Published: 18 January 2022
Accepted: 01 September 2021
Revised: 01 August 2021
Received: 01 March 2021

SURVEY

Biomedical Question Answering: A Survey of Approaches and Challenges

QIAO JIN, Tsinghua University, Beijing, China

ZHENG YUAN, Tsinghua University, Beijing, China

GUANGZHI XIONG, Tsinghua University, Beijing, China

QIANLAN YU, Tsinghua University, Beijing, China

HUAIYUAN YING, Tsinghua University, Beijing, China

CHUANQI TAN, Alibaba Group Holding Limited, Hangzhou, Zhejiang, China

View all

Biomedical Question Answering: A Survey of Approaches and Challenges

QIAO JIN, ZHENG YUAN, GUANGZHI XIONG, QIANLAN YU, and HUAIYUAN YING, Tsinghua University, China
CHUANQI TAN, MOSHA CHEN, and SONGFANG HUANG, Alibaba Group, China
XIAOZHONG LIU, Indiana University Bloomington, USA
SHENG YU, Tsinghua University, China

Automatic Question Answering (QA) has been successfully applied in various domains such as search engines and chatbots. Biomedical QA (BQA), as an emerging QA task, enables innovative applications to effectively perceive, access, and understand complex biomedical knowledge. There have been tremendous developments of BQA in the past two decades, which we classify into five distinctive approaches: classic, information retrieval, machine reading comprehension, knowledge base, and question entailment approaches. In this survey, we introduce available datasets and representative methods of each BQA approach in detail. Despite the developments, BQA systems are still immature and rarely used in real-life settings. We identify and characterize several key challenges in BQA that might lead to this issue, and we discuss some potential future directions to explore.

CCS Concepts: • **Applied computing** → **Life and medical sciences**; • **Computing methodologies** → **Machine learning**; **Natural language processing;**

Additional Key Words and Phrases: Question answering, natural language processing, machine learning, biomedicine

ACM Reference format:
Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical Question Answering: A Survey of Approaches and Challenges. *ACM Comput. Surv.* 55, 2, Article 35 (January 2022), 36 pages.
https://doi.org/10.1145/3490238

1 INTRODUCTION

Biomedical knowledge acquisition is an important task in information retrieval and knowledge management. Professionals as well as the general public need effective assistance to access, understand, and consume complex biomedical concepts. For example, doctors always want to be

Qiao Jin and Zheng Yuan work done during internship at Alibaba.
Authors' addresses: Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, and S. Yu (corresponding author), Tsinghua University, China; emails: {jqa14, yuanz17, xgz18, yuql18, yinghy18}@mails.tsinghua.edu.cn, syu@tsinghua.edu.cn; C. Tan (corresponding author), M. Chen, and S. Huang, Alibaba Group, China; emails: {chuanqi.tcq, chenmosha.cms, songfang.hsf}@alibaba-inc.com; X. Liu, Indiana University Bloomington; email: liu237@indiana.edu.

0360-0300/2022/01-ART35 $15.00
https://doi.org/10.1145/3490238

aware of up-to-date clinical evidence for the diagnosis and treatment of diseases under the scheme of Evidence-based Medicine [165], and the general public is becoming increasingly interested in learning about their own health conditions on the Internet [54].

Traditionally, **Information Retrieval (IR)** systems, such as PubMed, have been used to meet such information needs. However, classical IR is still not efficient enough [71, 77, 99, 164]. For instance, Russell-Rose and Chamberlain [164] reported that it requires four expert hours to answer complex medical queries using search engines. Compared with the retrieval systems that typically return a list of relevant documents for the users to read, **Question Answering (QA)** systems that provide direct answers to users' questions are more straightforward and intuitive. In general, QA itself is a challenging benchmark **Natural Language Processing (NLP)** task for evaluating the abilities of intelligent systems to understand a question, retrieve and utilize relevant materials, and generate its answer. With the rapid development of computing hardware, modern QA models, especially those based on deep learning [30, 31, 42, 146, 171], achieve comparable or even better performance than human on many benchmark datasets [67, 83, 154, 155, 215] and have been successfully adopted in general domain search engines and conversational assistants [150, 236].

The **Text REtrieval Conference (TREC)** QA Track has triggered the modern QA research [197], when QA models were mostly based on IR. Zweigenbaum [241] first identified the distinctive characteristics of BQA over general domain QA. Later, many classic BQA systems have been proposed, such as EPoCare [134], PICO-(P: patient/problem, I: intervention, C: comparison, O: outcome) and knowledge-extraction-based BQA systems [38–40], MedQA [220], Terol et al. [191], Weiming et al. [202], **Health on the Net QA (HONQA)** [34], AskHERMES [25] and so on. Such systems employ complex pipelines with numerous question, document, and answer processing modules, which is typically reflected by the IBM Watson system [51]. BioASQ [193] is a cornerstone challenge that has been running annually since 2013 for the evaluation of biomedical natural language understanding systems. A variety of BQA systems have been proposed in BioASQ, improving QA performance from approximately 20% top factoid mean reciprocal rank and list F-measure in BioASQ 1 to approximately 50% in BioASQ 8 [129]. Notably, the landscape of BioASQ participating models has been re-shaped by several NLP methodological revolutions: 1. The introduction of distributed word representations [115, 116]; 2. Deep learning-based QA models such as **Bi-Directional Attention Flow (Bi-DAF)** [171]; 3. Large-scale **Pre-trained Language Models (PLMs)** represented by **Embeddings for Language Models (ELMo)** [146] and **bidirectional encoder representations from transformers (BERT)** [42]. Currently, almost all top-performing BQA systems use the biomedical PLMs (e.g., BioBERT [98]), in their systems. Furthermore, various other BQA challenges and datasets have been introduced to further facilitate BQA research in different directions, e.g.: LiveQA-Med [1], MEDIQA [16, 167] for consumer health, emrQA [138] for clinical BQA, VQA-Rad [97], VQA-Med [4] and PathVQA [65] for visual BQA.

Despite the tremendous developments, BQA is still immature and faces several key challenges:

- **Dataset Scale, Annotation & Difficulty:** Most current BQA models utilize deep learning and are thus data-hungry. However, annotating large-scale biomedical corpora or knowledge bases is prohibitively expensive. As a result, current expert-annotated BQA datasets are small in size, with only hundreds to few thousand QA instances. To build large-scale datasets, many works have attempted to automatically collect BQA datasets, but their utility is limited and their annotation quality is not guaranteed. Furthermore, questions of most current BQA datasets do not require complex reasoning to answer.
- **Domain Knowledge Not Fully Utilized:** There are rich biomedical resources that encapsulate different types of domain knowledge, including large-scale corpora, various biomedical KBs, domain-specific NLP tools, and images. Unfortunately, most BQA models fail to utilize

Table 1. Characteristics of Different BQA Contents

Content	Main User	Question motivation	Answer style
Scientific (Section 2.1)	–	Learning cutting-edge scientific advances	Professional-level
Clinical (Section 2.2)	Professionals	Assisting clinical decision-making	Professional-level
Consumer health (Section 2.3)	General public	Seeking advice or knowledge	Consumer-understandable
Examination (Section 2.4)	–	Testing biomedical knowledge	Mostly choices

"–" denotes no specific users.

them effectively. As a result, biomedical domain knowledge is not fully utilized, which can be potentially solved by fusing different BQA approaches.

- **Lack of Explainability:** Since biomedicine is a highly specialized domain, ideal systems should not only return the exact answers (e.g.: "yes"/"no"), but also provide the explanations for giving such answers. However, there are only a few BQA systems that are explainable.
- **Evaluation Issues:** Qualitatively, current evaluations mainly focus on certain modules, e.g., **Machine Reading Comprehension (MRC)**, within a complete QA pipeline. Quantitatively, most evaluation metrics do not consider rich biomedically synonymous relationships.
- **Fairness and Bias:** Most machine-learning-based BQA systems learn from historical data, such as scientific literature and electronic medical records, which can be potentially biased and out of date. However, studies on BQA fairness and model transparency are quite sparse.

This article is organized as follows: We first describe the scope of this survey in Section 2; we then give an overview of the surveyed BQA approaches in Section 3. Various methods and datasets have been proposed for each BQA approach, and they are systematically discussed in Sections 4–8; To conclude, we summarize several challenges of BQA and discuss potential future directions in Section 9.

2 SURVEY SCOPE

Biomedicine is a broad domain that covers a range of biological and medical sciences. Since the term is often loosely used by the community, we specifically define several sub-domains of biomedicine, namely, scientific, clinical, consumer health, and examination, as the focus of this survey. Each content type is defined by the most distinctive characteristics of their corresponding users, questions, and expected answers, as shown in Table 1. It should be noted that the content types are not mutually exclusive, but most of our surveyed works belong to only one of them. In this section, we introduce these contents in Sections 2.1–2.4, and we also describe some related surveys with the focus of different scopes in Section 2.5. Several typical QA examples for each content type are shown in Table 2. The datasets are selected from hand literature search on PubMed and Google Scholar with keywords such as "biomedical," "biological," "medical," "clinical," "health," and "question answering." For each included dataset paper, we also checked their references and papers citing it. We describe mostly the methods with state-of-the-art performance on the surveyed datasets.

2.1 Scientific

Scientific QA addresses cutting-edge questions whose answers need to be extracted or inferred from scientific literature, e.g.: "Which cells express G protein-coupled receptors?" Most of the new findings in the biomedical field are published in the form of scientific literature, whose size is growing at an unprecedented pace; for example, MEDLINE,[1] a bibliographic database of life

[1] https://www.nlm.nih.gov/bsd/medline.html.

Table 2. Typical Question-answer Examples of Different Content Types

Type/Dataset	Question	Context	Answer
Scientific			
BioASQ	Is the protein Papilin secreted?	[...] secreted extracellular matrix proteins, mig-6/papilin [...]	Yes
Biomed-Cloze	Helicases are motor proteins that unwind double stranded _?_ into [...]	Defects in helicase function have been associated with [...]	nucleic acid
Clinical			
emrQA	Has the patient ever had an abnormal BMI?	08/31/96 [...] BMI: 33.4 Obese, high risk. Pulse: 60. resp. rate: 18	BMI: 33.4 Obese, high risk
CliCR	If steroids are used, great caution should be exercised on their gradual tapering to avoid _?_	[...] Thereafter, tapering of corticosteroids was initiated with no clinical relapse. [...]	relapse
Consumer			
MedQuAD	Who is at risk for Langerhans Cell Histiocytosis?	NA	Anything that increases your risk of [...]
MEDIQA-AnS	What is the consensus of medical doctors as to whether asthma can be cured? And do you have [...]	Asthma Overview Asthma is a chronic lung disease that causes episodes of wheezing [...]	Asthma is a chronic disease. This means that it can be treated but not cured. [...]
Examination			
HEAD-QA	The antibiotic treatment of choice for [...] is	1. Gentamicin; 2. Erythromycin; 3. Ciprofloxacin; 4. Cefotaxime	4. Cefotaxime

sciences, contains references to over 30M articles, and about 2.7K articles are added each day in 2019. Given the huge number of scientific publications, it is almost impossible to manually read all relevant studies and give comprehensive answers to scientific questions, so automatic answering of scientific questions is vital.

The BQA community's fight against COVID-19 is a great example of scientific QA. There has been a surge of COVID-19-related publications [102] that human experts find difficult to keep up with. Consequently, it is important to develop automatic methods for natural language understanding of them. To facilitate NLP studies on the COVID-19 literature, Wang et al. [199] released the CORD-19 corpus, which contains more than 280K papers about the novel coronavirus. Many BQA datasets have been introduced to help develop and evaluate models that answer COVID-19-related questions, e.g.: COVID-QA and COVIDQA datasets and the EPIC-QA challenge. Several resources and methods [148, 186, 231] have been introduced to tackle the COVID-19 QA by the QE approach (Section 8).

The most distinctive feature of scientific BQA is that large-scale corpora like PubMed and PubMed Central are freely available, which contain 4.5B and 13.5B tokens, respectively. In contrast, the entire English Wikipedia contains only 2.5B tokens. Besides, documents in PubMed and PubMed Central are semi-structured—they have sections of background, introduction, methods, conclusion, and so on, which can be potentially exploited in building domain-specific datasets. Consequently, the largest expert-annotated BQA dataset—BioASQ, and most large-scale (semi-)automtically constructed BQA datasets are all scientific BQA datasets (discussed in Section 9.1). Further exploiting the scale and structure of the scientific literature to design novel BQA tasks remains an interesting direction to explore.

2.2 Clinical

Clinical QA focuses on answering healthcare professionals' questions about medical decision-making for patients. Ely et al. [49] find the most frequent clinical questions are: 1. What is the drug of choice for condition x? (11%); 2. What is the cause of symptom x? (8%); 3. What test is indicated in situation x? (8%); 4. What is the dose of drug x? (7%); 5. How should I treat condition x (not limited to drug treatment)? (6%); 6. How should I manage condition x (not specifying diagnostic or therapeutic)? (5%); 7. What is the cause of physical finding x? (5%); 8. What is the

cause of test finding x? (5%); 9. Can drug x cause (adverse) finding y? (4%); 10. Could this patient have condition x? (4%).

Most of the clinical questions shown above are generic (cases 1–9) and largely non-specific to patients. In this case, clinical QA is similar to consumer health QA (Section 2.3). If the questions are specific to certain patients (e.g.: case 10), then their **Electronic Medical Records (EMRs)** should be provided. EMRs store all health-related data of each patient in both structured (i.e.: tables) and unstructured (i.e.: medical notes) formats. Due to the complexity and size of the EMR data, it is time-consuming and ineffective for the doctors to manually check the EMRs for clinical questions about the patient. Clinical QA systems can meet such information needs by quickly and accurately answering these questions. The difficulty of clinical BQA largely lies in the annotation of QA pairs, where considerable medical expertise and reasoning across clinical notes should be required to answer the questions [152]. For this, Pampari et al. [138] use expert-annotated templates (e.g.: "Has the patient ever been on {medication}?") with the existing i2b2 dataset annotations[2] (e.g.: "[...] Flagyl <medication> [...]") to build the first large-scale EMR BQA dataset emrQA. Yue et al. [225] analyze the emrQA dataset and find: 1. the answers are usually incomplete; 2. the questions are often answerable without using domain knowledge. Both are caused by the dataset collection approach of emrQA. Another large-scale clinical QA dataset, CliCR [187], is built by cloze generation (Section 9.1).

Roberts and Patra [161] show that the structured information of EMRs can be effectively queried by semantic parsing, where the goal is to map the natural language questions to their logic forms [86], e.g.: Q: "Was her ankle sprain healed?" Logic form: is_healed(latest(lambda(ankle sprain))). To tackle the clinical QA of structured EMR data, Soni et al. [179] annotate a dataset of 1K clinical questions with their logic forms. Some paraphrasing-based data augmentation methods are also introduced to improve the performance of semantic parsers of EMR questions [180, 181]. Wang et al. [200] propose TREQS, a two-stage generation model based on the sequence-to-sequence model and the attentive-copying mechanism and show its effectiveness on their MIMICSQL dataset for the question-to-SQL (table-based) task. Based on MIMICSQL dataset, Park et al. [143] propose a question-to-SPARQL (graph-based) dataset: MIMIC-SPARQL∗. TREQS also performs better on the graph-based dataset.

