Article

What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams

Di Jin [1,*], Eileen Pan [1], Nassim Oufattole [1], Wei-Hung Weng [1], Hanyi Fang [2] and Peter Szolovits [1]

[1] Computer Science and Artificial Intelligence, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; eileenp@mit.edu (E.P.); nassim@mit.edu (N.O.); ckbjimmy@mit.edu (W.-H.W.); psz@mit.edu (P.S.)

[2] Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430074, China; fanghanyi@hust.edu.cn

[*] Correspondence: jindi15@mit.edu

Abstract: Open domain question answering (OpenQA) tasks have been recently attracting more and more attention from the natural language processing (NLP) community. In this work, we present the first free-form multiple-choice OpenQA dataset for solving medical problems, MEDQA, collected from the professional medical board exams. It covers three languages: English, simplified Chinese, and traditional Chinese, and contains 12,723, 34,251, and 14,123 questions for the three languages, respectively. We implement both rule-based and popular neural methods by sequentially combining a document retriever and a machine comprehension model. Through experiments, we find that even the current best method can only achieve 36.7%, 42.0%, and 70.1% of test accuracy on the English, traditional Chinese, and simplified Chinese questions, respectively. We expect MEDQA to present great challenges to existing OpenQA systems and hope that it can serve as a platform to promote much stronger OpenQA models from the NLP community in the future.

Keywords: natural language processing; open-domain question answering; multi-choice question answering; clinical question answering

1. Introduction

Question answering (QA) is a fundamental task in Natural Language Processing (NLP), which requires models to answer a particular question. When given the context text associated with the question, language pre-training based models such as BERT [1], RoBERTa [2], and ALBERT [3] have achieved nearly saturated performance on most of the popular datasets [4–8]. However, real-world scenarios for QA are usually much more complex and one may not have a body of text already labeled as containing the answer to the question. In this scenario, models are required to find and extract relevant information to questions from large-scale text sources such as a search engine [9] and Wikipedia [10]. This type of task is generally called as open-domain question answering (OpenQA), which has recently attracted lots of attention from the NLP community [11–13] but still remains far from being solved.

Most previous works for OpenQA focus on datasets in which answers are in the format of spans (several consecutive tokens) and can be found based on the information explicitly expressed in the provided text [9,10,14,15]. As a more challenging task, free-form multiple-choice OpenQA datasets such as ARC [16] and OpenBookQA [17] contain a significant percentage of questions focusing on the implicitly expressed facts, events, opinions, or emotions in the retrieved text. To answer these questions, models need to perform logical reasoning over the information presented in the retrieved text and in some cases even need to integrate some prior knowledge. Such datasets can motivate a general QA algorithm that can read any given question, find relevant evidence from a knowledge bank, and conduct logical reasoning to obtain the answer. Unfortunately, these OpenQA

datasets consist of questions that require only elementary or middle school level knowledge (e.g., "Which object would let the most heat travel through?"), so even excellent models trained on them may be unable to support more sophisticated real-world scenarios.

To this end, we introduce a new OpenQA dataset, MEDQA, for solving medical problems, representing a demanding real-world scenario. Questions in this dataset are collected from medical board exams in US, Mainland China, and Taiwan, where human doctors are evaluated on their professional knowledge and ability to make clinical decisions. Questions in these exams are varied and generally require a deep understanding of related medical concepts learned from medical textbooks to answer. Table 1 shows two typical examples. An OpenQA model must learn to find relevant information from the large collection of text materials we assembled from medical textbooks, reason over them, and make decisions about the right answer. Taking the first question in Table 1 as instance, to obtain the correct answer, "Chlamydia trachomatis", the OpenQA model first needs to retrieve the relevant evidence as shown in this table from a large collection of medical textbooks, read over the question body, pay attention to the key findings for this patient— no evident signs of *urethritis*, finding of *pyuria*, and positive *leukocyte esterase test*—and infer the correct answer after extensive reasoning.

Table 1. Two examples of MEDQA (Medical Question Answering). The correct answer among options is marked in bold font. Key words in the question and evidence text to help answer the questions are highlighted in italic font. Evidence for both examples are from the textbook "Harrison's Principles of Internal Medicine".

Question	A 27-year-old male presents to urgent care complaining of pain with urination. He reports that the pain started 3 days ago. He has never experienced these symptoms before. He *denies gross hematuria or pelvic pain*. He is sexually active with his girlfriend, and they consistently use condoms. When asked about recent travel, he admits to recently returning from a "boys' trip" in Cancun where he had *unprotected sex* 1 night with a girl he met at a bar. The patient's medical history includes type I diabetes that is controlled with an insulin pump. His mother has rheumatoid arthritis. The patient's temperature is 99 °F (37.2 °C), blood pressure is 112/74 mmHg, and pulse is 81/min. On physical examination, there are no lesions of the penis or other body rashes. No costovertebral tenderness is appreciated. A urinalysis reveals no blood, glucose, ketones, or proteins but is *positive for leukocyte esterase*. A urine microscopic evaluation shows a *moderate number of white blood cells* but no casts or crystals. A urine culture is negative. Which of the following is the most likely cause for the patient's symptoms?
Options	**A: Chlamydia trachomatis**, B: Systemic lupus erythematosus, C: Mycobacterium tuberculosis, D: Treponema pallidum
Evidence	At least one-third of male patients with *C. trachomatis* urethral infection have *no evident signs or symptoms of urethritis*. ... Such patients generally have *pyuria* ..., a *positive leukocyte esterase test*, ...
Question	A 57-year-old man presents to his primary care physician with a 2-month history of *right upper and lower extremity weakness*. He noticed the weakness when he started falling far more frequently while running errands. Since then, he has had *increasing difficulty* with walking and lifting objects. His past medical history is significant only for well-controlled hypertension, but he says that some members of their *family have had musculoskeletal problems*. His right upper extremity shows *forearm atrophy* and *depressed reflexes* while their right lower extremity is *hypertonic with a positive Babinski sign*. Which of the following is most likely associated with the cause of this patient's symptoms?
Options	A: HLA-B8 haplotype, B: HLA-DR2 haplotype, **C: Mutation in SOD1**, D: Mutation in SMN1, E: Viral infection
Evidence	1. The manifestations of ALS ... *insidiously* developing *asymmetric weakness*, usually first evident distally in one of the limbs. 2. ... *hyperactivity of the muscle-stretch reflexes (tendon jerks)* and, often, *spastic resistance to passive movements* ... 3. *Familial ALS (FALS)*... clinically indistinguishable from sporadic ALS... Genetic studies have identified mutations in multiple genes, including cytosolic enzyme *SOD1*...

