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 23–49. [CrossRef]
16. Sankarasubramaniam, Y.; Ramanathan, K.; Ghosh, S. Text summarization using Wikipedia. *Inf. Process. Manag.* **2014**, *50*, 443–461. [CrossRef]
17. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef] [PubMed]
18. Derici, C.; Aydin, Y.; Yenialaca, Ç.; AYDIN, N.Y.; Kartal, G.; Özgür, A.; Güngör, T. A closed-domain question answering framework using reliable resources to assist students. *Nat. Lang. Eng.* **2018**, *24*, 725–762. [CrossRef]
19. Bauer, M.A.; Berleant, D. Usability survey of biomedical question answering systems. *Hum. Genom.* **2012**, *6*, 1–4. [CrossRef]
20. Cao, Y.G.; Cimino, J.J.; Ely, J.; Yu, H. Automatically extracting information needs from complex clinical questions. *J. Biomed. Inform.* **2010**, *43*, 962–971. [CrossRef] [PubMed]
21. Roberts, K.; Patra, B.G. A semantic parsing method for mapping clinical questions to logical forms. In Proceedings of the American Medical Informatics Association Annual Symposium Proceedings, Washington, DC, USA, 4–8 November 2017; p. 1478.
22. Sarrouti, M.; El Alaoui, S.O. A passage retrieval method based on probabilistic information retrieval model and UMLS concepts in biomedical question answering. *J. Biomed. Inform.* **2017**, *68*, 96–103. [CrossRef]

23. Ye, D.; Zhang, S.; Wang, H.; Cheng, J.; Zhang, X.; Ding, Z.; Li, P. Multi-level composite neural networks for medical question answer matching. In Proceedings of the 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), Guangzhou, China, 18–21 June 2018; pp. 139–145. [CrossRef]
24. Niu, Y.; Hirst, G. Analysis of Semantic Classes in Medical Text for Question Answering. Available online: https://www.aclweb.org/anthology/W04-0509.pdf (accessed on 15 March 2021).
25. Dai, D.; Tang, J.; Yu, Z.; Wong, H.S.; You, J.; Cao, W.; Hu, Y.; Chen, C.P. An inception convolutional autoencoder model for chinese healthcare question clustering. *IEEE Trans. Cybern.* **2019**, *51*, 2019–2031. [CrossRef] [PubMed]
26. Cai, L.Q.; Wei, M.; Zhou, S.T.; Yan, X. Intelligent Question Answering in Restricted Domains Using Deep Learning and Question Pair Matching. *IEEE Access* **2020**, *8*, 32922–32934. [CrossRef]
27. Mishra, A.; Jain, S.K. A survey on question answering systems with classification. *J. King Saud Univ. Comput. Inf. Sci.* **2016**, *28*, 345–361. [CrossRef]
28. Athenikos, S.J.; Han, H. Biomedical question answering: A survey. *Comput. Methods Programs Biomed.* **2010**, *99*, 1–24. [CrossRef]
29. Zhang, X.; Wu, J.; He, Z.; Liu, X.; Su, Y. Medical exam question answering with large-scale reading comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
30. Hu, Z.; Zhang, Z.; Yang, H.; Chen, Q.; Zuo, D. A deep learning approach for predicting the quality of online health expert question-answering services. *J. Biomed. Inform.* **2017**, *71*, 241–253. [CrossRef]
31. Lee, M.; Cimino, J.; Zhu, H.R.; Sable, C.; Shanker, V.; Ely, J.; Yu, H. Beyond information retrieval—Medical question answering. In Proceedings of the American Medical Informatics Association Annual Symposium Proceedings, Washington, DC, USA, 11–15 November 2006; p. 469.
32. Yu, H.; Lee, M.; Kaufman, D.; Ely, J.; Osheroff, J.A.; Hripcsak, G.; Cimino, J. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *J. Biomed. Inform.* **2007**, *40*, 236–251. [CrossRef] [PubMed]
33. Sharma, S.; Patanwala, H.; Shah, M.; Deulkar, K. A survey of medical question answering systems. *Int. J. Eng. Tech. Res.* **2015**, *3*, 131–133.