Radiology and pathology images play a vital role in the diagnosis and treatment of diseases. Clinical QA also contains VQA tasks, e.g.: VQA-Rad [97], VQA-Med [4], and PathVQA [65], which help doctors to analyze a large amount of images required for medical decision-making and population screening.

Ely et al. [48] also study the obstacles that prevent physicians from answering their clinical questions and find that doubting whether the answer exists is the most commonly (11%) reported reason for not pursuing the answers, and the most common obstacle in pursuing the answer is the failure to find the needed information in the selected resources (26%). Both problems can be solved by the clinical QA system. Currently, the main challenge for building such systems is the lack of large-scale expert-annotated datasets that reflect the real demands in the clinic. Apart from the high-price of deriving such annotations, there are also privacy and ethical issues for releasing them, especially when the datasets are based on EMRs. Future clinical QA datasets should have larger scales, less noise, and more diversity.

2.3 Consumer Health

Consumer health questions are typically raised by the general public on search engines, where online medical services provide people with great convenience, as they are not limited by time and

[2] https://www.i2b2.org/NLP/DataSets/.

space. As a result, rapidly increasing numbers of consumers are asking health-related questions on the Internet: According to one report released by the Pew Research Center [54], over one-third of American adults have searched online for medical conditions that they might have. Many try to find answers to their medical questions before going to a doctor or making decisions about whether to go to a doctor, and their information needs range from self-diagnosis to finding medications. It is vitally important to provide accurate answers for such questions, because consumers are unable to judge the quality of medical contents. Considering the contradiction between the great demands of consumers and the scarcity of medical experts, an automatic answering system is helpful for sharing medical resources to provide online medical service.

Some works [192, 228, 229] have exploited the doctors' answers to patients' questions on online medical consultation websites, e.g., XunYiWenYao,[3] to build large-scale consumer health QA datasets. These datasets are formatted as multi-choice BQA, where the task is to find the relevant or adopted answers. However, the quality of such datasets is questionable, since the answers are written by users from online communities and the forum data has intrinsic noise. Remarkable efforts have been made by NLM's **Consumer Health Information and Question Answering (CHIQA)** project.[4] CHIQA [41] is aimed at dealing with a vast number of consumer requests (over 90K per year) by automatically classifying the requests and answering their questions. It also provides various datasets to develop consumer health BQA methods, including question decomposition and type, named entity, and spelling error datasets.

For consumer health QA, understanding the questions of consumers is a vital but difficult step: Such questions might contain many spelling and grammar errors, non-standard medical terms, and multiple focuses [160, 228]. For example, Ben Abacha and Demner-Fushman [15] find that consumers often submit long and complex questions that lead to substantial false positives in answer retrieval. To tackle it, they introduce the MeQSum corpus,[5] which contains 1K summarized consumer health questions and achieves the best 44.16% ROUGE-1 score using pointer-generator network [170] with semantic augmentation from question datasets. However, most consumers have no biomedical domain knowledge, so the returned answers should be not only accurate but also explainable (Section 9.5), posing further challenges for consumer health QA.

2.4 Examination

Many general domain QA datasets that are extracted from examinations have been introduced [89, 95, 145, 176]. Similarly, Examination BQA that addresses automatic answering of medical examination questions has also been explored. For example, in many countries, medical licensure requires the passing of specific examinations, e.g.: USMLE[6] in the U.S. Test items in examinations often take the form of multi-choice questions, and answering them requires comprehensive biomedical knowledge. Several datasets have been released that exploit such naturally existing QA data, e.g.: HEAD-QA [196] and NLPEC [101]. Usually, no contexts are provided for such questions and automatic answering of them requires the systems to find supporting materials (e.g.: texts, images, and KBs) as well as reason over them. However, the real utility of examination QA is still questionable.

2.5 Related Surveys

Athenikos and Han [10] systematically review BQA systems, mainly classic ones published before 2010. Content-wise, they classify BQA into biological QA and medical QA, which roughly

[3] http://xywy.com/.
[4] https://lhncbc.nlm.nih.gov/project/consumer-health-question-answering.
[5] https://github.com/abachaa/MeQSum.
[6] https://www.usmle.org/.

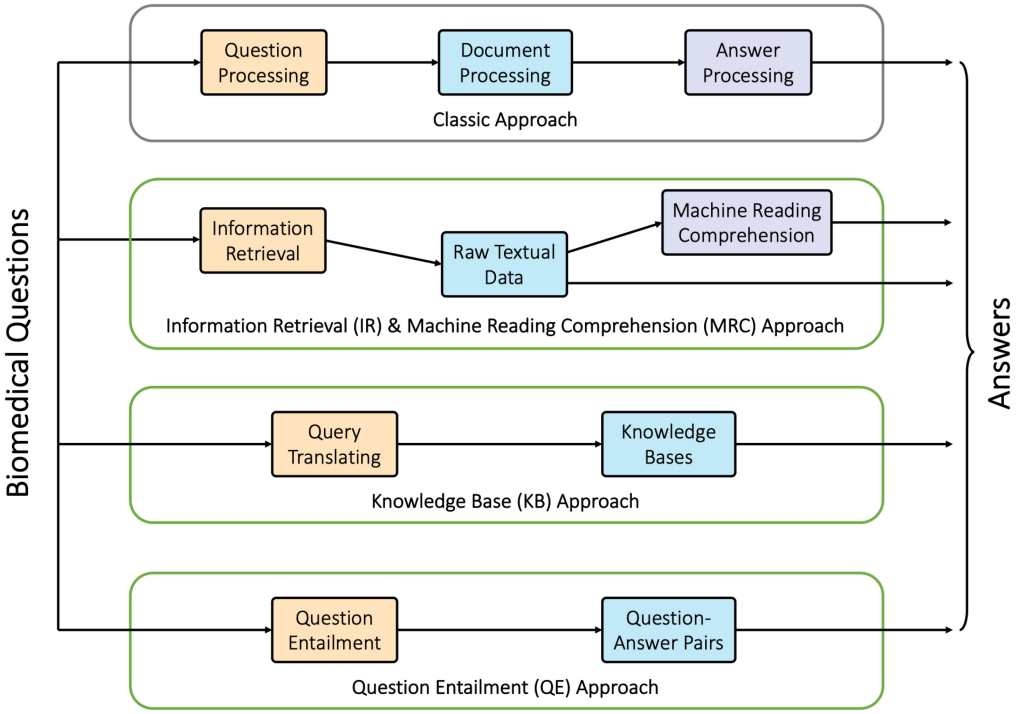

Fig. 1. Overview of major biomedical question answering approaches. Boxes indicate methods or resources.

correspond to our scientific and clinical content types, respectively. Bauer and Berleant [13] and Sharma et al. [173] briefly compare several classic BQA systems. Neves and Leser [130] present a detailed survey of QA for biology, which we classify as scientific BQA in this article. This survey also discusses various biological BQA systems. Recently, Nguyen [132] identified several challenges in consumer health BQA, including the lack of publicly available datasets, term ambiguity, incontinuous answer spans, and the lack of BQA systems for patients. Nguyen [132] propose a research plan to build a BQA system for consumer self-diagnosis. Kaddari et al. [84] present a brief survey that discusses several scientific BQA datasets and methods.

3 BQA APPROACH OVERVIEW

The fundamental task of BQA is to answer a given questions about biomedicine. In this survey, each BQA approach denotes a distinctive means to tackle the task. We briefly define different BQA approaches in this section, and a high-level overview of them is shown in Figure 1.

We first define the **Classic BQA approach** from a historical perspective. Mainly due to the lack of modular QA datasets (e.g., MRC BQA datasets), systems of this approach typically: 1. contain many sub-tasks and follow the pipeline of the question, document, and answer processing, similar to IBM's Watson system [51]; 2. use many rule-based and/or sparse-feature-based machine learning modules in one system; 3. are evaluated on small-scale private datasets. Since most of the classic BQA systems are surveyed in detail by Athenikos and Han [10], we just briefly introduce them in Section 4.

Table 3. Characteristics of Different BQA Approaches

BQA Approach	Supporting resources	Answer form
IR BQA approach (Section 5)	Document collections	Specific documents
MRC BQA approach (Section 6)	Specific documents (contexts)	Y/N; Extraction; Generation
KB BQA approach (Section 7)	Knowledge bases	Biomedical entities/relations
QE BQA approach (Section 8)	Answered questions (FAQs)	Existing answers of similar questions

Y/N: Yes/No.

Besides the classic BQA approach, other BQA approaches tackle the task using specific supporting resources that are included in standard, public datasets. They typically contains a collection of datasets and methods, and we define several BQA approaches below:

- **Information Retrieval (IR) approach:** where systems retrieve relevant documents to answer the questions;
- **Machine Reading Comprehension (MRC) approach:** where systems read given contexts about the questions to predict the answers. The contexts of MRC approach can be provided by the IR approach;
- **Knowledge Base (KB) approach:** where systems either explicitly translate the input questions to RDF queries to search the KBs or implicitly use the integrated knowledge from certain biomedical KBs to get the answers;
- **Question Entailment (QE) approach:** where systems find similar questions that have been previously answered in a Q-A pair database and re-use their answers to answer the given question;

Characteristics of these approaches are summarized in Table 3.

4 CLASSIC BQA

In this section, We briefly introduce several representative classic BQA systems and point the readers to the BQA survey by Athenikos and Han [10] for more details.

Traditionally, QA approaches consist of three main parts [72]: 1. Question processing, where systems (a) determine the type of the question and the corresponding type of the expected answer and (b) form queries that are fed to certain document retrieval systems; 2. Document processing, where systems (a) retrieve relevant documents from the queries generated in the previous step and (b) extract answer candidates from the relevant documents; 3. Answer processing, where systems rank the candidate answers based on certain criteria. Although recently some of these modules have become independent QA approaches (e.g., IR, MRC), classic BQA still remains a distinctive class of approach in this survey, because most of these works describe a whole system that includes all these subtasks. We show an overview of classic BQA methods in Table 4 with their specific methods for question, document, and answer processing.

PICO-based: Niu et al. [134] explore PICO-based BQA in the EPoCare project using simple keyword-based retrievers. Demner-Fushman and Lin further study the PICO-based semantic BQA for the practice of **Evidence-Based Medicine (EBM)** in a series of works [38–40], where the core step involves searching PubMed articles that are annotated with extracted medical knowledge. Huang et al. [74] study the feasibility of using the PICO format to represent clinical questions and conclude that PICO is primarily focused on therapy-type clinical questions and unsuitable for representing the others (e.g.: prognosis, etiology).

Natural-language-based: To tackle a broader range of topics, most other classic BQA systems accept natural language questions: The medical definitional question answering system (MedQA,

Table 4. Overview of Classic BQA Systems (Listed in Chronological Order)

System	Content	Question Processing	Document Processing	Answer Processing
EPoCare [134]	Clinical	PICO format	Keyword-based retriever	–
Takahashi et al. [188]	Scientific	Question type classification; Query formulation	MySQL retriever	–
Demner-Fushman and Lin [38–40]	Clinical	PICO-format; Query formulation	Knowledge extraction; Semantic matching	Semantic clustering & summarization
MedQA [99, 220]	Consumer	Question type classification	Lucene retriever	Answer extraction & summarization
BioSquash [175]	Scientific	Semantic annotation	Semantic annotation; Graph construction	Sentence selection, clustering and post-processing
Terol et al. [191]	Clinical	Question and answer type classification; Logic form extraction	–	Answer generation based on logic forms
Weiming et al. [202]	Clinical	Concept and relation recognition	Lucene retriever	Semantic interpretation and clustering based on relations
HONQA [34]	Consumer	Question, expected answer and medical type classification	–	–
Lin et al. [106]	Scientific	Question type classification; Query expansion	Google-interfacing retriever	NER- & SRL-based
EAGLi [57]	Scientific	Query and target categorization	PubMed e-utils and EasyIR [9]	Extracting concepts
AskHERMES [25]	Clinical	Topic classification with MetaMap [219]	BM25 retriever; Longest common subsequence extractor	Content Clustering
MiPACQ [23]	Clinical	Semantic annotation	Lucene retriever	ML-based re-ranking

"–": no special processing steps.

Lee et al. [99], Yu et al. [220]) is the first fully implemented BQA system that generates answers by extractive summarization for users' definitional questions from large text corpora. BioSquash [175] is adapted from the general domain summarizer Squash [114] and is focused on QA-oriented summarization of biomedical documents. Terol et al. [191] utilize logic forms for BQA, where the core step is to derive the logic forms of questions and utilize them to generate answers. HONQA [34] is a French/English bilingual BQA system that focuses on the supervised classification of question and answer types for question answering. Lin et al. [106] explore answering biomolecular event questions with named entities using syntactic and semantic feature matching. Gobeill et al. [57] generate 200 questions from biomedical relational databases to evaluate their EAGLi platform. Cao et al. [25] propose the askHERMES system, a BQA system that performs several semantic analyses, including question topic classification and content clustering, to provide extractive summaries for clinical questions.

The classic BQA approaches rely heavily on rule-based models and various ad hoc modules in their complex pipelines. Although these might be necessary in industry-level applications, they are hard to develop and maintain in academic settings. In addition, most classic BQA systems have not been validated on large-scale public datasets. With the introduction of various BQA datasets that are focused on specific BQA topics or steps, only a few BQA systems that tackle the complete question-to-answer BQA task have been proposed recently, which will be discussed as the modular evaluation issue in Section 9.6.

5 INFORMATION RETRIEVAL BQA

Information Retrieval (IR) BQA denotes the approach that uses *IR BQA Methods* to retrieve relevant text snippets from certain *Document Collections* for the given question, where the retrieved snippets can be either directly used as answers or further fed to MRC models (Section 6). We also discuss several *IR BQA Datasets* that are summarized in Table 5.

Table 5. Overview of the Information Retrieval Biomedical Question Answering Datasets (Listed in Alphabetical Order)

Dataset	Size	Metric	State-of-the-art (%)	Content	Format
BioASQ Task B Phase A [193]	3.7K	MAP	33.04 (document)/68.21 (snippet)[1]	Scientific	Retrieval
BiQA [96]	7.4K	MAP	–	Consumer	Retrieval
EPIC-QA[2]	45	MAP	–	Sci. & Con.	Retrieval
HealthQA [239]	7.5K	MRR	87.88 [239]	Consumer	Retrieval
TREC Genomics [68, 69]	28 (06), 36 (07)	MAP	54.39 (06)/32.86 (07) [70]	Scientific	Retrieval

[1]Batch 2 of BioASQ Task 8 b Phase A. [2]https://bionlp.nlm.nih.gov/epic_qa/.

5.1 Document Collections

PubMed[7] and PubMed Central[8] are the most widely used corpora. Both were developed and are maintained by the **National Library of Medicine (NLM)** of the U.S. PubMed provides free access to more than 30M citations for biomedical literature, where each citation mainly contains the paper title, author information, abstract, and semantic indices like MeSH (introduced in Section 7.1). **PubMed Central (PMC)** includes full-texts of over 6M biomedical articles in addition to the information provided in the PubMed citations.

More specific corpora are typically used for sub-domain BQA to filter out potential noise in larger corpora, e.g.: CORD-19 [199] for EPIC-QA Task A and Alzheimer's Disease Literature Corpus for QA4MRE-Alzheimer [125].