To provide benchmarks for MEDQA, we implement several state-of-the-art methods for the OpenQA task based on the standard system design, which consists of two components: a document retriever for finding relevant evidence text, and a document reader that performs machine comprehension over the retrieved evidence. Experimental results have shown that even the best method powered by the large pre-trained models [1] can only achieve 36.7%, 42.0%, and 70.1% of test accuracy on the questions collected from US, Mainland China, and Taiwan, respectively, indicating the great challenge of

this dataset. Through both quantitative and qualitative analysis, we find that the performance of document retriever should be the bottleneck since its current form cannot conduct multi-hop reasoning over the retrieving process. Our hope is that MEDQA can serve as a platform to encourage researchers to develop a general OpenQA model that can solve complex real-world questions via abundant logical reasoning in both retrieval and comprehension stages.

2. Related Work

Traditionally, QA tasks have been designed to be text-dependent [4,6,18], where a model is required to comprehend a given text to answer questions. Those given texts relevant to the questions are specially curated by people, which is infeasible for real-world applications where annotations of relevant context are expensive to obtain. This gave rise to the birth of open-domain QA (OpenQA) task, where models must both find and comprehend the context for answering the questions. As a preliminary trial, Chen et al. [10] proposes the DrQA model, which uses a text retriever to obtain relevant documents from Wikipedia, and further applies a trained reading comprehension model to extract the answer from the retrieved documents. Afterwards, researchers have introduced more sophisticated models, which either aggregate all informative evidence [11,12] or filter out those irrelevant retrieved texts [19,20] to better predict the answers. Benefiting from the power of neural networks, these models have achieved remarkable progress in OpenQA.

Much of the previous OpenQA work focuses on the datasets whose answers are spans from the retrieved documents [9,10,14]. In general, most questions concern the facts that are explicitly expressed in the text, offering an advantage to systems that rely mostly on surface word matching [21].

To promote more advanced reading skills, another research line has studied OpenQA tasks in a free-form multiple-choice form [16,17]. These benchmarks allow a more comprehensive evaluation of different higher-level reading skills such as logical reasoning and prior knowledge integration. In particular, real-world exams such as SAT and Gaokao are ideal sources for constructing this kind of OpenQA datasets [16,17,22]. Our proposed dataset follows this line but differs in three aspects:

- The source of our dataset is designed to examine the doctors' professional capability and thus contains a significant number of questions that require multi-hop logical reasoning, which helps push the development of reading comprehension models along this direction [23].
- Our dataset is the first publicly available large-scale multiple-choice OpenQA dataset for the medical problems, where extensive prior domain-specific knowledge is anticipated for the model. It can thus contribute to the emerging field where a general language model will need to be combined with world knowledge.
- Our dataset is cross-lingual, covering English and simplified/traditional Chinese, which contributes to the emerging field of cross-lingual natural language understanding.

There are several related medical QA datasets, which are summarized in Table 2. Among them, LiveQA [24], Medication QA [25], MEDIQA [26], and MedQuAD [27] contain consumer health related questions and the answers are obtained by searching healthcare websites such as MedlinePlus via the ChiQA system [28] and selecting a snippet of searching results out, where the answer retrieval is performed by keywords matching and complex reasoning is seldom involved. BioASQ [29] is similar to the SQuAQ dataset [4], where a span of text in the given context is used as the answer and thus no external knowledge source is needed. emrQA [30] aims to rank a list of Electronic Medical Record (EMR) text lines to find the best line as the answer and the ground truth answers are obtained via an algorithm rather than human annotation. Overall, none of these related datasets have been formulated as an OpenQA problem. Last but not least, Zhang et al. [22] and Ha and Yaneva [31] have previously worked on the Chinese and English versions of our proposed dataset, respectively. However, the former did not release any data while the

latter only released 454 questions for public use, while we publicize a large-scale dataset to promote more powerful deep models.

Table 2. Comparison of our dataset with existing medical QA datasets. In terms of the answer format, "retrieval" means that the answer is a snippet of retrieval result by searching relevant websites; "ranking" means that the answer is a list of candidates and the task is to rank them with higher ranking ones being better answers. Automatic annotation is obtained via an algorithm.

Datasets	LiveQA	Medication QA	BioASQ	emrQA	MEDIQA	MedQuAD	MedQA
Question Source	consumer health	consumer health	biomedical	EMR	consumer health	consumer health	medical&clinical
Answer Format	retrieval	retrieval	span based&binary	ranking	ranking	retrieval	multiple choice
Annotation Method	manual	manual	manual	automatic	manual	automatic	manual
Dataset Size	660	674	3743	455,837	383	47,457	61,097
OpenQA?	No	No	No	No	No	No	Yes

3. Data

3.1. Task Formulation

The task is defined by its three components:

Question: question in text, either in one sentence asking for a certain piece of knowledge, or in a long paragraph starting with a description of the patient condition.

Answer candidates: multiple answer options are given for each question, of which only one should be chosen as the most appropriate.