34. Kodra, L.; Meçe, E.K. Question answering systems: A review on present developments, challenges and trends. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*. [CrossRef]
35. Xiong, C.; Su, M. IARNN-Based Semantic-Containing Double-Level Embedding Bi-LSTM for Question-and-Answer Matching. *Comput. Intell. Neurosci.* **2019**, *2019*, 6074840. [CrossRef]
36. Neves, M.; Leser, U. Question answering for biology. *Methods* **2015**, *74*, 36–46. [CrossRef]
37. Dina, D.F.; Yassine, M.; Asma, B.A. Consumer health information and question answering: Helping consumers find answers to their health-related information needs. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 194–201. [CrossRef]
38. Ray, S.K.; Singh, S.; Joshi, B.P. A semantic approach for question classification using WordNet and Wikipedia. *Pattern Recognit. Lett.* **2010**, *31*, 1935–1943. [CrossRef]
39. Gupta, P.; Gupta, V. A survey of text question answering techniques. *Int. J. Comput. Appl.* **2012**, *53*, 1–8. [CrossRef]
40. Seena, I.; Sini, G.; Binu, R. Malayalam question answering system. *Procedia Technol.* **2016**, *24*, 1388–1392. [CrossRef]
41. Hamed, S.K.; Ab Aziz, M.J. A Question Answering System on Holy Quran Translation Based on Question Expansion Technique and Neural Network Classification. *J. Comput. Sci.* **2016**, *12*, 169–177. [CrossRef]
42. Kolomiyets, O.; Moens, M.F. A survey on question answering technology from an information retrieval perspective. *Inf. Sci.* **2011**, *181*, 5412–5434. [CrossRef]
43. Tsatsaronis, G.; Balikas, G.; Malakasiotis, P.; Partalas, I.; Zschunke, M.; Alvers, M.R.; Weissenborn, D.; Krithara, A.; Petridis, S.; Polychronopoulos, D.; et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.* **2015**, *16*, 1–28. [CrossRef] [PubMed]
44. Heie, M.H.; Whittaker, E.W.; Furui, S. Question answering using statistical language modelling. *Comput. Speech Lang.* **2012**, *26*, 193–209. [CrossRef]
45. García-Cumbreras, M.; Martínez-Santiago, F.; Ureña-López, L. Architecture and evaluation of BRUJA, a multilingual question answering system. *Inf. Retr.* **2012**, *15*, 413–432. [CrossRef]
46. Noh, J.; Kavuluru, R. Document retrieval for biomedical question answering with neural sentence matching. In Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, USA, 17–20 December 2018; pp. 194–201. [CrossRef]
47. Goodwin, T.R.; Harabagiu, S.M. Knowledge representations and inference techniques for medical question answering. *ACM Trans. Intell. Syst. Technol.* **2017**, *9*, 1–26. [CrossRef]
48. Soares, M.A.C.; Parreiras, F.S. A literature review on question answering techniques, paradigms and systems. *J. King Saud Univ. Comput. Inf. Sci.* **2020**, *32*, 635–646. [CrossRef]
49. Sarrouti, M.; El Alaoui, S.O. SemBioNLQA: A semantic biomedical question answering system for retrieving exact and ideal answers to natural language questions. *Artif. Intell. Med.* **2020**, *102*, 101767. [CrossRef] [PubMed]
50. Hou, W.; Tsai, B.H. An Answer Validation Concept Based Approach for Question Answering in Biomedical Domain. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Kaohsiung, Taiwan, 3–6 June 2014. [CrossRef]
51. Cao, Y.; Liu, F.; Simpson, P.; Antieau, L.D.; Bennett, A.S.; Cimino, J.; Ely, J.; Yu, H. AskHERMES: An online question answering system for complex clinical questions. *J. Biomed. Inform.* **2011**, *44*, 277–288. [CrossRef] [PubMed]

52. Cairns, B.L.; Nielsen, R.D.; Masanz, J.J.; Martin, J.H.; Palmer, M.S.; Ward, W.H.; Savova, G.K. The MiPACQ clinical question answering system. In Proceedings of the American Medical Informatics Association Annual Symposium Proceedings, Washington DC, USA, 22–26 October 2011; Volume 2011, p. 171.