5.2 IR BQA Datasets

BioASQ Task B Phase A: BioASQ Task B is named "Biomedical Semantic Question Answering" [193] and contains two phases correspond to the IR and MRC BQA approaches in our BQA classification: In the phase A (IR phase), systems retrieve relevant documents for the given question; in the phase B (MRC phase), systems use gold standard relevant documents and ontological data to answer the given questions (discussed in Section 6). Specifically, for the given question, BioASQ phase A participants shall return the relevant: 1. Concepts from certain ontologies such as MeSH; 2. Articles from PubMed and text snippets within the articles; 3. RDF triples from the Linked Life Data.[9] **Mean average precision (MAP)** is used as a main metric for the BioASQ retrieval phase, with slight modifications for snippet retrieval.

Lamurias et al. [96] introduces ***BiQA***, an IR BQA dataset containing 7.4K questions and 14.2K relevant PubMed articles collected from online QA forums (Stack Exchange[10] and Reddit[11]). They show that adding BiQA in the training set can marginally boost BioASQ Task B phase A test performance.

The **Epidemic Question Answering** (***EPIC-QA***) challenges[12] are also formatted as IR BQA, where the participants return a ranked list of sentences from expert and consumer corpora to answer questions about the COVID-19 pandemic.

HealthQA has 7.5K manually annotated questions that can be answered by one of the 7.3K web articles [239].

The ***TREC Genomics Tracks*** in 2006 and 2007 tackle the BQA with IR approach [68–70]: 28 and 36 topic questions (e.g.: "What [GENES] are genetically linked to alcoholism?") are released in

[7]https://pubmed.ncbi.nlm.nih.gov/.
[8]https://www.ncbi.nlm.nih.gov/pmc/.
[9]http://linkedlifedata.com/.
[10]https://stackexchange.com/.
[11]https://www.reddit.com/.
[12]https://bionlp.nlm.nih.gov/epic_qa/.

2006 and 2007, respectively, and the participating systems are required to retrieve passages from 162K full-text articles from Highwire Press.[13]

IR BQA systems typically return a ranked list of documents as answers, and MAP is usually used as the evaluation metric:

$$\text{AP} = \frac{1}{m}\sum_{k=1}^{n} P@k \times \text{rel}(k), \qquad \text{MAP} = \frac{1}{N}\sum_{q=1}^{N} \text{AP}_q,$$

where m denotes the number of gold-standard relevant documents and n denotes the number of the returned documents. rel(k) is an indicator function that has the value 1 if the kth document is relevant otherwise 0.

5.3 IR BQA Methods

BioASQ Task B Phase A: Wishart [107] re-ranks and combines sentences from the retrieved documents to form the ideal answers for BioASQ task B phase B and generates exact answers from the ideal answers according to the question type. The USTB team [82] wins all batches in document, snippet, and concept retrieval in BioASQ 5. They use sequential dependence model [20], pseudo relevance feedback, fielded sequential dependence model [235], and divergence from randomness model [33]. The AUEB team proposes a series of models [22, 140, 141] that win most of the Task B Phase A challenges since BioASQ 6. At BioASQ 6, they [22] use the Position-Aware Convolutional Recurrent Relevance model [75] and the Deep Relevance Matching Model [59] for document retrieval and use the Basic Bi-CNN model [217] for snippet retrieval. They win 3/5 and 5/5 batches for retrieving documents and snippets in BioASQ 6, respectively. At BioASQ 7, they [140] combine the document and snippet retrieval system by modifying their BioASQ 6 system to also output the sentence-level (i.e.: snippet) relevance score in each document. They win 4/5 and 4/5 batches for retrieving documents and snippets in BioASQ 7, respectively. In BioASQ 8, they [141] continue to use this system and win 2/5 for document and 4/5 batches for snippet retrieval.

HealthQA: Zhu et al. [239] propose Hierarchical Attention Retrieval, a ranking model for biomedical QA that uses a deep attention mechanism at word, sentence, and document levels to compute a relevance score of the given query with each candidate document. With the proposed model, they achieve an MRR of 0.8788 on the HealthQA dataset.

Zhou and Yu [237] win the ***TREC 2006 Genomics Track***, where they first identify query concepts and retrieve relevant documents with concept-level and word-level similarity measurements, and then extract the answers. At ***TREC 2007 Genomics Track***, NLMinter [37] achieves the best performance. NLMinter is an interactive retriever where the fusion retrieval results are boosted by the document relevance feedback, which is determined by expert PubMed search and occasional examination of the abstracts.

5.4 Comments

Though Classic BQA approaches usually contain a retrieval step, IR BQA is still considered as a distinct approach, because the retrieved documents are directly used as answers and are evaluated by IR metrics. Traditional retrieval methods like TF-IDF have been well-studied and ubiquitously used in IR BQA approach. Future studies can focus more on PLM-based (re-)ranking methods [27, 104] and how to better bridge the IR and MRC models.

[13] https://www.highwirepress.com/.

Table 6. Types of Questions in BioASQ and Respective Examples

Type	Example Question	Example Context	Exact answer	Ideal answer
Yes/No	Is the protein Papilin secreted?	[...] and two genes encoding secreted extracellular matrix proteins, mig-6/papilin [...].	Yes	Yes, papilin is a secreted protein
Factoid	Name synonym of Acrokeratosis paraneoplastica.	Acrokeratosis paraneoplastic (Bazex syndrome) is a rare, but [...]	Bazex syndrome	Acrokeratosis paraneoplastic (Bazex syndrome) is a rare [...]
List	List Hemolytic Uremic Syndrome Triad.	Atypical hemolytic uremic syndrome (aHUS) is a rare disease characterized by the triad of [...]	anaemia, renal failure, thrombocytopenia	Hemolytic uremic syndrome (HUS) is a clinical syndrome characterized by [...]
Summary	What is the effect of TRH on myocardial contractility?	Thyrotropin-releasing hormone (TRH) improved [...]	NA	TRH improves myocardial contractility

6 MACHINE READING COMPREHENSION BQA

Machine Reading Comprehension (MRC) is a well-studied BQA task, where the models answer questions about given textual contexts. *MRC BQA Datasets* are typically specialized in content and have predetermined answer format, so most *MRC BQA Methods* developed on them are end-to-end neural models.

6.1 MRC BQA Datasets

Many MRC BQA datasets have been proposed, and we show an overview of them in Table 7.

BioASQ Task B Phase B: It provides the largest and most widely used manually annotated MRC BQA dataset: Starting from 2013, BioASQ annotates about 500 test QA instances each year, which will be included in the training set of the following years. Currently, BioASQ 2020 consists of 3,243 training QA instances and at least 500 test instances. Questions in BioASQ are typically scientific questions (Section 2.1). There are four types of QA instances in BioASQ: factoid, list, yes/no, and summary. Factoid, list, and yes/no instances have both *exact* and *ideal* answers: *Exact* answers are short answers that directly answer the questions, e.g., single and multiple biomedical entities for factoid and list questions, respectively; "yes" or "no" for yes/no questions. *Ideal* answers are *exact* answers written in complete sentences, e.g.: "Yes, because [...]." The main evaluation metrics for yes/no, factoid, list, and summary questions are accuracy, MRR, mean F-score, and manual score, respectively. We show several examples of BioASQ instances in Table 6.

Question Answering for Machine Reading Evaluation (***QA4MRE***) holds a sub-task on machine reading of biomedical texts about Alzheimer's disease [125]. This task provides only a test dataset with 40 QA instances, and each instance contains one question, one context and five answer choices.

Cloze-style questions require the systems to predict the missing spans in contexts (e.g.: Q: "Protein X suppresses immune systems by inducing _____ of immune cells."; A: "apoptosis"). There are many large-scale cloze-style MRC BQA datasets that are automatically constructed, such as CliCR, Biomed-Cloze, BioRead, BMKC, and BioMRC.

COVIDQA [190] is a QA dataset specifically designed for COVID-19. It has 124 question-article pairs translated from the literature review page of Kaggle's COVID-19 Open Research Dataset Challenge,[14] where relevant information for each category or subcategory in the review is presented.

[14] https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/tasks.

Table 7. Overview of the Machine Reading Comprehension Biomedical Question Answering Datasets (Listed in Alphabetical Order)

Dataset	Size	Metric	State-of-the-art (%)	Content	Format
BioASQ Task B Phase B [193]	3.7K	F1, MRR, List F1, Manual	90.3, 39.7, 52.3, 4.39/5 [129]	Scientific	Y/N; Extraction; Generation
Biomed-Cloze [43]	1M	–	–	Scientific	Extraction
BioMRC [142]	812K	Acc	80.06 (dev)/79.97 (test)[1] [142]	Scientific	Extraction
BioRead [139]	16.4M	Acc	47.06 (dev)/51.52 (test) [139]	Scientific	Extraction
BMKC [91]	473K (T); 370K (LS)	Acc	T: 85.5 (val)/83.6 (test); LS: 80.1 (val)/77.3 (test) [91]	Scientific	Extraction
CliCR [187]	100K	EM, F1	55.2, 59.8 [147]	Clinical	Extraction
COVIDQA [190]	124	P@1, R@3, MRR	30.6 [28], 47.7 [185], 41.5 [190]	Scientific	Extraction
COVID-QA [124]	2K	EM, F1	37.2, 64.7 [157]	Scientific	Extraction
EBMSummariser [122]	456	ROUGE-1,2,SU4	39.85, 24.50, 22.59 [172]	Clinical	Generation
emrQA [138]	455K	EM, F1	76.1, 81.7 [163]	Clinical	Extraction
MASH-QA [238]	34.8K	EM, F1	29.49, 64.94 [238]	Consumer	Extraction
MEDHOP [205]	2.5K	Acc	63.2 [76]	Scientific	Multi-choice
MEDIQA-AnS [167]	552 (single); 156 (multi)	ROUGE-1,2,L; BLEU	Extractive[2]: 29, 15, 12, 9; Abstractive: 32, 12, 8, 9 [167]	Consumer	Generation
MEDIQA-QA [16]	3K	Acc, P, MRR	79.49 [66], 84.02 [66], 96.22 [149]	Consumer	Multi-choice
ProcessBank [17]	585	Acc	66.7 [17]	Scientific	Multi-choice
PubMedQA [81]	212K	Acc, F1	68.08, 52.72 [81]	Scientific	Y/N
QA4MRE-Alz [125]	40	$c@1$	76 [19]	Scientific	Multi-choice

[1]Results on BioMRC lite. [2]Results on single document summarization.

COIVD-QA [124] is another COVID-19 QA dataset with 2K question-answer pairs annotated by biomedical experts. The annotation is similar to that of SQuAD, while the answers in COVID-QA tend to be longer, as they generally come from longer texts.

Molla and Santiago-Martinez [121], Mollá et al. [122] build the ***EBMSummariser***, a summarization dataset of 456 instances for EBM, from the Clinical Inquiries section of the Journal of Family Practice[15]: Each instance contains a clinical question, a long "bottom-line" answer, the answer's evidence quality, and a short justification of the long answer.

MASH-QA [238] is a dataset based on consumer health domain and is designed for extracting information from texts that span across a long document. It utilizes long and comprehensive healthcare articles as context to answer generally non-factoid questions. Different from the existing MRC datasets with short single-span answers, many answers in MASH-QA are several sentences long and excerpted from multiple parts or spans of the long context.

MEDHOP [205] is a multi-hop MRC BQA dataset, where each instance contains a query of a subject and a predicate (e.g.: "Leuprolide, interacts_with, ?"), multiple relevant and linking documents, and a set of answer options extracted from the documents. Reasoning over multiple documents is required for the model to answer the question.

MEDIQA-QA [16] is the dataset of the QA subtask of MEDIQA 2019 shared task that has 400 questions and 3K associate answers. It is obtained by submitting medical questions to the consumer health QA system CHiQA, then re-ranks the answers by medical experts. The task of MEDIQA-QA dataset is to filter and improve the ranking of answers, making it a multi-choice QA task. ***MEDIQA-AnS*** [167], however, is a summarization dataset. It provides extractive and abstractive versions of single and multi-document summary of the answer passages from MEDIQA-QA.

ProcessBank [17] contains multi-choice questions along with relevant biological paragraphs. The paragraphs are annotated with "process," a directed graph $(\mathscr{T}, \mathscr{A}, \mathscr{E}_{tt}, \mathscr{E}_{ta})$, where nodes $\mathscr{T}$

[15]https://www.mdedge.com/familymedicine/clinical-inquiries.

are token spans denoting the occurrence of events, nodes $\mathscr{A}$ are token spans denoting entities in the process, and the latter two are edges describing event relations and semantic roles, respectively.

Jin et al. [81] build the ***PubMedQA*** dataset from PubMed articles that use binary questions as titles (e.g.: "Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?") and have structured abstracts. The conclusive parts of the abstracts are the long answers, while the main task of PubMedQA is to predict their short forms, i.e., yes/no/maybe, using the abstracts without the conclusive parts as contexts.

6.2 MRC BQA Methods

In this section, we first introduce the top-performing systems in each year of the BioASQ challenge to reflect the landscape changes of MRC BQA methods. We then briefly describe SoTA models of other surveyed MRC BQA datasets.

BioASQ: The first two BioASQ challenges [12, 144] use a Watson-motivated baseline [203] that ensembles multiple scoring functions to rank the relevant concepts with type coercion to answer the given questions.

The Fudan system [233] of BioASQ 3B contains three major components: 1. A question analysis module that mainly extracts semantic answer types of questions; 2. Candidates generating by PubTator [201] and Stanford POS tools[16]; 3. Candidates ranking based on word frequency. The SNU team [32] directly combines the retrieved relevant passages to generate the ideal answer and achieve state-of-the-art performance. At BioASQ 4B: The HPI team [169] proposes an algorithm based on LexRank [50] to generate ideal answers, which only uses biomedical named entities in the similarity function. They win 1/5 batch in the ideal answer generation. At BioASQ 5B: The UNCC team [18] uses lexical chaining-based extractive summarization to achieve the highest ROUGE scores for ideal answer generation, with 0.7197 ROUGE-2 and 0.7141 ROUGE-SU4.

The CMU OAQA team describes a series of works for BioASQ [29, 213, 216]. At BioASQ 3B, they propose a three-layered architecture [213] where: The first layer contains domain-independent QA components such as input/output definition, intermediate data objects; the second layer has implementations of biomedical domain-specific materials such as UMLS and MetaMap [8]; the third design layer is BioASQ-specific, including the end-to-end training and testing pipeline for the task. The core components of the pipeline are the answer type prediction module and the candidate answer scoring module based on supervised learning. At BioASQ 4B, they extend their BioASQ 3B system with general-purpose NLP annotators, machine-learning-based search result scoring, collective answer re-ranking and yes/no answer prediction [216]. The CMU OAQA team is focused on ideal answer generation [29] at BioASQ 5B, using extractive summarization tools such as Maximal Marginal Relevance [26] and Sentence Compression [52] with biomedical ontologies such as UMLS and SNOMED-CT. BioAMA [174] further improves the ROUGE score by 7% for ideal answer generation than Chandu et al. [29] by combining effective IR-based techniques and diversification of relevant snippets.

The Macquarie University team has participated in BioASQ 5–8 with the focus of ideal answer generation by extractive summarization for yes/no, factoid, list, and summary questions [117–120]. At BioASQ 5B, Mollá [117] observes that a trivial baseline that returns the top retrieved snippets as the ideal answer is hard to beat. At BioASQ 6B, Mollá [118] shows that using LSTM-based deep learning methods that predict the F1 ROUGE-SU4 score of an individual sentence and the ideal answer achieves the best results. At BioASQ 7B, Mollá and Jones [119] observe that sentence-level classification task works better than regression task for finding the extractive summary sentences.