Document collection: a collection of text material extracted from a variety of sources and organized into paragraphs, which contains the knowledge and information to help find the answers.

This task is to determine the best answer to the question among the candidates, relying on the documents.

3.2. Data Collection

3.2.1. Questions and Answers

We collected the questions and their associated answer candidates from the National Medical Board Examination in the USA (https://www.usmle.org/, Accessed on 10 March 2021), Mainland China (http://www.nmec.org.cn, Accessed on 5 April 2021), and Taiwan (https://wwwq.moex.gov.tw/exam/wFrmExam-QandASearch.aspx, Accessed on 23 March 2021). For convenience, we denote the datasets from these three sources as USMLE, MCMLE, and TWMLE, respectively. These tests assess a physician's ability to apply knowledge, concepts, and principles, and the ability to demonstrate fundamental patient-centered skills. We include problems from both real exams and mock tests; all are freely accessible online for public usage. Details about the sources that we collected data from are described in the Appendix A.

We remove duplicate problems and randomly split the data based on questions, with 80% training, 10% development, and 10% test. The overall statistics of the dataset are summarized in Table 3. To comply with fair use of law (https://www.copyright.gov/fair-use/more-info.html, Accessed on 23 March 2021), we shuffle the order of answer options and randomly delete one of the wrong options for each question for USMLE and MCMLE datasets, which results in four options with one right option and three wrong options. Percentages of each option as the correct answer for the development set are summarized in the Appendix A.

Table 3. Overall statistics of MEDQA. Question/option length and vocabulary/character size are calculated in tokens for English and in characters for Chinese. Vocabulary/character size is measured on the combination of questions and options. "Avg./Max. option len." represents "Average/Maximum option length".

	USMLE	MCMLE	TWMLE
Number of options per question	4	4	4
Avg./Max. option len.	3.5/45	7.3/100	20.6/210
Avg./Max. question len.	116.6/530	45.7/333	61.0/1950
vocab/character size	63,317	3263	3588
Number of questions			
Train	10,178	27,400	11,298
Development	1272	3425	1412
Test	1273	3426	1413
All	12,723	34,251	14,123

3.2.2. Document Collection

Extensive medical knowledge is needed to answer every question in our MEDQA data. For people to obtain answers for these questions, they need to obtain necessary knowledge from a volume of medical textbooks during years of training. Similarly, for a machine learning model to be successful in this task, we need to grant it access to the same collection of text materials as human have. Therefore, for USMLE, we prepared text materials from a total of 18 English medical textbooks that have been widely used by medical students and USMLE takers, whereas for MCMLE, we collected 33 simplified Chinese medical textbooks designated as the official textbooks for preparing the medical licensing exam in Mainland China. For TWMLE, since medical students in Taiwan use the same textbooks as those in USA for exam preparation, USMLE and TWMLE would use the same document collection for solving questions. We will release the textbooks we collected upon the license agreement of research use only.

All textbooks we collected are originally in PDF format and we converted them into digital text via OCR. We performed some clean-up pre-processing over the converted text such as misspelling correction, and then divided all text into paragraphs. Table 4 summarizes the statistics of the document collection.

Table 4. Overall statistics of the document collection. USMLE and TWMLE share the same document collection. Token number and vocabulary size are counted in tokens for English and in characters for Chinese.

Metric	USMLE/TWMLE	MCMLE
# of books	18	33
# of paragraphs	231,581	116,216
# of tokens	12,727,711	14,730,364
Vocabulary size	245,851	4695
Avg./Max. paragraph length	55.0/1234	126.7/9082

To evaluate whether these collected text materials can cover enough knowledge to answer those questions, we randomly extracted 100 questions from the development set of all three datasets and let two medical experts with the MD degree annotate how many of them can be answered by the evidence from our prepared text materials. Table 5 summarizes the results of this evaluation. From this table, we see that our collected text materials can provide enough information for answering most of the questions in our data.

Table 5. Percentage of questions that human experts can find enough evidence from our collected text material to obtain correct answers for by randomly annotating 100 samples from the development set.

	USMLE	MCMLE	TWMLE
Coverage (%)	88.0	100.0	87.0

3.3. Data Analysis

Based on a preliminary analysis of our proposed data, we find that it poses unique challenges for language understanding compared with existing OpenQA datasets, elaborated below:

- **Professional Knowledge:** For most existing QA datasets, the question answering process relies largely on the basic understanding of language and general world knowledge reasoning. Some works have revealed that the large-scale pretrained language models bear a certain level of common-sense and symbolic reasoning capability besides their linguistic knowledge [32,33], which may significantly contribute to the remarkable performance of current QA models. However, the answering of every question in our dataset needs abundant professional domain-specific knowledge, particularly medical knowledge, which forces the model to have a deep understanding of the extracted context.

- **Diversity of Questions:** The field of clinical medicine is diverse enough for questions to be asked about a wide range of topics. In general, there are two categories of questions: (1). The question is asking for a single piece of knowledge, for instance, via the question "Which of the following symptoms belongs to schizophrenia?" (2). The question first describes a patient's condition and then asks for *the most probable diagnosis/the most appropriate treatment/the examination needed/the mechanism of certain conditions/the possible outcome for a certain treatment*, etc. Table 1 shows two typical examples for type 2. Table 6 summarizes the percentages of these two types of questions for each dataset by annotating randomly selected 100 questions from the development set. Typically, type 1 questions need one-step reasoning while type 2 questions require multi-hop reasoning and are thus much more complicated than type 1 ones, imposing challenges not only to the reading comprehension model but also to the relevant text retrieval module. For example, in order to solve the first question in Table 1, the model needs to first extract, understand, and interpret the symptoms of this patient among a long paragraph of description, then match these symptoms to millions of medical knowledge text snippets and find out the most relevant one, and finally understand the evidence sentence for answer selection.