53. Burns, C.S.; Nix, T.; Shapiro, R.M.; Huber, J.T. MEDLINE search retrieval issues: A longitudinal query analysis of five vendor platforms. *PLoS ONE* **2021**, *16*, e0234221. [CrossRef]
54. Ni, Y.; Zhu, H.; Cai, P.; Zhang, L.; Qui, Z.; Cao, F. CliniQA : Highly Reliable Clinical Question Answering System. *Stud. Health Technol. Inform.* **2012**, *180*, 215–219. [CrossRef] [PubMed]
55. Zhu, X.; Yang, X.; Chen, H. A Biomedical Question Answering System Based on SNOMED-CT. In Proceedings of the International Conference on Knowledge Science, Engineering and Management, Changchun, China, 17–19 August 2018. [CrossRef]
56. Murdock, J.W.; Tesauro, G.; Tj, I. Statistical approaches to question answering in Watson. In *Mathematics Awareness Month Theme Essay*; IBM TJ Watson Research Center: Ossining, NY, USA, 2012.
57. Wren, J.D. Question answering systems in biology and medicine-the time is now. *Bioinformatics* **2011**, *27*, 2025–2026. [CrossRef] [PubMed]
58. Ni, J.; Chen, Y.; Chen, Y.; Zhu, J.; Ali, D.; Cao, W. A Survey on Theories and Applications for Self-Driving Cars Based on Deep Learning Methods. *Appl. Sci.* **2020**, *10*, 2749. [CrossRef]
59. Yan, Y.; Zhang, B.; Li, X.; Liu, Z. List-wise learning to rank biomedical question-answer pairs with deep ranking recursive autoencoders. *PLoS ONE* **2020**, *15*, e0242061. [CrossRef]
60. Zhang, S.; Zhang, X.; Wang, H.; Cheng, J.; Li, P.; Ding, Z. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. *Appl. Sci.* **2017**, *7*, 767. [CrossRef]
61. He, J.; Fu, M.; Tu, M. Applying deep matching networks to Chinese medical question answering: A study and a dataset. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 91–100. [CrossRef]
62. Luo, Y. Recurrent neural networks for classifying relations in clinical notes. *J. Biomed. Inform.* **2017**, *72*, 85–95. [CrossRef]
63. Asma, B.A.; Dina, D.F. A question-entailment approach to question answering. *BMC Bioinform.* **2019**, *20*, 1–23. [CrossRef]
64. Iyyer, M.; Boyd-Graber, J.; Claudino, L.; Socher, R.; Daumé III, H. A neural network for factoid question answering over paragraphs. In Proceedings of the 2014 Conference on Empirical METHODS in Natural Language processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 633–644. [CrossRef]
65. Cai, R.; Zhu, B.; Ji, L.; Hao, T.; Yan, J.; Liu, W. An CNN-LSTM attention approach to understanding user query intent from online health communities. In Proceedings of the 2017 IEEE international conference on data mining workshops (ICDMW), New Orleans, LA, USA, 18–21 November 2017; pp. 430–437. [CrossRef]
66. Zhang, Y.; Lu, W.; Ou, W.; Zhang, G.; Zhang, X.; Cheng, J.; Zhang, W. Chinese medical question answer selection via hybrid models based on CNN and GRU. *Multimed. Tools Appl.* **2020**, *79*, 14751–14776. [CrossRef]
67. Ibrahim, Y.; Wang, H.; Bai, M.; Liu, Z.; Wang, J.; Yang, Z.; Chen, Z. Soft Error Resilience of Deep Residual Networks for Object Recognition. *IEEE Access* **2020**, *8*, 19490–19503. [CrossRef]
68. Abacha, A.B.; Zweigenbaum, P. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Inf. Process. Manag.* **2015**, *51*, 570–594. [CrossRef]
69. Habimana, O.; Li, Y.; Li, R.; Gu, X.; Yu, G. Sentiment analysis using deep learning approaches: An overview. *Sci. China Inf. Sci.* **2020**, *63*, 1–36. [CrossRef]
70. He, B.; Guan, Y.; Dai, R. Convolutional Gated Recurrent Units for Medical Relation Classification. In Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 3–6 December 2018; pp. 646–650. [CrossRef]
71. Stroh, E.; Mathur, P. Question Answering Using Deep Learning. Available online: http://cs224d.stanford.edu/reports/StrohMathur.pdf (accessed on 15 March 2021).