[16] https://nlp.stanford.edu/software/tagger.shtml.

In recent years of BioASQ, transfer learning has gained increasing attention, where models are first pre-trained on large-scale general domain QA datasets or BQA datasets and then fine-tuned on the BioASQ training set. Wiese et al. [207] achieve state-of-the-art performance on factoid questions and competitive performance on list questions by transferring the FastQA model pre-trained by SQuAD to BioASQ. Dhingra et al. [43] show significant performance improvement over purely supervised learning by pre-training the GA-Reader [44] on an automatically generated large-scale cloze BQA dataset (Section 9.1) and then fine-tuning it on BioASQ. Du et al. [46, 47] have similar observations with transfer learning from the SQuAD dataset. Kang [88] shows transfer learning from NLI datasets also benefits BioASQ performance on yes/no (+ 5.59%), factoid (+ 0.53%), and list (+13.58%) questions. Generally, two main components are ubiquitously used in top-performing systems of the current BioASQ 8 challenge [129]: 1. domain-specific pre-trained language models [218], such as BioBERT; 2. task-specific QA datasets that can (further) pre-train the used models, such as SQuAD for extractive QA and PubMedQA for yes/no QA.

In summary, with the introduction of large-scale MRC datasets like SQuAD [154, 155], a variety of neural MRC models have been proposed that incrementally improve the task performance, such as DCN [211], Bi-DAF [171], FastQA [204]. Contextualized word embeddings pre-trained by **language models (LM)** such as ELMo [146] and BERT [42] show significant improvements on various NLP tasks including MRC. Pre-trained LMs on biomedical corpora, such as BioELMo [80], BioBERT [98], SciBERT [14], clinical BERT [7, 73], and PubMedBERT [58], further improve their in-domain performance. Probing experiments and analyses by Jin et al. [80] indicate that better encoding of biomedical entity-type and relational information leads to the superiority of domain-specific pre-trained embeddings.

Various methods have also been developed for other MRC BQA datasets. Here, we briefly discuss their representative SoTA methods as shown in Table 7.

BioRead: Pappas et al. [139] train the AOA Reader [35] on BioReadLite, which computes the mutual information between query and context and places another attention layer over the document-level attention to achieve attended attention for the final prediction. They achieve the best accuracy of 0.5152. ***BioMRC:*** It is the updated version of BioRead. Pappas et al. [142] use SciBERT [14] and maximize scores of all mentions of each entity in the passage, achieving SoTA accuracy of 0.7997. ***BMKC:*** Based on Attention Sum Reader architecture [85], Kim et al. [91] present a new model that combines pre-trained knowledge and information of entity types. They also develop an ensemble method to integrate results from multiple independent models, which gets the accuracy of 0.836 on BMKC_T and 0.773 on BMKC_LS.

CliCR: Pham et al. [147] show that language models have better performance with systematic modification on cloze-type datasets. They replace *@placeholder* with [MASK] and train BioBERT [98] on the modified dataset to obtain the SoTA EM of 0.552 and F1-score of 0.598.

COVIDQA: Chakravarti et al. [28] fine-tune pre-trained language models on the Natural Questions dataset [94] with attention-over-attention strategy and attention density layer. They try its zero-shot transfer and achieve P@1 of 0.306. Su et al. [185] combine HLTC-MRQA [184] with BioBERT to rank context sentences to get the evidence and obtain R@3 of 0.477. Tang et al. [190] achieve MRR of 0.415 by fine-tuning T5 [151] on MS MARCO [135]. ***COVID-QA:*** Reddy et al. [157] propose an example generation model for the training of MRC and fine-tune RoBERTa-large [109] on SQuAD2.0 [154], NQ, and their generated training examples, which achieves EM of 0.372 and F1-score of 0.647.

EBMSummariser: Sarker et al. [166] extract three sentences using hand-crafted features such as sentence length, position, and question semantics for the EBMSummariser dataset, achieving ROUGE-L F-score of 0.168. ShafieiBavani et al. [172] utilize both UMLS and WordNet to summarize

medical evidence for queries and achieve ROUGE-1 of 0.3985, ROUGE-2 of 0.2450, and ROUGE-SU4 of 0.2259 on EBMSummariser.

emrQA: Rongali et al. [163] use rehearsal and elastic weight consolidation to improve domain-specific training, which can benefit the performance of models in both general domain and domain-specific tasks. They achieve EM of 0.761 and F1-score of 0.817.

MASH-QA: Zhu et al. [238] propose MultiCo to select sentences across the long contexts to form answers. MultiCo combines a query-based sentence selection approach with an inter-sentence attention mechanism and achieves EM of 0.2949 and F1-score of 0.6494 on single-span MASH-QA dataset.

MEDHOP: Huo and Zhao [76] propose a Sentence-based Circular Reasoning approach that establishes a information path with sentence representation. They also implement a nested mechanism to systematically represent semantics, which improves the model performance significantly and achieves an accuracy of 0.632.

MEDIQA-AnS: Savery et al. [167] train BART [100] on the BioASQ data to achieve SOTA results.
MEDIQA-QA: He et al. [66] infuse disease knowledge into pre-trained language models like BERT and achieve accuracy of 0.7949 and precision of 0.8402. Pugaliya et al. [149] train their end-to-end system in a multi-task setting and use the pretrained RQE and NLU modules to extract the best entailed questions and best candidate answers. They achieve MRR of 0.9622.

ProcessBank: Berant et al. [17] first predict a structure representing the process in the given paragraph, then they map each question into queries and compare them with the predicted structure. They achieve the accuracy of 0.667.

PubMedQA: Jin et al. [81] take the multi-phase fine-tuning schedule with long answer as additional supervision and achieve accuracy of 0.6808 and F1-score of 0.5272.

6.3 Comments

BioASQ is still the well-recognized benchmark and the "go-to" dataset for MRC BQA because of its careful design, expert annotations, large size, and highly active community. Future models could explore developing pre-training methods that utilize richer biomedical knowledge than the raw texts (Section 9.4). Additionally, collecting harder datasets/datasets that require other types of reasoning still remains an interesting future direction (Section 9.2).

7 KNOWLEDGE-BASE BQA

KBQA (Knowledge-Base QA) refers to answering questions using entities or relation information from knowledge bases [55]. In biomedical domain, various large-scale biomedical KBs have been introduced, and one of their objectives is to assist with BQA. Typically, one can convert natural language questions to SPARQL queries[17] and use them to search the KBs for the answers. In this section, we first introduce the *existing knowledge bases* that have been used for KB BQA, and then introduce the *KB BQA datasets* and *KB BQA methods* developed on them.

7.1 Existing Knowledge Bases

We define biomedical KBs as databases that describe biomedical entities and their relations, which can usually be stored by subject-predicate-object triples. Biomedical KBs can be used for enhancing text representations [80, 223, 224] and improving performances for BQA [101] (not only KB BQA). Substantial efforts have been made towards building biomedical KBs, including ontologies such as **Medical Subject Headings (MeSH)**[18] for biomedical text topics, **International**

[17] https://www.w3.org/TR/rdf-sparql-query/.
[18] https://www.ncbi.nlm.nih.gov/mesh/.

Table 8. Overview of KB BQA Datasets (Listed in Alphabetical Order)

Dataset	Size	Metric	State-of-the-art (%)	Content	Format
Bioinformatics [177]	30	F1	60.0 [177]	Scientific	Generation
QALD-4 task 2 [194]	50	F1	99.0 [111]	Consumer	Generation

Classification of Diseases (ICD)[19] for diseases, and **Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT**, Stearns et al. [183]**)** for medical terms. The **Unified Medical Language System (UMLS)**[20] is a metathesaurus that integrates nearly 200 different biomedical KBs such as MeSH and ICD. Biomedical KB is a big topic, and we refer the interested readers to References [87, 133].

7.2 KB BQA Datasets

KB BQA datasets provide a list of biomedical questions and several biomedical KBs. One should generate a SPARQL query for each question, and the answers are evaluated by query results. We summarize existing KB BQA datasets and show an overview of them in Table 8.

QALD-4 task 2: Unger et al. [194] provide 50 natural language biomedical questions and request SPARQL queries to retrieve answers from SIDER,[21] Drugbank,[22] and Diseasome,[23] where most questions require integrating knowledge from multiple databases to answer. An example natural language question is, "Which genes are associated with breast cancer?," and a possible query can be:

```
SELECT DISTINCT ?x
WHERE {
 diseases:1669 diseasome:associatedGene ?x .
}
```

Bioinformatics contains 30 biomedical queries with different complexity, and the database searching are restricted in Bgee[24] and OMA.[25] The natural language questions include multiple concepts that lead to longer and more complicated SPARQL queries.

7.3 KB BQA Methods

In this section, we introduce the well-performing KB BQA methods applied on ***QALD-4 task 2*** and ***Bioinformatics***.

QALD-4 task 2: Marginean [111] wins QALD-4 task 2 by introducing the GFMed that is built with Grammatical Framework [156] and Description Logic constructors, achieving 0.99 F1 on the test set. GFMed performs extraordinary well in QALD-4 task 2, since it is highly customized to this dataset. CANaLI [113] designs an semantic automaton to parse the questions in specified form and achieves F1 of 0.92 on QALD-4 task 2. Questions not in specified form are ignored by CANaLI. Zhang et al. [232] exploit KBs to find out candidate relation, type, entity + relation triple

[19] https://www.who.int/classifications/icd/.
[20] https://www.nlm.nih.gov/research/umls.
[21] http://sideeffects.embl.de/.
[22] https://www.drugbank.com/.
[23] http://wifo5-03.informatik.uni-mannheim.de/diseasome/.
[24] https://bgee.org/sparql.
[25] https://sparql.omabrowser.org/sparql.

patterns in the questions. They select and align triple patterns using integer linear programming and achieve F1 of 0.88 on QALD-4 task 2. Hamon et al. [61] establish a complex pipeline to translate questions using existing NLP tools and semantic resources, and it achieves F1 of 0.85 on QALD-4 task 2.

Bioinformatics: Sima et al. [177] propose Bio-SODA, which converts natural language questions into SPARQL queries without training data. Bio-SODA generates a list of query graphs based on matched entities in the question and ranks the query graphs considering semantic and syntactic similarity and node centrality. Bio-SODA achieves F1 of 0.60 on QALD-4 task 2 and 0.60 on Bioinformatics.

Classic BQA systems also require natural language question translation systems to query KB. Rinaldi et al. [159] adapt the general domain ExtrAns system [123] to genomics domain. They first convert the documents to **Minimal Logical Forms (MLFs)** and use them to construct a KB during the offline phase. In the online QA phase, the system also converts the given question to MLFs by the same mechanism and then gets the answer by searching the built MLFs KB. Abacha and Zweigenbaum [5, 6] propose MEANS for medical BQA, which converts questions to SPARQL queries with a pipeline of classifying question types, finding Expected Answers Types, question simplification, medical entity recognition, extracting semantic relations, and constructing SPARQL-based on entities and semantic relations. Kim and Cohen [90] introduce Linked Open Data Question Answering system to generate SPARQL queries for SNOMED-CT by predicate-argument relations from sentences.

7.4 Comments

Existing KB BQA datasets are limited by size, making it hard to train learning-based methods. As a result, most top-performing KB BQA methods apply complex pipeline methods including entity extraction, relation extraction, entity alignment, and entity typing to construct queries. To leverage the potential of end-to-end deep learning methods, more labeled datasets are required for training a supervised Seq2seq question translation model.

8 QUESTION ENTAILMENT BQA

Harabagiu and Hickl [62] show that recognizing textual entailment can be used to enhance open QA systems. The QE approach for BQA is essentially a nearest-neighbor method that uses the answers of similar and already answered questions (e.g.: **Frequently Asked Questions, FAQs**) to answer the given question. We will discuss three main components of this approach in this section: 1. Models that can recognize similar questions, i.e., *QE BQA Methods*; 2. datasets of similar *Question-Question Pairs* for training QE models; 3. datasets of answered questions, i.e., *Question-Answer Pairs*, which can be used to answer new questions.

8.1 QE BQA Methods

QE is formally defined by Abacha and Demner-Fushman [2] as: a question Q_A entails a question Q_B if every answer to Q_B is also a correct answer to Q_A. **Natural language inference (NLI)** is a relevant NLP task that predicts whether the relation of entailment, contradiction, or neutral holds between a pair of sentences. In the general domain, predicting question-question similarity is an active research area with potential applications in question recommendation and community question answering [127, 128].

Luo et al. [110] propose the SimQ system to retrieve similar consumer health questions on the Web using UMLS-annotated semantic and AQUA-parsed [24] syntactic features of the questions. CMU OAQA [198] use a dual entailment approach with bidirectional **recurrent neural networks (RNN)** and attention mechanism to predict question similarity; Abacha and Demner-Fushman [3]

use feature-based logistic regression classifier and deep learning models that pass the concatenation of two question embeddings to multiple ReLU layers [126] for recognizing QE; Zhu et al. [240] fine-tune pre-trained language models to classify question pairs and conduct transfer learning from NLI to boost the performance.

8.2 Question-question Pairs

Training QE models needs datasets of question pairs annotated with entailment (similarity) labels. Towards this end, Abacha and Demner-Fushman [2] introduce the clinical-QE dataset[26] that contains over 8K training biomedical question pairs with similarity labels. The questions are from clinical questions collected by Ely et al. [49], Consumer Health Questions and FAQs from NLM and NIH websites, respectively. This dataset is also used as the RQE dataset in the MEDIQA challenge with slight changes. Poliak et al. [148], Sun and Sedoc [186], Zhang et al. [231] build question-question relevance datasets along with their FAQ datasets on COVID-19.

In general, only limited efforts have been made to build biomedical QE datasets, which results in the lack of training instances. Many works instead consider a transfer learning approach to pre-train the QE models on other text pair tasks, including biomedical NLI datasets like MedNLI [162], general domain QE datasets like SemEval-cQA [127, 128], and general domain NLI datasets such as SNLI [21] and MultiNLI [208].

8.3 Question-answer Pairs

QE approach relies heavily on large databases of question-answer pairs with high quality to answer unseen questions. For this, Abacha and Demner-Fushman [3] build MedQuAD, a collection of 47,457 question-answer pairs from trusted websites e.g.: https://www.cancer.gov/. Using MedQuAD for BQA can protect users from misleading and harmful health information, since most answers are well-curated. Moreover, several FAQ datasets have been introduced for answering COVID-19 related questions [148, 186, 231]. However, since expert-curated answers are expensive to collect, such question-answer pair datasets might be limited in size. Online health and QA communities such as WebMD[27] and Quora[28] provide large amounts of QA pairs, and many large-scale BQA datasets have been built using online doctor-patient QA data [63, 192, 228, 229]. These materials can be potentially used in QE approach, but the quality of user-provided answers should be carefully checked.

8.4 Comments

The most important components for the QE BQA approach are the datasets of question-question (Q-Q) and question-answer (Q-A) pairs. However, these datasets are currently limited in scale or quality. To tackle this issue, methods for automatically collecting large-scale Q-Q and Q-A datasets with high quality should be explored (Section 9.1).