- **Complex Reasoning over Multiple Evidence:** Many questions in our data involve complex multi-hop reasoning over several evidence snippets. For instance, the second example is a typical question that requires multiple steps of reasoning over three evidences, where from the symptoms and signs we can know that this patient is highly likely an ALS case by looking at the evidence 1 and 2. Afterwards, from evidence 3, we know that the SOD1 is the possible genetic mutation for familial ALS, which is the correct answer.

- **Noisy Evidence Retrieval:** Retrieving relevant information from large-scale text is much more challenging than reading a short piece of text. Passages from textbooks often do not directly give answers to questions and many passages retrieved by the most widely adopted term-matching based information retrieval (IR) systems turn out to be noisy distractors and not relevant, especially for type 2 questions. For those questions involving multi-hop reasoning, models must identify all relevant information scattered in different passages, and missing any single piece of evidence would lead to failures.

Table 6. Percentages of type 1 and type 2 questions by annotating 100 randomly samples data from the development set. Type 1 questions ask about a single knowledge point, whereas type 2 questions simulate the realistic clinical settings by studying a patient case.

Question Types	USMLE	MCMLE	TWMLE
Type 1	2.0	75.0	69.0
Type 2	98.0	25.0	31.0

4. Approaches

We implement both classical rule-based methods and recent state-of-the-art neural network based models.

4.1. Rule-Based Methods

We first propose two rule-based methods that do not involve a training process.

4.1.1. Pointwise Mutual Information (PMI)

This method is based on the PMI score function [34], which measures the strength of association between two n-grams x and y and is defined as:

$$PMI(x,y) = log \frac{p(x,y)}{p(x)p(y)},$$

where $p(x,y)$ is the joint probability that x and y occur together in our document collection C, within a certain window of text (we use a 10 word window in both left and right directions); $p(x)$ is the probability that x occurs in C. In practice, we use the frequency to represent the probability. The larger this PMI score, the stronger the association between x and y.

This method extracts unigrams, bigrams, trigrams, and skip-bigrams from the question q and each answer option a_i and calculates the average PMI score over all pairs of question n-grams and answer option n-grams. The answer option with the highest average PMI score will be picked as the prediction.

4.1.2. Information Retrieval (IR)

As the preliminary trial, we adopt a standard off-the-shelf text retrieval system built upon Apache Lucene, Elasticsearch, using inverted index lookup followed by BM25 ranking. Specifically, for each question q and each answer option a_i, we send $q + a_i$ as a query to the search engine, and return the search engine's score for the top-N retrieved sentences. This is repeated for all options to score them all, and the option with the highest score is selected. We denote this version as IR-ES.

To seek better IR performance, we enhance the above using a BM25 re-weighting mechanism inspired by Chen et al. [10]. The updated scoring function is detailed in the Appendix B. Our best performing system uses unigram counts. For English questions, we perform Snowball stemming to both documents and questions; and we use the MetaMap (https://metamap.nlm.nih.gov/, Accessed on 2 April 2020) tool to identify and remove the non-medically-related words out of the questions since they are not useful for medical evidence retrieval. Details of this step can be found in Appendix C. We denote this version as IR-CUSTOM.

4.2. Neural Models

Following the DrQA system by Chen et al. [10], this line of models consists of two components: (1) the Document Retriever module for finding relevant passages and (2) a machine comprehension model, Document Reader, for obtaining the answer by reading the small collection of passages.

4.2.1. Document Retriever

We use the best IR system developed in Section 4.1.2 and obtain the top-N ranked passages from the large-scale document collection C, concatenating them into a long sequence c. Then for each question and option pair $qa_i = q + a_i$, qa_i and c are then passed to the Document Reader for reasoning and decision-making.

4.2.2. Document Reader

We implement the following widely used document reader models:

MAX-OUT:

Following Mihaylov et al. [17], we first use the same bi-directional gated recurrent unit (BiGRU) model to encode both the context c and the question and option pair qa_i, and then perform max-pooling to obtain the final representation vectors $\mathbf{h_c} \in \mathbb{R}^h$ and $\mathbf{h_{qa_i}}$. We then use the following equation to calculate the probability score of how likely this option a_i is to answer this question q given the retrieved context c:

$$\mathbf{h} = [\mathbf{h_c}; \mathbf{h_{qa_i}}; \mathbf{h_c} \cdot \mathbf{h_{qa_i}}; |\mathbf{h_c} - \mathbf{h_{qa_i}}|],$$
$$p(q, a_i|c) = W_1(tanh(W_2\mathbf{h})) \in \mathbb{R}^1,$$

where $[\cdot; \cdot]$ represents the concatenation operation; $W_1 \in \mathbb{R}^{1 \times h}$ and $W_2 \in \mathbb{R}^{h \times 4h}$ are weight matrices to be learned. We compute such a probability score for each option and select the option with the highest score.

Fine-Tuning Pre-Trained Language Models:

We also apply the framework of fine-tuning a pretrained language model such as BERT [1] on our data following Jin et al. [35], Yan et al. [36]. Specifically, we construct the input sequence by concatenating [CLS], tokens in c, [SEP], tokens in qa_i, [SEP], where [CLS] and [SEP] are the classifier token and sentence separator in a pre-trained language model, respectively. We denote the hidden state output by the first token as $\mathbf{h} \in \mathbb{R}^h$ and obtain the unnormalized log probability $p(q, a_i|c) = W\mathbf{h} \in \mathbb{R}^1$, where $W \in \mathbb{R}^{1 \times h}$ is the weight matrix. We obtain the final prediction by applying a softmax layer over the unnormalized log probabilities of all options associated with q and picking the option with the highest probability.

5. Experiments

5.1. Experimental Settings

Dealing with TWMLE Data:

Since TWMLE uses the same document collection as USMLE for solving problems, we translate the questions in TWMLE from traditional Chinese to English via Google Translation and then use the same models as USMLE.