72. Reddy, A.C.O.; Madhavi, K. Convolutional recurrent neural network with template based representation for complex question answering. *Int. J. Electr. Comput. Eng.* **2020**, *10*, 2710. [CrossRef]
73. Duan, N.; Tang, D.; Chen, P.; Zhou, M. Question generation for question answering. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; pp. 866–874. [CrossRef]
74. Ray, S.K.; Shaalan, K. A review and future perspectives of arabic question answering systems. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 3169–3190. [CrossRef]
75. Xie, W.; Ding, R.; Yan, J.; Qu, Y. A mobile-based question-answering and early warning system for assisting diabetes management. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 9163160. [CrossRef]
76. Pampari, A.; Raghavan, P.; Liang, J.; Peng, J. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2357–2368. [CrossRef]
77. Bhaskar, P.; Pakray, P.; Banerjee, S.; Banerjee, S.; Bandyopadhyay, S.; Gelbukh, A.F. Question Answering System for QA4MRE@CLEF 2012. Available online: http://ceur-ws.org/Vol-1178/CLEF2012wn-QA4MRE-BhaskarEt2012b.pdf (accessed on 15 March 2021).
78. Zhu, M.; Ahuja, A.; Juan, D.C.; Wei, W.; Reddy, C.K. Question Answering with Long Multiple-Span Answers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, Online Event, 16–20 November 2020; pp. 3840–3849. [CrossRef]

79. Katz, B.; Felshin, S.; Yuret, D.; Ibrahim, A.; Lin, J.; Marton, G.; Mcfarland, A.J.; Temelkuran, B. Omnibase: Uniform Access to Heterogeneous Data for Question Answering. In Proceedings of the International Conference on Application of Natural Language to Information Systems, Salford, UK, 26–28 June 2002. [CrossRef]
80. Quantin, C.; Jaquet-Chiffelle, D.O.; Coatrieux, G.; Benzenine, E.; Allaert, F.A. Medical record search engines, using pseudonymised patient identity: An alternative to centralised medical records. *Int. J. Med. Inform.* **2011**, *80*, e6–e11. [CrossRef] [PubMed]
81. Karpagam, K.; Saradha, A.; Manikandan, K.; Madusudanan, K. Text Summarization using QA Corpus for User Interaction Model QA System. *Int. J. Educ. Manag. Eng.* **2020**, *10*, 33. [CrossRef]
82. Soni, S.; Roberts, K. A paraphrase generation system for ehr question answering. In Proceedings of the 18th BioNLP Workshop and Shared Task, Florence, Italy, 1 August 2019; pp. 20–29. [CrossRef]
83. Nguyen, V. Question answering in the biomedical domain. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Florence, Italy, 28 July–2 August 2019; pp. 54–63. [CrossRef]
84. Yoo, H.; Chung, K. PHR based diabetes index service model using life behavior analysis. *Wirel. Pers. Commun.* **2017**, *93*, 161–174. [CrossRef]
85. Chavez, S.; Fedele, D.; Guo, Y.; Bernier, A.; Smith, M.; Warnick, J.; Modave, F. Mobile apps for the management of diabetes. *Diabetes Care* **2017**, *40*, e145–e146. [CrossRef]
86. Miller, D.D.; Brown, E.W. Artificial intelligence in medical practice: The question to the answer? *Am. J. Med.* **2018**, *131*, 129–133. [CrossRef]