9 CHALLENGES AND FUTURE DIRECTIONS

In this section, we discuss the challenges identified in Section 1 with the surveyed works. We also propose some interesting and potential future directions to explore. Section 9.1 and Section 9.2 involve dataset scale and difficulty, respectively. In Section 9.3, we discuss visual BQA, which is an active and novel research field that is gaining more attention. We explain why domain knowledge is not fully utilized in BQA and how the fusion of different BQA approaches can potentially solve it

[26] https://github.com/abachaa/RQE_Data_AMIA2016.
[27] https://www.webmd.com/.
[28] https://www.quora.com/.

in Section 9.4. In Section 9.5, we study different forms of explainability of BQA systems. In Section 9.6, we discuss two main issues of BQA system evaluation: Qualitatively, what parts of the systems are evaluated and, quantitatively, how they are evaluated. Last but not the least, we discuss the fairness and bias issues of BQA in Section 9.7.

9.1 Dataset Collection

Annotating large-scale BQA datasets is prohibitively expensive, since it requires intensive expert involvement. As a result, the majority of BQA datasets are automatically collected. There are three main approaches for automatically collecting BQA datasets: question generation, cloze generation, and exploiting existing QA pairs.

Question Generation: **Question generation (QG)** can automatically generate questions from given contexts [45], which can be utilized to build QA datasets. Yue et al. [226] apply the QG approach to synthesize clinical QA datasets without human annotations and show that the generated datasets can be used to improve BQA models on new contexts.

Cloze Generation: Cloze-type QA is the task of predicting a removed word or phrase in a sentence, usually using a detailed context. Several biomedical QA datasets have been collected by cloze generation, such as CliCR, BioRead, BMKC, and BioMRC. Most of them follow a similar process to Hermann et al. [67] for generating the CNN and Daily Mail datasets: 1. Recognizing biomedical named entities appearing in both summary sentences (e.g., article titles, article learning points) and their detailed contexts (e.g., article abstracts) with tools like MetaMap; 2. Masking the recognized named entities in the summary sentences; 3. The task is to predict the named entities using the masked summary sentences and the contexts. The generated datasets are typically large-scale (ranging from 100K to 16.4M instances) and thus can be used for pre-training [43] or as a task itself [91, 139, 142]. When used for pre-training, cloze-type QA is actually a special type of language modeling that predicts only biomedical entities and is conditioned on the corresponding contexts.

Exploiting Existing QA Pairs: Another widely used approach for dataset collection is to exploit naturally existing QA Pairs or exploiting domain-specific corpora structures. Biomedical question-answer pairs can be found in a variety of contexts: For example, PubMedQA [81] collects citations in PubMed whose titles are questions and uses the conclusive part of the abstracts as ideal answers. MedQA [230], HEAD-QA [196], and NLPEC [101] are built from QA pairs in medical examinations. MedQuAD collects expert-curated FAQs from trusted medical websites. cMedQA [228, 229] and ChiMed [192] exploit doctor-patient QA pairs on online healthcare communities.

9.2 Dataset Difficulty

BQA datasets should be difficult to evaluate the non-trivial reasoning abilities of BQA systems. In this section, we discuss three types of advanced reasoning abilities: answerability, multi-hop, and numeric reasoning.

Answerability reasoning: Almost all current BQA datasets and methods focus on answerable questions. However, not all questions in biomedicine are answerable; the fact that only answerable questions are evaluated can be exploited by BQA systems to get high performance without the expected reasoning process (e.g., by identifying the only text snippet in the context that is consistent with the expected lexical answer type). In general domain, a strong neural baseline drops from 86% F1 on SQuAD v1.1 to 66% F1 after unanswerable questions are added [154]. It remains an interesting direction to add unanswerable questions in BQA datasets and test the robustness of BQA systems under such settings.

Table 9. Overview of Biomedical VQA Datasets (Listed in Alphabetical Order)

Dataset	Size	Metric	State-of-the-art (%)	Content	Format
PathVQA [65]	32.8K	Acc, BLEU-1, BLEU-2, BLEU-3	68.2, 32.4, 22.8, 17.4 [65]	Clinical	Generation
VQA-Med [4]	15.3K	Acc, BLEU	64.0, 65.9 [158]	Clinical	Generation
VQA-Rad [97]	3.5K	Acc	60.0 (open)/79.3 (close) [227]	Clinical	Generation

Multi-hop reasoning: Answering real biomedical questions often requires multi-hop reasoning. For example, doctors might ask, "What tests shall we conduct for patients with [certain symptoms]?" To answer this question, models must: 1. infer the possible diseases that the patient might have and 2. find the available tests that are needed for the differential diagnosis. However, to the best of our knowledge, only MEDHOP evaluates multi-hop reasoning abilities, while almost all other BQA datasets focus on single-hop reasoning.

Numeric reasoning: Numbers are widely used in modern biomedical literature, so understanding the numeric contents in texts is necessary to correctly answer non-trivial scientific questions. Jin et al. [81] show that nearly all questions from PubMed article titles require quantitative reasoning to answer. While about 3/4 of them have text descriptions of the statistics, e.g., "significant differences," about 1/5 only have numbers. For this, Wu et al. [210] re-annotate the PubMedQA and BioASQ datasets with numerical facts as answers and show that adding numerical encoding scheme improves BioBERT performance on their dataset. However, most current BQA systems treat numbers as text tokens and do not have specific modules to process them.

9.3 Biomedical VQA

Biomedical VQA is a novel BQA task. In biomedical VQA, questions are asked about images, which are ubiquitously used and play a vital role in clinical decision-making. Since manual interpretation of medical images is time-consuming and error-prone, automatically answering natural language questions about medical images can be very helpful. VQA is a novel variant of QA task that requires both NLP techniques for question understanding and **Computer Vision (CV)** techniques for image representation. General VQA is an active research area, and there have been many recent survey articles [60, 182, 209]. Here, we mainly focus on *Biomedical VQA Datasets* and their corresponding *Multi-Modal Methods* that fuse NLP and CV methods.

Biomedical VQA Datasets. We show an overview of biomedical VQA datasets in Table 9 and an instance in Figure 2. **VQA-Rad** [97] is the first VQA dataset for radiology. It contains 3.5K QA pairs that are annotated by clinicians and the images are from MedPix.[29] Questions in VQA-Rad are classified into modality, plane, organ system, abnormality, object/condition presence, and so on. Answer formats include multi-choices and generations in VQA-Rad. The **VQA-Med** [4] dataset is automatically constructed using captions of the radiology images. There are 15.3K questions in VQA-Med that are restricted to be about only one element and should be answerable from the corresponding images. VQA-Med concentrates on the most common questions in radiology, which include categories of modality, plane, organ system, and abnormality. Yes/no and WH-questions are included in VQA-Med. **PathVQA** [65] semi-automatically extracts 32.8K pathology images and generates the corresponding answers from textbooks.

Multi-modal Methods. Typically, for biomedical VQA, images and texts are separately encoded, and a multi-modal pooling mechanism is often used to obtain the mixed representations for

[29] https://medpix.nlm.nih.gov/home.

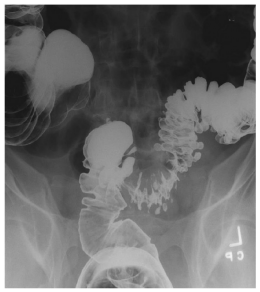

Fig. 2. An instance of VQA-Med.

generating the answers. ***Image encoders:*** VGGNet [178] and ResNet [64] are commonly used for image feature extraction. Ren and Zhou [158], Yan et al. [212] adopt global average pooling [105] on VGGNet for image encoding to prevent over-fitting on small datasets. To overcome data limitation of images, Nguyen et al. [131] apply model-agnostic meta-learning [53] and convolutional denoising auto-encoder [112] to initialize CNN layers on VQA-Rad and achieve 43.9 and 75.1 accuracy on open-ended and close-ended questions, respectively. ***Text encoders:*** Questions are usually encoded by a recurrent network or a pre-trained language model similar other BQA approaches. The co-attention mechanism is used for finding important words and regions to enhance both textual and visual representation. Stacked attention networks [214] use text representation to query visual representation multiple times for obtaining multi-step reasoning. ***Multi-modal pooling:*** It is crucial to combine features from the visual and textual encoders. Direct concatenation of them can serve as a baseline. Multi-modal compact bilinear pooling [56], multi-modal factorized bilinear pooling [221], and multi-modal factorized high-order pooling [222] are often used for feature fusion in VQA. Recently, several multi-modal pre-trained models have been proposed that use transformers [103, 189] to generate visual and textual representations in the general domain. Similarly, Ren and Zhou [158] introduce the CGMVQA model that feeds VGGNet and word embedding features into a single transformer for classification or generation on VQA-Med, achieving the accuracy of 0.640 and BLEU of 0.659.

9.4 Domain Knowledge Utilization

There are a variety of biomedical domain-specific materials and tools that can be used in BQA, including: 1. Large-scale corpora such as PubMed and PMC that contain millions of freely available biomedical articles; 2. Various biomedical KBs such as UMLS and DrugBank; 3. Many domain-specific NLP tools, e.g., MetaMap and SemRep, for identifying biomedical entities and relations, respectively. Each kind of resource has its advantages and disadvantages: Biomedical raw textual resources are extremely large-scale, but their quality cannot be assured. Specific textual resources, e.g., FAQs from NLM websites, are regularly maintained and thus of high quality, but they are limited in scale, since maintaining and collecting them is expensive. KBs have high quality and intensive knowledge, but most of them are sparse and incomplete.

However, the above-mentioned resources have not been fully utilized by current BQA systems. As shown in Table 10, different BQA approaches use only one or two different types of resources, but not all of them. For example, IR, MRC, and QE BQA systems typically use textual data, while KB BQA systems mainly use the KBs. Biomedical NLP tools are mostly used in classic BQA systems. Since each resource only encodes certain types of biomedical knowledge, only by fusing different BQA approaches can systems fully utilize the domain knowledge.

Table 10. Utilized Domain Knowledge by Different BQA Approaches

BQA Approach	Texts	Images	KBs	BioNLP tools
Information Retrieval	Document collections	–	–	–
Machine Reading Comprehension	Raw documents (Contexts)	–	–	–
Knowledge Base	–	–	✓	Used for KB construction
Question Entailment	Existing FAQs	–	–	–
Visual Question Answering	–	✓	–	–
Classic	Document collections	–	Ontologies	✓

The KMQA model [101] combines the IR-MRC approach and the KB approach by using co-attention mechanism and a novel knowledge acquisition algorithm. Their ablation experiments show that only using texts achieves 64.6% accuracy on the development set; only using knowledge bases achieves 45.3%; using both texts and knowledge bases achieves 71.1%. This shows the effectiveness of the fusion of different BQA approaches. However, it still remains an underexplored area.

9.5 Answer Explainability

Explainability is a vital property of healthcare applications. An ideal BQA model should not only have high accuracy in predicting the exact answers, but also be able to provide explanations or evidence for giving such answers. This improves the answer reliability and enables further fact checking.

Each BQA approach has its intrinsic way for answer explanation: For the IR approach, the retrieved documents can be considered as evidence; for the KB approach, the reasoning paths in the KBs provide explainability; for the QE approach, users are directly pointed to similar questions that have already been answered. Though controversial [78, 206], the attention mechanism [11] that is ubiquitously used in modern BQA systems provides at least some level of explainability. To evaluate explicit answer explanations, Phase B of BioASQ challenges also require the participants to submit "ideal answers," i.e., answers that include both the exact answers and explanations, in addition to exact answers.

Zhang et al. [234] generate explanations with medical KB paths that link the entities extracted from consumer health questions and doctors' answers. The path representations are learned by a translation-based method, and the weights of reasoning paths for specific QA pairs are generated by a hierarchical attention network. They also use the entity figures return from Google for better entity representation and consumer understanding. Liu et al. [108] present the MURKE model to solve HEAD-QA, which iteratively selects the most relative document to reformulate the question, where the series of modified questions can be considered as an interpretable reasoning chain.

9.6 Evaluation Issues

Modular evaluation: Most current evaluations are modular, because they only evaluate certain parts of the full BQA system, e.g., for the IR-MRC BQA approach, BioASQ Task B phase A only evaluates the IR methods and the Phase B provides gold standard contexts and only evaluates the MRC methods. The majority of BioASQ teams only participate in one phase [129]. However, in real settings: 1. it is impossible to have the relevant documents; 2. state-of-the-art MRC BQA systems might not perform well, given non-perfect retrieved documents [104]. As a result, closing the gap between system modules by combining the evaluations is vital to test the real utility of BQA systems.

Table 11. Overview of the BQA Datasets That Contain No Supporting Materials (Listed in Alphabetical Order)

Dataset	Size	Metric	State-of-the-art (%)	Content	Format
ChiMed [192]	24.9K	Acc	98.32 (rel.)/84.24 (adopt.) [192]	Consumer	Multi-choice
cMedQA [228]	54K	P@1	65.35 (dev)/64.75 (test) [228]	Consumer	Multi-choice
cMedQA v2 [229]	108K	P@1	72.1 (dev)/72.1 (test) [229]	Consumer	Multi-choice
HEAD-QA [196]	6.8K	Acc	44.4 (supervised)/46.7 (unsupervised) [108]	Examination	Multi-choice
LiveQA-Med [1]	738	avgScore	82.7 [2]	Consumer	Generation
MedQA [230]	235K	Acc	75.8 (dev)/75.3 (test) [230]	Examination	Multi-choice
MEDQA [79]	61K	Acc	MC: 69.3 (dev)/70.1 (test); TW: 42.2 (dev)/42.0 (test); US: 36.1 (dev)/36.7 (test) [79]	Examination	Multi-choice
NLPEC [101]	2.1K	Acc	71.1 (dev)/61.8 (test) [101]	Examination	Multi-choice
webMedQA [63]	63K	P@1, MAP	66.0, 79.5 [63]	Consumer	Multi-choice

In the general domain, Chen et al. [30] propose the Machine Reading at Scale task for the complete IR-MRC QA evaluation. They show that the performance of a complete QA system that reads all Wikipedia might have a large drop compared to its MRC component that reads only the gold standard contexts, e.g.: from 69.5% EM to 27.1% on the development set of SQuAD. In the biomedical domain, many datasets that only contain questions and answers have been proposed. We list these datasets in Table 11, most of which are related to Consumer health (5/10) or Examination (4/10), because their dataset sources typically have no supporting materials for the answers. It should be noted that other types of BQA datasets can also be converted to such datasets by removing the supporting materials (document collections, contexts, FAQs, etc.).

Olelo [92] and Bio-AnswerFinder [137] are complete QA systems that participate in the BioASQ challenge. Olelo is proposed as an integrated web application for QA-based exploration of biomedical literature. For each user question, Olelo uses the HPI system at BioASQ 2016 (Schulze et al. 169, described in Section 6) to retrieve relevant abstracts and return the answers, as well as the entity-supported summarizations of the abstracts [168]. Bio-AnswerFinder uses iterative document retrieval by LSTM-enhanced keyword queries and BERT-based answer ranking. The system performance is comparable to a BioASQ 5 SoTA MRC system for factoid questions (38.1% vs. 40.5% MRR, Wiese et al. [207]), but is still lower than BioBERT (38.1%, 48.3%). The baselines of ChiMed, cMedQA, and webMedQA use answer matching models without explicit supporting materials, and the baselines provided by HEAD-QA, MedQA, and MEDQA are basically combined IR-MRC approach (Section 5 and Section 6). Since most of the current BQA evaluations only focus on the MRC, tasks that involve both retrieving relevant contents and comprehending over them should be further explored.