MAX-OUT:

We use spaCy as the English tokenizer and HanLP as the Chinese tokenizer. We use the 200-dimensional word2vec word embeddings induced from PubMed and PMC texts [37] for English text, and use the 300-dimensional Chinese fastText word embeddings [38]. The maximum sequence length for a passage is limited to 450 tokens while that for the question and answer pair is limited to 150 tokens.

Fine-Tuning Pre-Trained Language Models:

We use the following pre-trained language models for Chinese: Chinese BERT-Base (denoted as BERT-BASE-ZH) released by Google [1], Chinese BERT-Base with whole word masking during pre-training over larger corpora (denoted as BERT-BASE-WWM-EXT) [39], multilingual BERT-Base (uncased, denoted as MBERT-BASE) [1], Chinese RoBERTa-Large

with whole word masking over the same corpora as BERT-Base-wwm-ext (denoted as RoBERTa-Large-wwm-ext) [39].

For English, we consider the following pre-trained models: English BERT-Base (denoted as BERT-Base-En) [1], English BioBERT-Base and BioBERT-Large that fine-tune the English BERT models further over the bio-medical literature from PubMed (denoted as BioBERT-Base/Large) [40], English Clinical BERT-Base that fine-tunes the BERT-Base model further over the clinical notes extracted from MIMIC-III (denoted as Clinical-BERT) [41], bio-medical RoBERTa-Base by adapting RoBERTa-Base to bio-medical scientific papers from the Semantic Scholar corpus (denoted as BioRoBERTa-Base) [42], English RoBERTa-Large (RoBERTa-Large) [2]. We did not try further fine-tuning BERT models on our collected textbooks since they only contain 12-15 M tokens, which are far less than the typical corpus size for model pre-training (containing over billions of tokens).

We set the learning rate and effective batch size (product of batch size and gradient accumulation steps) to 2×10^{-5} and 18 for base models, and 1×10^{-5} and 6 for large models. We truncate the longest sequence to 512 tokens after sentence-piece tokenization (we only truncate context). We fine-tune English/Chinese models for 8/16 epochs, set the warm-up steps to 1000, and keep the default values for the other hyper-parameters [1].

5.2. Baseline Results

Tables 7 and 8 summarize the performance of all baselines on the three datasets. By looking at these two tables, we observe:

- Our customized IR system outperforms the off-the-shelf version and the improvement is larger on the English questions.
- For the MCMLE dataset, pretrained models significantly outperform non-pretrained models and all neural models outperform the non-neural baselines. However, for the USMLE and TWMLE datasets, the non-pretrained neural models (Max-out) cannot even surpass the IR baseline and, most surprisingly, most of the pretrained models for the USMLE dataset cannot even beat the IR baseline.

Table 7. Performance of baselines in accuracy (%) on the MCMLE dataset.

Methods	Dev	Test
Chance	25.0	25.0
PMI	36.6	36.9
IR-ES	38.3	37.2
IR-Custom	39.1	37.8
Max-out	51.8	50.9
BERT-Base-Zh	66.5	65.8
BERT-Base-wwm-ext	64.4	64.0
MBERT-Base	62.1	62.3
RoBERTa-Large-wwm-ext	**69.3**	**70.1**

Overall, even the strongest pretrained model (BioBERT-Large, RoBERTa-Large) cannot harvest good scores on any of the three datasets, validating the great challenge of our proposed data. Notably, we did not include human performance for comparison due to the high variance of scores for human examinees (best medical students can earn almost full marks while worse ones cannot pass the exams).

Table 8. Performance of baselines in accuracy (%) on the USMLE and TWMLE datasets.

Datasets	USMLE		TWMLE	
	Dev	**Test**	**Dev**	**Test**
CHANCE	25.0	25.0	25.0	25.0
PMI	29.8	31.1	30.8	31.1
IR-ES	34.0	35.5	24.9	26.8
IR-CUSTOM	**38.3**	36.1	35.1	34.8
MAX-OUT	28.9	28.6	29.4	27.8
BERT-BASE-EN	33.9	34.3	34.3	33.3
CLINICALBERT-BASE	33.7	32.4	33.4	32.1
BIOROBERTA-BASE	35.1	36.1	38.7	36.9
BIOBERT-BASE	34.3	34.1	41.6	41.1
ROBERTA-LARGE	35.2	35.0	39.6	39.3
BIOBERT-LARGE	36.1	**36.7**	**42.2**	**42.0**

5.3. Error Analysis

5.3.1. Quantitative Analysis

Since our approaches to the proposed data involve two stages, i.e., document retrieval and reading comprehension, both stages could be the potential error sources. We first check the performance of the IR-based document retrieval stage by letting two medical doctors annotate whether the top 25 retrieved paragraphs by IR-CUST contain enough evidence for answering the questions over 100 randomly selected samples from the development set. We have three annotation levels: full evidence (evidence is totally complete to derive the answer), partial evidence (evidence is useful but not complete), and no evidence (no evidence is found at all). Table 9 summarizes the results of the annotation. We also summarize the percentages of samples for which we find full, partial, or no evidence individually for type 1 and type 2 questions on the MCMLE and TWMLE datasets. From this table, we see that only the MCMLE dataset has a good retrieval recall while for the majority of samples in USMLE, we cannot find any evidence in the retrieved text. Moreover, among those samples that can find full or partial evidence, we calculate the percentage of samples that find the evidence in the top N (1, 5, 10, 15) retrieved paragraphs, as shown in Table 10. Overall, we have two findings: (1). The MCMLE and TWMLE datasets have much higher percentage of type 1 questions than USMLE, which is positively correlated with their much better IR retrieval performance; (2). Such poor evidence retrieval performance for the USMLE dataset should be the main cause of the extremely low baseline models' performance as shown in Table 8 (neural models are even beaten by the IR baseline).