Evaluation metrics: In extractive and generative BQA, current metrics do not consider the synonyms of biomedical concepts. For example, if the ground truth answer is "kidney diseases," then "renal diseases" should conceptually be an exact match, and "renal insufficiency" should be rated as relevant. However, if we use EM in practice, then both "renal diseases" and "renal insufficiency" have a score of 0; if we use F1, BLEU, or ROGUE, then "renal diseases" is only a partial match and "renal insufficiency" has a score of 0. Wiese et al. [207] report that their model predictions of 10 among 33 analyzed questions are synonyms to the gold standard answers, but are not counted as right in BioASQ evaluation.

There are two potential approaches to solve this problem: 1. From the annotation side, we can manually annotate more gold standard answers. This approach is expensive, but the quality is guaranteed; 2. From the metrics side, it is worth exploring to infuse domain knowledge (e.g.: UMLS ontologies) into current evaluation metrics. For example, to consider the rich concept synonyms in

biomedicine during evaluation, Šuster and Daelemans [187] also evaluate QA models by a cosine similarity metric between the mean word vectors of the ground truth and the predicted answer.

9.7 Fairness and Bias

Fairness and bias are serious and vital issues in machine learning. One cannot be more cautious in removing all potential biases, e.g., racial and gender biases, when developing healthcare applications like BQA. Here, we discuss the fairness and bias issues of BQA from the NLP and the biomedical side. From the NLP side: Word embeddings [116] are ubiquitously used in NLP models, but such embeddings result in biased analogies like: "man" is to "doctor" as "woman" is to "nurse." Similar trends have been observed [93] in contextualized word representations like BERT. From the biomedical side, since most current BQA models learn from historically collected data (e.g.: EMRs, scientific literature), populations that have experienced structural biases in the past might be vulnerable under incorrect predictions [153].

Some works have been done in general NLP and biomedical machine learning domains, but only a little progress has been made in the BQA domain, and the majority of them study non-English BQA. Unlike English BQA, non-English BQA suffers additional challenges mainly from the lack of domain-specific resources; much less scientific literature is available in non-English languages; general NLP tools are scarce for non-English languages, let alone biomedical domain-specific ones. Delbecque et al. [36], Jacquemart and Zweigenbaum [77] present preliminary studies of BQA in French. Olvera-Lobo and Gutiérrez-Artacho [136] evaluate multilingual QA system HONQA and find that English questions are answered much better than French and Italian. Researchers also introduce multi-lingual BQA datasets for low-resource languages: He et al. [63], Tian et al. [192] Zhang et al. [228–230] for Chinese, Vilares and Gómez-Rodríguez [196] for Spanish, and Veisi and Shandi [195] for Persian.

However, current works are far from enough, and our community should seriously take fairness and bias issues into account when introducing new BQA datasets and algorithms in the future.

REFERENCES

[1] Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at TREC 2017 LiveQA. In *Proceedings of the Text Retrieval Conference (TREC)*.

[2] Asma Ben Abacha and Dina Demner-Fushman. 2016. Recognizing question entailment for medical question answering. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association.

[3] Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinform.* 20, 1 (2019), 511.

[4] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. 2019. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *CLEF 2019 Working Notes*.

[5] Asma Ben Abacha and Pierre Zweigenbaum. 2012. Medical question answering: translating medical questions into sparql queries. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. 41–50.

[6] Asma Ben Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic web technologies. *Inf. Process. Manag.* 51, 5 (2015), 570–594.

[7] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, 72–78. DOI: https://doi.org/10.18653/v1/W19-1909

[8] Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association.

[9] Alan R. Aronson, Dina Demner-Fushman, Susanne M. Humphrey, and Jimmy J. Lin. 2005. Fusion of knowledge-intensive and statistical approaches for retrieving and annotating textual genomics documents. In *Proceedings of the Text Retrieval Conference (TREC)*.

[10] Sofia J. Athenikos and Hyoil Han. 2010. Biomedical question answering: A survey. *Comput. Meth. Prog. Biomed.* 99, 1 (2010), 1–24.

[11] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[12] Georgios Balikas, Ioannis Partalas, Axel-Cyrille Ngonga Ngomo, Anastasia Krithara, Eric Gaussier, and George Paliouras. 2014. Results of the BioASQ tasks of the question answering lab at CLEF 2014. In *CLEF 2014 Working Notes*.

[13] Michael A. Bauer and Daniel Berleant. 2012. Usability survey of biomedical question answering systems. *Hum. Genom.* 6, 1 (2012), 17.

[14] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 3615–3620. DOI: https://doi.org/10.18653/v1/D19-1371

[15] Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2228–2234. DOI: https://doi.org/10.18653/v1/P19-1215

[16] Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, 370–379. DOI: https://doi.org/10.18653/v1/W19-5039

[17] Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1499–1510. DOI: https://doi.org/10.3115/v1/D14-1159

[18] Abhishek Bhandwaldar and Wlodek Zadrozny. 2018. UNCC QA: Biomedical Question Answering system. In *Proceedings of the 6th BioASQ Workshop*. Association for Computational Linguistics, 66–71. DOI: https://doi.org/10.18653/v1/W18-5308

[19] Pinaki Bhaskar, Partha Pakray, Somnath Banerjee, Samadrita Banerjee, Sivaji Bandyopadhyay, and Alexander F. Gelbukh. 2012. Question answering system for QA4MRE@ CLEF 2012. In *Proceedings of the CLEF Online Working Notes/Labs/Workshop*.

[20] Ludovic Bonnefoy, Romain Deveaud, and Patrice Bellot. 2012. Do social information help book search? In *Workshop Pre-proceedings INEX'12*.

[21] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 632–642. DOI: https://doi.org/10.18653/v1/D15-1075

[22] George Brokos, Polyvios Liosis, Ryan McDonald, Dimitris Pappas, and Ion Androutsopoulos. 2018. AUEB at BioASQ 6: Document and snippet retrieval. In *Proceedings of the 6th BioASQ Workshop*. Association for Computational Linguistics, 30–39. DOI: https://doi.org/10.18653/v1/W18-5304

[23] Brian L. Cairns, Rodney D. Nielsen, James J. Masanz, James H. Martin, Martha S. Palmer, Wayne H. Ward, and Guergana K. Savova. 2011. The MiPACQ clinical question answering system. In *AMIA Annual Symposium Proceedings*, Vol. 2011. American Medical Informatics Association.

[24] David Campbell and Stephen Johnson. 2002. A transformational-based learner for dependency grammars in discharge summaries. In *Proceedings of the ACL-02 Workshop on Natural Language Processing in the Biomedical Domain*. Association for Computational Linguistics, 37–44. DOI: https://doi.org/10.3115/1118149.1118155

[25] YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J. Cimino, John Ely, and Hong Yu. 2011. AskHERMES: An online question answering system for complex clinical questions. *J. Biomed. Inform.* 44, 2 (2011), 277–288.

[26] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 335–336.

[27] Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. 2020. BioMedBERT: A pre-trained biomedical language model for QA and IR. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 669–679. DOI: https://doi.org/10.18653/v1/2020.coling-main.59

[28] Rishav Chakravarti, Anthony Ferritto, Bhavani Iyer, Lin Pan, Radu Florian, Salim Roukos, and Avi Sil. 2020. Towards building a robust industry-scale question answering system. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*. International Committee on Computational Linguistics, 90–101. Retrieved from https://www.aclweb.org/anthology/2020.coling-industry.9.

[29] Khyathi Chandu, Aakanksha Naik, Aditya Chandrasekar, Zi Yang, Niloy Gupta, and Eric Nyberg. 2017. Tackling biomedical text summarization: OAQA at BioASQ 5B. In *Proceedings of the BioNLP 2017*. Association for Computational Linguistics, 58–66. DOI: https://doi.org/10.18653/v1/W17-2307

[30] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1870–1879. DOI: https://doi.org/10.18653/v1/P17-1171

[31] Jianpeng Cheng, Li Dong, and Mirella Lapata. 2016. Long short-term memory-networks for machine reading. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 551–561. DOI: https://doi.org/10.18653/v1/D16-1053

[32] Sungbin Choi. 2015. SNUMedinfo at CLEF QA track BioASQ 2015. In *CLEF 2015 Working Notes*.

[33] Stéphane Clinchant and Eric Gaussier. 2009. Bridging language modeling and divergence from randomness models: A log-logistic model for IR. In *Proceedings of the Conference on the Theory of Information Retrieval*. Springer, 54–65.

[34] Sarah Cruchet, Arnaud Gaudinat, and Célia Boyer. 2008. Supervised approach to recognize question type in a QA system for health. *Stud. Health Technol. Inform.* 136 (2008), 407.

[35] Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-Attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 593–602. DOI: https://doi.org/10.18653/v1/P17-1055

[36] T. Delbecque, P. Jacquemart, and P. Zweigenbaum. 2005. Indexing UMLS semantic types for medical question-answering. *Stud. Health Technol. and Inform.* 116 (2005), 805–810.

[37] Dina Demner-Fushman, S. Humphrey, Nicholas C. Ide, R. Loane, James G. Mork, P. Ruch, M. Ruiz, L. H. Smith, W. Wilbur, and A. Aronson. 2007. Combining resources to find answers to biomedical questions. In *Proceedings of the Text Retrieval Conference (TREC)*.

[38] Dina Demner-Fushman and Jimmy Lin. 2005. *Knowledge Extraction for Clinical Question Answering: Preliminary Results*. AAAI Workshop - Technical Report (01 2005).

[39] Dina Demner-Fushman and Jimmy Lin. 2006. Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 841–848. DOI: https://doi.org/10.3115/1220175.1220281

[40] Dina Demner-Fushman and Jimmy Lin. 2007. Answering clinical questions with knowledge-based and statistical techniques. *Comput. Ling.* 33, 1 (2007), 63–103.

[41] Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: Helping consumers find answers to their health-related information needs. *J. Amer. Med. Inform. Assoc.* 27, 2 (2020), 194–201.

[42] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[43] Bhuwan Dhingra, Danish Danish, and Dheeraj Rajagopal. 2018. Simple and effective semi-supervised question answering. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. Association for Computational Linguistics, 582–587. DOI: https://doi.org/10.18653/v1/N18-2092

[44] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1832–1846. DOI: https://doi.org/10.18653/v1/P17-1168

[45] Xinya Du, Junru Shao, and Claire Cardie. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1342–1352. DOI: https://doi.org/10.18653/v1/P17-1123

[46] Yongping Du, Wenyang Guo, and Yiliang Zhao. 2019. Hierarchical question-aware context learning with augmented data for biomedical question answering. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 370–375.

[47] Yongping Du, Bingbing Pei, Xiaozheng Zhao, and Junzhong Ji. 2018. Hierarchical multi-layer transfer learning model for biomedical question answering. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 362–367.

[48] John W. Ely, Jerome A. Osheroff, M. Lee Chambliss, Mark H. Ebell, and Marcy E. Rosenbaum. 2005. Answering physicians' clinical questions: Obstacles and potential solutions. *J. Amer. Med. Inform. Assoc.* 12, 2 (2005), 217–224.

[49] John W. Ely, Jerome A. Osheroff, Paul N. Gorman, Mark H. Ebell, M. Lee Chambliss, Eric A. Pifer, and P. Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *Bmj* 321, 7258 (2000), 429–432.

[50] Günes Erkan and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* 22 (2004), 457–479.
[51] D. A. Ferrucci. 2012. Introduction to "This is Watson." *IBM J. Res. Devel.* 56, 3.4 (2012), 1:1–1:15. DOI: https://doi.org/10.1147/JRD.2012.2184356
[52] Katja Filippova, Enrique Alfonseca, Carlos A. Colmenares, Lukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with LSTMs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 360–368. DOI: https://doi.org/10.18653/v1/D15-1042
[53] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. *arXiv preprint arXiv:1703.03400* (2017).
[54] Susannah Fox and Maeve Duggan. 2012. Health Online 2013. *Pew Res. Internet Proj. Rep.* (01 2012).
[55] Bin Fu, Yunqi Qiu, Chengguang Tang, Yang Li, Haiyang Yu, and Jian Sun. 2020. A survey on complex question answering over knowledge base: Recent advances and challenges. *arXiv preprint arXiv:2007.13069* (2020).
[56] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847* (2016).
[57] Julien Gobeill, E. Patsche, D. Theodoro, A.-L. Veuthey, C. Lovis, and P. Ruch. 2009. Question answering for biology and medicine. In *Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine*. IEEE, 1–5.
[58] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779* (2020).
[59] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Bruce Croft. 2016. A deep relevance matching model for ad hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (CIKM'16)*. Association for Computing Machinery, New York, NY, 55–64. DOI: https://doi.org/10.1145/2983323.2983769
[60] Akshay Kumar Gupta. 2017. Survey of visual question answering: Datasets and techniques. *arXiv preprint arXiv:1705.03865* (2017).
[61] Thierry Hamon, Natalia Grabar, and Fleur Mougin. 2017. Querying biomedical linked data with natural language questions. *Seman. Web* 8, 4 (2017), 581–599.
[62] Sanda Harabagiu and Andrew Hickl. 2006. Methods for using textual entailment in open-domain question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 905–912. DOI: https://doi.org/10.3115/1220175.1220289
[63] Junqing He, Mingming Fu, and Manshu Tu. 2019. Applying deep matching networks to Chinese medical question answering: A study and a dataset. *BMC Med. Inform. Decis.-mak.* 19, 2 (2019), 52.
[64] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.
[65] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286* (2020).
[66] Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 4604–4614. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.372
[67] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 1693–1701.
[68] William Hersh, Aaron Cohen, Lynn Ruslen, and Phoebe Roberts. 2007. TREC 2007 genomics track overview. In *Proceedings of the Text Retrieval Conference (TREC)*.
[69] William Hersh, Aaron M. Cohen, Phoebe Roberts, and Hari Krishna Rekapalli. 2006. TREC 2006 genomics track overview. In *Proceedings of the Text Retrieval Conference (TREC)*.
[70] William Hersh and Ellen Voorhees. 2009. TREC genomics special issue overview. *Inf. Retr.* 12, 1 (Feb. 2009), 1–15. DOI: https://doi.org/10.1007/s10791-008-9076-6
[71] William R. Hersh, M. Katherine Crabtree, David H. Hickam, Lynetta Sacherek, Charles P. Friedman, Patricia Tidmarsh, Craig Mosbaek, and Dale Kraemer. 2002. Factors associated with success in searching MEDLINE and applying evidence to answer clinical questions. *J. Amer. Med. Inform. Assoc.* 9, 3 (2002), 283–293.
[72] L. Hirschman and R. Gaizauskas. 2001. Natural language question answering: The view from here. *Nat. Lang. Eng.* 7, 4 (Dec. 2001), 275–300. DOI: https://doi.org/10.1017/S1351324901002807
[73] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).

[74] Xiaoli Huang, Jimmy Lin, and Dina Demner-Fushman. 2006. Evaluation of PICO as a knowledge representation for clinical questions. In *AMIA Annual Symposium Proceedings*, Vol. 2006. American Medical Informatics Association.
[75] Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. PACRR: A position-aware neural IR model for relevance matching. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 1049–1058. DOI: https://doi.org/10.18653/v1/D17-1110
[76] Lijun Huo and Xiang Zhao. 2020. A sentence-based circular reasoning model in multi-hop reading comprehension. *IEEE Access* 8 (2020), 174255–174264.
[77] P. Jacquemart and P. Zweigenbaum. 2003. Towards a medical question-answering system: A feasibility study. *Stud. Health Technol. Inform.* 95 (2003), 463.
[78] Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 3543–3556. DOI: https://doi.org/10.18653/v1/N19-1357
[79] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081* (2020).
[80] Qiao Jin, Bhuwan Dhingra, William Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*. 82–89.
[81] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2567–2577. DOI: https://doi.org/10.18653/v1/D19-1259
[82] Zan-Xia Jin, Bo-Wen Zhang, Fan Fang, Le-Le Zhang, and Xu-Cheng Yin. 2017. A multi-strategy query processing approach for biomedical question answering: USTB_PRIR at BioASQ 2017 Task 5B. In *Proceedings of the Biomedical Natural Language Processing Workshop (BioNLP)*. Association for Computational Linguistics, 373–380. DOI: https://doi.org/10.18653/v1/W17-2348
[83] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1601–1611. DOI: https://doi.org/10.18653/v1/P17-1147
[84] Z. Kaddari, Y. Mellah, J. Berrich, T. Bouchentouf, and M. G. Belkasmi. 2020. Biomedical question answering: A survey of methods and datasets. In *Proceedings of the 4th International Conference On Intelligent Computing in Data Sciences (ICDS)*. 1–8. DOI: https://doi.org/10.1109/ICDS50568.2020.9268742
[85] Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention Sum Reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 908–918. DOI: https://doi.org/10.18653/v1/P16-1086
[86] Aishwarya Kamath and Rajarshi Das. 2018. A survey on semantic parsing. *arXiv preprint arXiv:1812.00978* (2018).
[87] Maulik R. Kamdar and Mark A. Musen. 2020. An empirical meta-analysis of the life sciences (Linked?) open data on the web. *arXiv preprint arXiv:2006.04161* (2020).
[88] Jaewoo Kang. 2020. Transferability of natural language inference to biomedical question answering. *arXiv preprint arXiv:2007.00217* (2020).
[89] Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. *arXiv preprint arXiv:1604.06076* (2016).
[90] Jin-Dong Kim and K. Bretonnel Cohen. 2013. Natural language query processing for SPARQL generation: A prototype system for SNOMED CT. In *Proceedings of Biolink*, Vol. 32. Academia.
[91] Seongsoon Kim, Donghyeon Park, Yonghwa Choi, Kyubum Lee, Byounggun Kim, Minji Jeon, Jihye Kim, Aik Choon Tan, and Jaewoo Kang. 2018. A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis. *JMIR Medical Inform.* 6, 1 (2018).
[92] Milena Kraus, Julian Niedermeier, Marcel Jankrift, Sören Tietböhl, Toni Stachewicz, Hendrik Folkerts, Matthias Uflacker, and Mariana Neves. 2017. Olelo: A web application for intuitive exploration of biomedical literature. *Nucleic Acids Res.* 45, W1 (2017), W478–W483.
[93] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W. Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the 1st Workshop on Gender Bias in Natural Language Processing*. Association for Computational Linguistics, 166–172. DOI: https://doi.org/10.18653/v1/W19-3823
[94] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput. Ling.* 7 (2019), 453–466.

[95] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 785–794. DOI: https://doi.org/10.18653/v1/D17-1082

[96] A. Lamurias, D. Sousa, and F. M. Couto. 2020. Generating biomedical question answering corpora from Q A Forums. *IEEE Access* 8 (2020), 161042–161051. DOI: https://doi.org/10.1109/ACCESS.2020.3020868

[97] Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* 5, 1 (2018), 1–10.

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

[99] Minsuk Lee, James Cimino, Hai Ran Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. 2006. Beyond information retrieval—Medical question answering. In *AMIA Annual Symposium Proceedings*, Vol. 2006. American Medical Informatics Association.

[100] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7871–7880. DOI: https://doi.org/10.18653/v1/2020.acl-main.703

[101] Dongfang Li, Baotian Hu, Qingcai Chen, Weihua Peng, and Anqi Wang. 2020. Towards medical machine reading comprehension with structural knowledge and plain text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1427–1438. DOI: https://doi.org/10.18653/v1/2020.emnlp-main.111

[102] Guanqiao Li, Yangzhong Zhou, Junyi Ji, Xiaozhen Liu, Qiao Jin, and Linqi Zhang. 2020. Surging publications on the COVID-19 pandemic. *Clin. Microbiol. Infect.* 27, 3 (2020).

[103] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv:1908.03557 [cs.CV]*

[104] Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. 2020. Pretrained transformers for text ranking: BERT and beyond. *arXiv preprint arXiv:2010.06467* (2020).

[105] Min Lin, Qiang Chen, and Shuicheng Yan. 2013. Network in network. *arXiv preprint arXiv:1312.4400* (2013).

[106] Ryan T. K. Lin, Justin Liang-Te Chiu, Hong-Jei Dai, Min-Yuh Day, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2008. Biological question answering with syntactic and semantic feature matching and an improved mean reciprocal ranking measurement. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*. IEEE, 184–189.

[107] Yifeng Liu. 2013. The University of Alberta participation in the BioASQ challenge: The Wishart system. In *Proceedings of the 1st Workshop Bio-Medical Semantic Indexing Question Answering, Conference Labs Evaluation Forum*. 1–4.

[108] Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S. Yu. 2020. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *arXiv preprint arXiv:2008.02434* (2020).

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

[110] Jake Luo, Guo-Qiang Zhang, Susan Wentz, Licong Cui, and Rong Xu. 2015. SimQ: real-time retrieval of similar consumer health questions. *J. Med. Internet Res.* 17, 2 (2015).

[111] Anca Marginean. 2017. Question answering over biomedical linked data with grammatical framework. *Seman. Web* 8, 4 (2017), 565–580.

[112] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. 2011. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Proceedings of the International Conference on Artificial Neural Networks*. Springer, 52–59.

[113] Giuseppe M. Mazzeo and Carlo Zaniolo. 2016. Question answering on RDF KBs using controlled natural language and semantic autocompletion. *Seman. Web* 1 (2016), 1–5.

[114] Gabor Melli, Yang Wang, Yudong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar, and Fred Popowich. 2005. Description of SQUASH, the SFU question answering summary handler for the DUC-2005 summarization task. *Safety* 1 (2005), 14345754.

[115] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). Retrieved from http://arxiv.org/abs/1301.3781.

[116] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*. 3111–3119.

[117] Diego Mollá. 2017. Macquarie University at BioASQ 5b—Query-based summarisation techniques for selecting the ideal answers. In *Proceedings of the Biomedical Natural Language Processing Workshop (BioNLP)*. Association for Computational Linguistics, 67–75. DOI: https://doi.org/10.18653/v1/W17-2308

[118] Diego Mollá. 2018. Macquarie University at BioASQ 6b: Deep learning and deep reinforcement learning for query-based summarisation. In *Proceedings of the 6th BioASQ Workshop*. Association for Computational Linguistics, 22–29. DOI: https://doi.org/10.18653/v1/W18-5303

[119] Diego Mollá and Christopher Jones. 2019. Classification betters regression in query-based multi-document summarisation techniques for question answering. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 624–635.

[120] Diego Molla, Christopher Jones, and Vincent Nguyen. 2020. Query focused multi-document summarisation of biomedical texts. *arXiv preprint arXiv:2008.11986* (2020).

[121] Diego Molla and Maria Elena Santiago-Martinez. 2011. Development of a corpus for evidence based medicine summarisation. In *Proceedings of the Australasian Language Technology Association Workshop*. 86–94. Retrieved from https://www.aclweb.org/anthology/U11-1012.

[122] Diego Mollá, María Elena Santiago-Martínez, Abeed Sarker, and Cécile Paris. 2016. A corpus for research in text processing for evidence based medicine. *Lang. Resour. Eval.* 50, 4 (2016), 705–727.

[123] Diego Mollá, Rolf Schwitter, Michael Hess, and Rachel Fournier. 2000. ExtrAns, an answer extraction system. In *T.A.L.* 41, 2 (2000), 1–25.

[124] Timo Moller, Anthony Reina, Raghavan Jayakumar, and Malte Pietsch. 2020. COVID-QA: A question answering dataset for COVID-19. Retrieved from https://openreview.net/forum?id=JENSKEEzsoU.

[125] Roser Morante, Martin Krallinger, Alfonso Valencia, and Walter Daelemans. 2012. Machine reading of biomedical texts about Alzheimer's disease. In *CLEF 2012 Conference and Labs of the Evaluation Forum-question Answering For Machine Reading Evaluation (QA4MRE)*, J. Forner (Ed.). CEUR-WS, 1–14.

[126] Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. 807–814.

[127] Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics, 27–48. DOI: https://doi.org/10.18653/v1/S17-2003

[128] Preslav Nakov, Lluís Màrquez, Alessandro Moschitti, Walid Magdy, Hamdy Mubarak, Abed Alhakim Freihat, Jim Glass, and Bilal Randeree. 2016. SemEval-2016 Task 3: Community question answering. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval)*. Association for Computational Linguistics. 525–545. DOI: https://doi.org/10.18653/v1/S16-1083

[129] Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Martin Krallinger, Carlos Rodriguez-Penagos, Marta Villegas, and Georgios Paliouras. 2020. Overview of BioASQ 2020: The eighth BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro (Eds.). Springer International Publishing, Cham, 194–214.

[130] Mariana Neves and Ulf Leser. 2015. Question answering for biology. *Methods* 74 (2015), 36–46.

[131] Binh D. Nguyen, Thanh-Toan Do, Binh X. Nguyen, Tuong Do, Erman Tjiputra, and Quang D. Tran. 2019. Overcoming data limitation in medical visual question answering. In *Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, 522–530.

[132] Vincent Nguyen. 2019. Question answering in the biomedical domain. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, 54–63. DOI: https://doi.org/10.18653/v1/P19-2008

[133] David N. Nicholson and Casey S. Greene. 2020. Constructing knowledge graphs and their biomedical applications. *Comput. Struct. Biotechnol. J.* 18 (2020), 1414.

[134] Yun Niu, Graeme Hirst, Gregory McArthur, and Patricia Rodriguez-Gianolli. 2003. Answering clinical questions with role identification. In *Proceedings of the ACL Workshop on Natural Language Processing in Biomedicine*. Association for Computational Linguistics. 73–80. DOI: https://doi.org/10.3115/1118958.1118968

[135] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics. 708–718. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.63

[136] María-Dolores Olvera-Lobo and Juncal Gutiérrez-Artacho. 2011. Multilingual question-answering system in biomedical domain on the web: An evaluation. In *Proceedings of the International Conference of the Cross-language Evaluation Forum for European Languages*. Springer, 83–88.

[137] Ibrahim Burak Ozyurt, Anita Bandrowski, and Jeffrey S. Grethe. 2020. Bio-AnswerFinder: A system to find answers to questions from biomedical texts. *Database* 2020 (2020).

[138] Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2357–2368. DOI: https://doi.org/10.18653/v1/D18-1258

[139] Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A new dataset for biomedical reading comprehension. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA). Retrieved from https://www.aclweb.org/anthology/L18-1439.

[140] Dimitris Pappas, Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2019. AUEB at BioASQ 7: document and snippet retrieval. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 607–623.

[141] Dimitris Pappas, Petros Stavropoulos, and Ion Androutsopoulos. 2020. AUEB-NLP at BioASQ 8: Biomedical document and snippet retrieval. In *CLEF 2020 Working Notes*.

[142] Dimitris Pappas, Petros Stavropoulos, Ion Androutsopoulos, and Ryan McDonald. 2020. BioMRC: A dataset for biomedical machine reading comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, 140–149. Retrieved from https://www.aclweb.org/anthology/2020.bionlp-1.15.

[143] Junwoo Park, Youngwoo Cho, Haneol Lee, Jaegul Choo, and Edward Choi. 2020. Knowledge graph-based question answering with electronic health records. *arXiv preprint arXiv:2010.09394* (2020).

[144] Ioannis Partalas, Eric Gaussier, Axel-Cyrille Ngonga Ngomo, et al. 2013. Results of the first BioASQ workshop. In *BioASQ@CLEF 2013*.

[145] Anselmo Penas, Yusuke Miyao, Alvaro Rodrigo, Eduard H. Hovy, and Noriko Kando. 2014. Overview of CLEF QA entrance exams task 2014. In *CLEF (Working Notes)*. CEUR-WS, 1194–1200.

[146] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2227–2237. DOI: https://doi.org/10.18653/v1/N18-1202

[147] Mai Phuong Pham et al. 2020. *Machine Comprehension for Clinical Case Reports*. Ph.D. Dissertation. Massachusetts Institute of Technology.

[148] Adam Poliak, Max Fleming, Cash Costello, Kenton W. Murray, Mahsa Yarmohammadi, Shivani Pandya, Darius Irani, Milind Agarwal, Udit Sharma, Shuo Sun, Nicola Ivanov, Lingxi Shang, Kaushik Srinivasan, Seolhwa Lee, Xu Han, Smisha Agarwal, and João Sedoc. 2020. Collecting verified COVID-19 question answer pairs. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.nlpcovid19-2.31

[149] Hemant Pugaliya, Karan Saxena, Shefali Garg, Sheetal Shalini, Prashant Gupta, Eric Nyberg, and Teruko Mitamura. 2019. Pentagon at MEDIQA 2019: Multi-task learning for filtering and re-ranking answers using language inference and question entailment. *arXiv preprint arXiv:1907.01643* (2019).

[150] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. AliMe Chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 498–503. DOI: https://doi.org/10.18653/v1/P17-2079

[151] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.

[152] Preethi Raghavan, Siddharth Patwardhan, Jennifer J. Liang, and Murthy V. Devarakonda. 2018. Annotating electronic medical records for question answering. *arXiv preprint arXiv:1805.06816* (2018).

[153] Alvin Rajkomar, Michaela Hardt, Michael D. Howell, Greg Corrado, and Marshall H. Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Ann. Intern. Med.* 169, 12 (2018), 866–872.

[154] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 784–789. DOI: https://doi.org/10.18653/v1/P18-2124

[155] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2383–2392. DOI: https://doi.org/10.18653/v1/D16-1264

[156] Aarne Ranta, Ali El Dada, and Janna Khegai. 2009. The GF resource grammar library. *Ling. Issues Lang. Technol.* 2, 2 (2009), 1–63.

[157] Revanth Gangi Reddy, Bhavani Iyer, Md Arafat Sultan, Rong Zhang, Avi Sil, Vittorio Castelli, Radu Florian, and Salim Roukos. 2020. End-to-end QA on COVID-19: Domain adaptation with synthetic training. *arXiv preprint arXiv:2012.01414* (2020).

[158] Fuji Ren and Yangyang Zhou. 2020. CGMVQA: A new classification and generative model for medical visual question answering. *IEEE Access* 8 (2020), 50626–50636.

[159] Fabio Rinaldi, James Dowdall, Gerold Schneider, and Andreas Persidis. 2004. Answering questions in the genomics domain. In *Proceedings of the Conference on Question Answering in Restricted Domains*. Association for Computational Linguistics, 46–53. Retrieved from https://www.aclweb.org/anthology/W04-0508.

[160] Kirk Roberts and Dina Demner-Fushman. 2016. Interactive use of online health resources: a comparison of consumer and professional questions. *J. Amer. Med. Inform. Assoc.* 23, 4 (2016), 802–811.