Table 9. Percentage of samples that can find full evidence, partial evidence, or no evidence in the top 25 retrieved paragraphs. For numbers within parenthesis, the left number is for type 1 while the right number is for type 2 questions. Since almost all questions are type 2 for USMLE, we do not have percentage numbers specifically for type 1 and 2 questions.

Types	USMLE	MCMLE	TWMLE
Full Evidence	24.0	75.0 (82.7/52.0)	60.0 (63.4/52.6)
Partial Evidence	8.0	21.0 (14.7/40.0)	16.7 (19.5/10.5)
No Evidence	68.0	4.0 (2.6/8.0)	23.3 (17.1/36.9)

Table 10. Percentage of samples that can find evidence in the top *N* (1, 5, 10, 15) retrieved paragraphs among samples that can find full or partial evidence.

Datasets	Top 1	Top 5	Top 10	Top 15
USMLE	0.0	31.2	56.2	81.2
MCMLE	66.7	92.7	96.9	100.0
TWMLE	0.0	71.7	91.3	93.5

5.3.2. Qualitative Analysis

Seeing such poor retrieval performance on the USMLE dataset, we wonder what could be the reasons. After taking a close look into the successful and failed samples in the retrieval stage, we can summarize one success pattern as well as two failure patterns for the majority of cases, described in the sections below. Remember that almost all questions in USMLE are about case studies (type 2 questions). Furthermore, the IR system always returns us snippets with each of which matching to only a small portion of the question text.

Success Patterns:

The question asks about the most probable diagnosis, which involves only one step of reasoning (inference from symptoms/signs/findings to diagnosis of diseases), and it is easy to constrain the possible diagnosis candidates to one or two based on some special condition terms. In this case, it is easy for the IR system to obtain useful evidence snippets by matching those key condition terms. For example, given a case of a stressful young female patient suffering from *recurrent headaches* that *alternately affect the right or left brain and are exacerbated by loud sounds or bright light*, along with *nausea*, the IR system can identify these typical, condition-specific terms (highlighted by italic font) and retrieve the suitable evidence about the disease of *migraine headache*.

Failure Patterns:

(1). The question still asks about the most probable diagnosis; however, each of the patient's symptoms is very common and could correspond to many possible diagnoses. In this case, the IR system may return miscellaneous diseases' descriptions, each of which can match part of the patient's symptoms, signs or other findings. However, none of these retrieved texts are relevant to the correct diagnosis. We provide one example in the Appendix D. (2). The question asks about the *the most appropriate treatment/the examination needed/the mechanism of certain condition*, etc., which all involve two-steps of reasoning. We need to first derive the diagnosis based on the patient's condition and then answer this question based on the inferred diagnosis. In this case, it is highly likely that the IR system only return evidence that enables the first step of reasoning (making the diagnosis) but not the second one. We also provide an example in the Appendix D.

Last but not the least, we did a qualitative analysis of the translation quality from traditional Chinese to English for TWMLE questions during experiments and found that most of the sentences are good enough while a few of them look not that fluent. However, since all medical terms in TWMLE questions are originally in English, such minor lack of fluency would not affect evidence retrieval and question answering.

6. Conclusions

We present the first open-domain multiple-choice question answering dataset for solving medical problems, MEDQA, collected from the real-world professional examinations. This dataset covers three languages: English, simplified Chinese, and traditional Chinese, and it requires extensive and advanced domain knowledge to answer questions. Together with the question data, we also collect and release a large-scale corpus from medical textbooks from which the reading comprehension models can obtain necessary knowledge for answering the questions. We implement several classic as well as state-of-

the-art methods as baselines to this dataset by cascading two steps: document retrieval and reading comprehension. Furthermore, experimental results demonstrate that even current high-performing approaches based on large-scale pre-trained models cannot achieve good performance on these data. For example, the best of our implemented models can achieve less than 45% of accuracy on the test sets of USMLE and TWMLE datasets. We anticipate more research efforts from the community can be devoted to this dataset so that future OpenQA models can be strong enough to solve such real-world complex problems.

Appendix A. Data Collection

For USMLE and MCMLE datasets, we scraped several websites that provide the question banks, and for TWMLE dataset, we downloaded the examination materials in PDF format directly from the official National Examination website in Taiwan and then converted them into digital format via Optical Character Recognition (OCR). Table A1 lists the websites that we obtained data from.

Table A1. Websites that we collected our data from.

Datasets	Websites
USMLE	https://step1.medbullets.com https://www.amboss.com/us/usmle https://www.lecturio.com/usmle-step-1
MCMLE	http://www.offcn.com/yixue/linchuang
TWMLE	https://wwwq.moex.gov.tw/exam/wFrmExamQandASearch.aspx

Table A2. Percentages of each option as the correct answer for the development sets.

Options	USMLE	MCMLE	TWMLE
A	22.5	25.9	20.3
B	27.1	24.8	25.8
C	26.1	27.7	29.0
D	24.3	21.5	24.9

Appendix B. Information Retrieval (IR)

In this section, we describe how the BM25 re-weighted BM25 scoring function works using the following equation:

$$s(Q, D) = \sum_{i=1}^{n} \frac{BM_{25}(q_i, D) \cdot IDF(q_i) \cdot f(q_i, Q) \cdot (k_Q + 1)}{f(q_i, Q) + k_Q \cdot (1 - b_Q + b_Q \cdot \frac{queryLen}{avgQueryLen})},$$

$$BM_{25}(q_i, D) = \frac{IDF(q_i) \cdot f(q_i, D) \cdot (k_D + 1)}{f(q_i, Q) + k_D \cdot (1 - b_D + b_D \cdot \frac{docLen}{avgDocLen})},$$

where $Q = \{q_1, q_2, ..., q_n\}$ is the query consisting of n query terms (each query term in our method is actually a n-gram); $IDF(q_i)$ the inverse document frequency of the query term q_i; $f(q_i, Q)$ and $f(q_i, D)$ represents the frequency of query term q_i in the query Q and the document D, respectively; k_D, k_Q, b_D, and b_Q are hyper-parameters and their values in our experiments are summarized in Table A3; *queryLen* and *docLen* are the length in tokens of the current query Q and document D, respectively; *avgQueryLen* and *avgDocLen* are the average length of all queries and documents.

Table A3. Hyper-parameters for the IR system tuned on the development set.

Datasets	k_Q	b_Q	k_D	b_D
USMLE	0.40	0.70	0.90	0.35
MCMLE	1.10	0.20	0.90	0.35
TWMLE	0.90	1.00	0.90	0.35

Appendix C. MetaMap for IR

Since type 2 questions are long and contain many words that are not useful for retrieving the relevant context, we adopt the MetaMap tool to filter out those unwanted words. MetaMap was developed by the National Library of Medicine (NLM) to map biomedical text to concepts in the Unified Medical Language System (UMLS). It uses a hybrid approach combining a natural language processing (NLP), knowledge-intensive approach and computational linguistic techniques. We can use it to extract and standardize medical concepts from any biomedical/clinical text. Table A4 shows one example by sending one question to MetaMap API and parsing the returned result, where the bold font highlighted words are identified medically-related entities and we concatenate them together as the processed question for retrieval.

Table A4. An example showing the extracted medically-related entities (highlighted in bold font) by MetaMap.

Question
A **61-year-old man** presents to the **emergency department** because he has developed **blisters** at **multiple locations** on their **body**. He says that the **blisters** appeared **several days ago after** a day of hiking in the **mountains with** his colleagues. When asked **about potential** triggering **events**, he says that he recently had an **infection** and was treated **with antibiotics** but he cannot recall the **name** of the **drug** that he took. In addition, he accidentally confused his **medication with one** of his wife's **blood thinner pills several days before** the **blisters** appeared. On **examination**, the **blisters** are **flesh-colored**, raised, and **widespread** on their **skin** but do **not** involve their **mucosal surfaces**. The **blisters** are **tense** to **palpation** and do **not separate with rubbing**. **Pathology** of the **vesicles** show that they continue **under the level of the epidermis**. Which of the following is the most **likely cause** of this **patient's blistering**?

Appendix D. Failure Patterns of IR

In Table A5, we showcase examples for each of the two failure patterns mentioned in the main text.

In this table, the first example illustrates the first failure pattern, where the model fails to perform the correct one-step reasoning for disease diagnosis since the symptoms and patient history mentioned in the case description are very common symptoms (e.g., cough with sputum, increased urinary frequency) and are not disease- or condition-specific. Although the better option for this case is to examine the patient's abdomen due to their age and the lacking information of their abdominal condition, the IR system can only

return some miscellaneous evidences about (1) a similar case, (2) lung cancer and smoking, (3) urinary tract infection, and (4) cholesterol issue, which can not actually answer the given question.

The second example for explaining the second failure pattern is a case about a mother with gestational diabetes/hypertension and her baby. In this case, the IR system can identify the medical condition of the mother and retrieve relevant information to this mother. However, the question is actually asking about the baby, which is only described in the very end of this question. Unfortunately, the IR system only returns the evidence for the first-step reasoning (mother's diagnosis) but not for the second-step (diagnose the baby's disease based on the mother's diagnosis).

Table A5. Two examples of failure patterns for the IR system. The first example demonstrates the failure of capturing the correct one-step reasoning, and the second example shows the failure of focusing on the question target while correctly identifying the medical condition of the given case. The correct option is marked by the bold font. The top 6 retrieved paragraphs are shown here.

Question	A 67-year-old man presents to a primary care clinic to establish care after moving from another state. According to his prior medical records, he last saw a physician 4 years ago and had no significant medical problems at that time. Records also show a normal EKG and normal colonoscopy results at that time. The patient reports feeling well overall, but review of systems is positive for 1 year of mild cough productive of clear sputum and 2 years of increased urinary frequency. He denies fever, chills, dyspnea, dysuria or hematuria. He denies illicit drug use but has been drinking approximately 1-2 beers per night and smoking 1 pack of cigarettes per day since age 20. Physical exam is unremarkable. Which of the following tests is indicated at this time?
Options	**A: Abdominal ultrasound**, B: Bladder ultrasound, C: Colonoscopy, D: Serum prostate specific antigen (PSA) testing, E: Sputum culture
Evidence	1. ... This patient is a 67-year-old man with weight loss of 10 pounds in 4 weeks and a 35 pack-year history of cigarette smoking. He quit smoking 10 years ago. He had left shoulder pain for 4 months with no dyspnea, cough, hemoptysis, or other symptoms. Massage and other musculoskeletal manipulation did not improve his symptoms. ... 2. Lung cancer, which was rare prior to 1900 with fewer than 400 cases described in the medical literature, is considered a disease of modern man. ... Tobacco consumption is the primary cause of lung cancer, ... Given the magnitude of the problem, it is incumbent that every internist has a general knowledge of lung cancer and its management. 3. The symptoms and signs of UTI vary markedly with age. ... 4. Cigarette smoking is a well-established risk factor in men and probably accounts for the increasing incidence and severity of atherosclerosis in women. ... 5. Smoking within 30 min of waking, smoking daily, smoking more cigarettes per day, and waking at night to smoke are associated with tobacco use disorder. ... Serious medical conditions, such as lung and other cancers, cardiac and pulmonary disease, perinatal problems, cough, shortness of breath, and accelerated skin aging, often occur. 6. Alcohol and cigarette smoking are well-known modifiers of cholesterol. ...
Question	A 37-year-old G1P1001 delivers a male infant at 9 pounds 6 ounces after a C-section for preeclampsia with severe features. The mother has a history of type II diabetes with a hemoglobin A1c of 12.8% at her first obstetric visit. Before this pregnancy, she was taking metformin, and during this pregnancy, she was started on insulin. At her routine visits, her glucose logs frequently showed fasting fingerstick glucoses above 120 mg/dL and postprandial values above 180 mg/dL. In addition, her routine third trimester culture for group B Streptococcus was positive. At 38 weeks and 4 days gestation, she was found to have a blood pressure of 176/103 mmHg and reported a severe headache during a routine obstetric visit. She denied rupture of membranes or vaginal bleeding. Her physician sent her to the obstetric triage unit, and after failure of several intravenous doses of labetalol to lower her blood pressure and relieve her headache, a C-section was performed without complication. Fetal heart rate tracing had been reassuring throughout her admission. Apgar scores at 1 and 5 min were 7 and 10. After one hour, the infant is found to be jittery; the infant's temperature is 96.1 °F (35.6 °C), blood pressure is 80/50 mmHg, pulse is 110/min, and respirations are 60/min. When the first feeding is attempted, he does not latch and begins to shake his arms and legs. After 20 s, the episode ends and the infant becomes lethargic. Which of the following is the most likely cause of this infant's presentation?