[161] Kirk Roberts and Braja Gopal Patra. 2017. A semantic parsing method for mapping clinical questions to logical forms. In *AMIA Annual Symposium Proceedings*, Vol. 2017. American Medical Informatics Association.

[162] Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 1586–1596. DOI: https://doi.org/10.18653/v1/D18-1187

[163] Subendhu Rongali, Abhyuday Jagannatha, Bhanu Pratap Singh Rawat, and Hong Yu. 2020. Improved pretraining for domain-specific contextual embedding models. *arXiv preprint arXiv:2004.02288* (2020).

[164] Tony Russell-Rose and Jon Chamberlain. 2017. Expert search strategies: The information retrieval practices of healthcare information professionals. *JMIR Med. Inform.* 5, 4 (2017).

[165] David L. Sackett. 1997. Evidence-based medicine. In *Seminars in Perinatology*, Vol. 21. Elsevier, 3–5.

[166] Abeed Sarker, Diego Mollá, and Cécile Paris. 2013. An approach for query-focused text summarisation for evidence based medicine. In *Artificial Intelligence in Medicine*, Niels Peek, Roque Marín Morales, and Mor Peleg (Eds.). Springer Berlin, 295–304.

[167] Max Savery, Asma Ben Abacha, Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *arXiv e-prints* (May 2020). *arXiv:2005.09067 [cs.CL]*.

[168] Frederik Schulze and Mariana Neves. 2016. Entity-Supported summarization of biomedical abstracts. In *Proceedings of the 5th Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM)*. The COLING 2016 Organizing Committee, 40–49. Retrieved from https://www.aclweb.org/anthology/W16-5105.

[169] Frederik Schulze, Ricarda Schüler, Tim Draeger, Daniel Dummer, Alexander Ernst, Pedro Flemming, Cindy Perscheid, and Mariana Neves. 2016. HPI question answering system in BioASQ 2016. In *Proceedings of the 4th BioASQ Workshop*. 38–44.

[170] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 1073–1083. DOI: https://doi.org/10.18653/v1/P17-1099

[171] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603* (2016).

[172] Elaheh ShafieiBavani, Mohammad Ebrahimi, Raymond Wong, and Fang Chen. 2016. Appraising UMLS coverage for summarizing medical evidence. In *Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, 513–524. Retrieved from https://www.aclweb.org/anthology/C16-1050.

[173] Samrudhi Sharma, Huda Patanwala, Manthan Shah, and Khushali Deulkar. 2015. A survey of medical question answering systems. *Int. J. Eng. Technic. Res.* 3, 2 (2015), 2321–0869.

[174] Vasu Sharma, Nitish Kulkarni, Srividya Pranavi, Gabriel Bayomi, Eric Nyberg, and Teruko Mitamura. 2018. BioAMA: Towards an end to end BioMedical question answering system. In *Proceedings of the Biomedical Natural Language Processing Workshop (BioNLP)*. Association for Computational Linguistics. 109–117. DOI: https://doi.org/10.18653/v1/W18-2312

[175] Zhongmin Shi, Gabor Melli, Yang Wang, Yudong Liu, Baohua Gu, Mehdi M. Kashani, Anoop Sarkar, and Fred Popowich. 2007. Question answering summarization of multiple biomedical documents. In *Proceedings of the Conference of the Canadian Society for Computational Studies of Intelligence*. Springer, 284–295.

[176] Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. Overview of the NTCIR-11 QA-Lab Task. In *Proceedings of the NTCIR Conference*.

[177] Ana Claudia Sima, Tarcisio Mendes de Farias, Maria Anisimova, Christophe Dessimoz, Marc Robinson-Rechavi, Erich Zbinden, and Kurt Stockinger. 2021. Bio-SODA: Enabling natural language question answering over knowledge graphs without training data. *arXiv preprint arXiv:2104.13744* (2021).

[178] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
[179] Sarvesh Soni, Meghana Gudala, Daisy Zhe Wang, and Kirk Roberts. 2019. Using FHIR to construct a corpus of clinical questions annotated with logical forms and answers. In *AMIA Annual Symposium Proceedings*, Vol. 2019. American Medical Informatics Association.
[180] Sarvesh Soni and Kirk Roberts. 2019. A paraphrase generation system for EHR question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. 20–29.
[181] Sarvesh Soni and Kirk Roberts. 2020. Paraphrasing to improve the performance of electronic health records question answering. *AMIA Summ. Translat. Sci. Proc.* 2020 (2020), 626.
[182] Yash Srivastava, Vaishnav Murali, Shiv Ram Dubey, and Snehasis Mukherjee. 2019. Visual question answering using deep learning: A survey and performance analysis. *arXiv preprint arXiv:1909.01860* (2019).
[183] Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. SNOMED clinical terms: Overview of the development process and project status. In *Proceedings of the AMIA Symposium*. American Medical Informatics Association.
[184] Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu, and Pascale Fung. 2019. Generalizing question answering system with pre-trained language model fine-tuning. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 203–211. DOI: https://doi.org/10.18653/v1/D19-5827
[185] Dan Su, Yan Xu, Tiezheng Yu, Farhad Bin Siddique, Elham Barezi, and Pascale Fung. 2020. CAiRE-COVID: A question answering and query-focused multi-document summarization system for COVID-19 scholarly information management. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*. Association for Computational Linguistics. DOI: https://doi.org/10.18653/v1/2020.nlpcovid19-2.14
[186] Shuo Sun and João Sedoc. 2020. An analysis of BERT FAQ retrieval models for COVID-19 infobot. (2020).
[187] Simon Šuster and Walter Daelemans. 2018. CliCR: A dataset of clinical case reports for machine reading comprehension. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 1551–1563. DOI: https://doi.org/10.18653/v1/N18-1140
[188] Kouji Takahashi, Asako Koike, and Toshihisa Takagi. 2004. Question answering system in biomedical domain. In *Proceedings of the 15th International Conference on Genome Informatics*. Citeseer, 161–162.
[189] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 5100–5111. DOI: https://doi.org/10.18653/v1/D19-1514
[190] Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly bootstrapping a question answering dataset for COVID-19. *arXiv preprint arXiv:2004.11339* (2020).
[191] Rafael M. Terol, Patricio Martínez-Barco, and Manuel Palomar. 2007. A knowledge based method for the medical question answering problem. *Comput. Biol. Med.* 37, 10 (2007), 1511–1521.
[192] Yuanhe Tian, Weicheng Ma, Fei Xia, and Yan Song. 2019. ChiMed: A Chinese medical corpus for question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, 250–260. DOI: https://doi.org/10.18653/v1/W19-5027
[193] George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinform.* 16, 1 (2015), 138.
[194] Christina Unger, Corina Forascu, Vanessa Lopez, Axel-Cyrille Ngonga Ngomo, Elena Cabrio, Philipp Cimiano, and Sebastian Walter. 2014. Question answering over linked data (QALD-4). In *Working Notes for CLEF 2014 Conference*. CEUR-WS.
[195] Hadi Veisi and Hamed Fakour Shandi. 2020. A Persian medical question answering system. *Int. J. Artif. Intell. Tools* 29, 06 (2020), 2050019.
[196] David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 960–966. DOI: https://doi.org/10.18653/v1/P19-1092
[197] Ellen M. Voorhees. 2001. The TREC question answering track. *Nat. Lang. Eng.* 7, 4 (2001), 361–378. DOI: https://doi.org/10.1017/S1351324901002789
[198] Di Wang and Eric Nyberg. 2017. CMU OAQA at TREC 2017 LiveQA: A neural dual entailment approach for question paraphrase identification. In *Proceedings of the Text Retrieval Conference (TREC)*.

[199] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, K. Funk, Rodney Michael Kinney, Ziyang Liu, W. Merrill, P. Mooney, D. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, B. Stilson, Alex D. Wade, Kuansan Wang, Christopher Wilhelm, Boya Xie, D. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 open research dataset. *ArXiv, arXiv:2004.10706v2*.

[200] Ping Wang, Tian Shi, and Chandan K. Reddy. 2020. Text-to-SQL generation for question answering on electronic medical records. In *Proceedings of the Web Conference*. Association for Computing Machinery, New York, NY, 350–361. DOI: https://doi.org/10.1145/3366423.3380120

[201] Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res.* 41, W1 (2013), W518–W522.

[202] Wang Weiming, Dawei Hu, Min Feng, and Liu Wenyin. 2007. Automatic clinical question answering based on UMLS relations. In *Proceedings of the 3rd International Conference on Semantics, Knowledge and Grid (SKG)*. IEEE, 495–498.

[203] Dirk Weissenborn, George Tsatsaronis, and Michael Schroeder. 2013. Answering factoid questions in the biomedical domain. (2013).

[204] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. 2017. Making neural QA as simple as possible but not simpler. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 271–280. DOI: https://doi.org/10.18653/v1/K17-1028

[205] Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Ling.* 6 (2018), 287–302. DOI: https://doi.org/10.1162/tacl_a_00021

[206] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 11–20. DOI: https://doi.org/10.18653/v1/D19-1002

[207] Georg Wiese, Dirk Weissenborn, and Mariana Neves. 2017. Neural domain adaptation for biomedical question answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 281–289. DOI: https://doi.org/10.18653/v1/K17-1029

[208] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 1112–1122. DOI: https://doi.org/10.18653/v1/N18-1101

[209] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.* 163 (2017), 21–40.

[210] Ye Wu, Tak-Wah Lam, Hing-Fung Ting, and Ruibang Luo. 2021. BioNumQA-BERT: Answering biomedical questions using numerical facts with a deep language representation model. In *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*.

[211] Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604* (2016).

[212] Xin Yan, Lin Li, Chulin Xie, Jun Xiao, and Lin Gu. 2019. Zhejiang university at ImageCLEF 2019 visual question answering in the medical domain. In *CLEF (Working Notes)*.

[213] Zi Yang, Niloy Gupta, Xiangyu Sun, Di Xu, Chi Zhang, and Eric Nyberg. 2015. Learning to answer biomedical factoid & list questions: OAQA at BioASQ 3B. (2015).

[214] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 21–29.

[215] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2369–2380. DOI: https://doi.org/10.18653/v1/D18-1259

[216] Zi Yang, Yue Zhou, and Eric Nyberg. 2016. Learning to answer biomedical questions: OAQA at BioASQ 4B. In *Proceedings of the 4th BioASQ Workshop*. Association for Computational Linguistics, 23–37. DOI: https://doi.org/10.18653/v1/W16-3104

[217] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. ABCNN: Attention-based convolutional neural network for modeling sentence pairs. *Trans. Assoc. Comput. Ling.* 4 (2016), 259–272. DOI: https://doi.org/10.1162/tacl_a_00097

[218] Wonjin Yoon, Jinhyuk Lee, Donghyeon Kim, Minbyul Jeong, and Jaewoo Kang. 2019. Pre-trained language model for biomedical question answering. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 727–740.

[219] Hong Yu and Yong-gang Cao. 2008. Automatically extracting information needs from ad hoc clinical questions. In *AMIA Annual Symposium Proceedings*, Vol. 2008. American Medical Informatics Association.

[220] Hong Yu, Minsuk Lee, David Kaufman, John Ely, Jerome A. Osheroff, George Hripcsak, and James Cimino. 2007. Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians. *J. Biomed. Inform.* 40, 3 (2007), 236–251.

[221] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 1821–1830.

[222] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. Neural Netw. Learn. Sys.* 29, 12 (2018), 5947–5959.

[223] Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*. Association for Computational Linguistics, 180–190. DOI: https://doi.org/10.18653/v1/2021.bionlp-1.20

[224] Zheng Yuan, Zhengyun Zhao, Haixia Sun, Jiao Li, Fei Wang, and Sheng Yu. 2021. CODER: Knowledge infused cross-lingual medical term embedding for term normalization. *arXiv:2011.02947 [cs.CL]*.

[225] Xiang Yue, Bernal Jimenez Gutierrez, and Huan Sun. 2020. Clinical reading comprehension: a thorough analysis of the emrQA dataset. *arXiv e-prints*, Article arXiv:2005.00574 (May 2020).

[226] Xiang Yue, Ziyu Yao, Simon Lin, Huan Sun, et al. 2020. CliniQG4QA: Generating diverse questions for domain adaptation of clinical question answering. *arXiv preprint arXiv:2010.16021* (2020).

[227] Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. 2020. Medical visual question answering via conditional reasoning. In *Proceedings of the 28th ACM International Conference on Multimedia*. 2345–2354.

[228] Sheng Zhang, Xin Zhang, Hui Wang, Jiajun Cheng, Pei Li, and Zhaoyun Ding. 2017. Chinese medical question answer matching using end-to-end character-level multi-scale CNNs. *Appl. Sci.* 7, 8 (2017), 767.

[229] Sheng Zhang, Xin Zhang, Hui Wang, Lixiang Guo, and Shanshan Liu. 2018. Multi-scale attentive interaction networks for Chinese medical question answer selection. *IEEE Access* 6 (2018), 74061–74071.

[230] Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[231] Xinliang Frederick Zhang, Heming Sun, Xiang Yue, Emmett Jesrani, Simon Lin, and Huan Sun. 2020. COUGH: A challenge dataset and models for COVID-19 FAQ retrieval. *arXiv preprint arXiv:2010.12800* (2020).

[232] Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao. 2016. A joint model for question answering over multiple knowledge bases. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[233] Yanchun Zhang, S. Peng, R. You, Z. Xie, B. Wang, and Shanfeng Zhu. 2015. The Fudan participation in the 2015 BioASQ challenge: Large-scale biomedical semantic indexing and question answering. In *CEUR Workshop Proceedings*, Vol. 1391. CEUR Workshop Proceedings.

[234] Yingying Zhang, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2019. Multi-modal knowledge-aware hierarchical attention network for explainable medical question answering. In *Proceedings of the 27th ACM International Conference on Multimedia (MM)*. Association for Computing Machinery, New York, NY, 1089–1097. DOI: https://doi.org/10.1145/3343031.3351033

[235] Nikita Zhiltsov, Alexander Kotov, and Fedor Nikolaev. 2015. Fielded sequential dependence model for ad hoc entity retrieval in the web of data. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 253–262.

[236] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of XiaoIce, an empathetic social chatbot. *Comput. Ling.* 46, 1 (2020), 53–93.

[237] Wei Zhou and Clement Yu. 2007. TREC genomics track at UIC. *Resource* 1 (2007), G2.

[238] Ming Zhu, Aman Ahuja, Da-Cheng Juan, Wei Wei, and Chandan K. Reddy. 2020. Question answering with long multiple-span answers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 3840–3849. DOI: https://doi.org/10.18653/v1/2020.findings-emnlp.342

[239] Ming Zhu, Aman Ahuja, Wei Wei, and Chandan K. Reddy. 2019. A hierarchical attention retrieval model for healthcare question answering. In *Proceedings of the World Wide Web Conference (WWW)*. Association for Computing Machinery, New York, NY, 2472–2482. DOI: https://doi.org/10.1145/3308558.3313699

[240] Wei Zhu, Xiaofeng Zhou, K. Wang, X. Luo, Xiepeng Li, Y. Ni, and G. Xie. 2019. PANLP at MEDIQA 2019: Pre-trained language models, transfer learning and knowledge distillation. In *Proceedings of the BioNLP@ACL Conference*.

[241] Pierre Zweigenbaum. 2003. Question answering in biomedicine. *Nat. Lang. Process. Quest. Answer.* 2005 (2003), 1–4.

Received March 2021; revised August 2021; accepted September 2021