Table A5. *Cont.*

Options	A: Transplacental action of maternal insulin, **B: β-cell hyperplasia**, C: Neonatal sepsis, D: Inborn error of metabolism, E: Neonatal encephalopathy
Evidence	1. At Parkland Hospital, women with diabetes are seen in a specialized obstetrical clinic every 2 weeks. During these visits, glycemic control records are evaluated and insulin adjusted. ... 2. Once pregnancy is established, glucose control should be managed more aggressively than in the nonpregnant state. In addition to dietary changes, this enhanced management requires more frequent blood glucose monitoring and often involves additional injections of insulin or conversion to an insulin pump. ... 3. 2 h postprandial values of 100 to 120 mg/dL, and mean the action profiles of commonly used short- and long-term daily glucose concentrations $<$ 110 mg/dL. ... 4. The gold standard for diagnosis of insulinoma is the 72 h monitored fast. ... 5. Treatment of gestational diabetes with a two-step strategy—dietary intervention followed by insulin injections if diet alone does not adequately control blood sugar ... 6. The rate of vertical transmission is reduced to less than 8% by chemoprophylaxis with a regimen of zidovudine to the mother (100 mg five times/24 h orally) started by 4 weeks gestation, ...

References

1. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 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, Minneapolis, MN, USA, 2–7 June 2019; pp. 4171–4186. [CrossRef]
2. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
3. Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; Soricut, R. Albert: A lite bert for self-supervised learning of language representations. *arXiv* **2019**, arXiv:1909.11942.
4. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, TX, USA, 1–5 November 2016; pp. 2383–2392. [CrossRef]
5. Rajpurkar, P.; Jia, R.; Liang, P. 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, Melbourne, Australia, 15–20 July 2018; pp. 784–789. [CrossRef]
6. Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; Hovy, E. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing; Association for Computational Linguistics, Copenhagen, Denmark, 7–11 September 2017; pp. 785–794. [CrossRef]
7. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 31 October–4 November 2018; pp. 2369–2380. [CrossRef]
8. Gao, S.; Agarwal, S.; Jin, D.; Chung, T.; Hakkani-Tur, D. From Machine Reading Comprehension to Dialogue State Tracking: Bridging the Gap. In Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI, Association for Computational Linguistics, Online, 9 July 2020; pp. 79–89. [CrossRef]
9. Dunn, M.; Sagun, L.; Higgins, M.; Güney, V.U.; Cirik, V.; Cho, K. SearchQA: A New QA Dataset Augmented with Context from a Search Engine. *arXiv* **2017**, arXiv:1905.05733.
10. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. 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, Vancouver, BC, Canada, 2017; pp. 1870–1879. [CrossRef]
11. Clark, C.; Gardner, M. Simple and Effective Multi-Paragraph Reading Comprehension. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers); Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 845–855. [CrossRef]
12. Wang, Z.; Ng, P.; Ma, X.; Nallapati, R.; Xiang, B. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. *arXiv* **2019**, arXiv:1908.08167.
13. Asai, A.; Hashimoto, K.; Hajishirzi, H.; Socher, R.; Xiong, C. Learning to Retrieve Reasoning Paths over Wikipedia Graph for Question Answering. *arXiv* **2020**, arXiv:1911.10470.
14. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. 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, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1601–1611. [CrossRef]
15. Dhingra, B.; Mazaitis, K.; Cohen, W.W. Quasar: Datasets for Question Answering by Search and Reading. *arXiv* **2017**, arXiv:1707.03904.

16. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457.

17. Mihaylov, T.; Clark, P.; Khot, T.; Sabharwal, A. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 31 October–4 November 2018; pp. 2381–2391. [CrossRef]

18. Reddy, S.; Chen, D.; Manning, C.D. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 249–266. [CrossRef]

19. Wang, S.; Yu, M.; Guo, X.; Wang, Z.; Klinger, T.; Zhang, W.; Chang, S.; Tesauro, G.; Zhou, B.; Jiang, J. R3: Reinforced Ranker-Reader for Open-Domain Question Answering. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

20. Das, R.; Dhuliawala, S.; Zaheer, M.; McCallum, A. Multi-step Retriever-Reader Interaction for Scalable Open-domain Question Answering. *arXiv* **2019**, arXiv:1905.05733.

21. Sun, K.; Yu, D.; Yu, D.; Cardie, C. Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. *Trans. Assoc. Comput. Linguist.* **2019**, *8*, 141–155. [CrossRef]

22. Zhang, X.; Wu, J.; He, Z.; Liu, X.; Su, Y. Medical exam question answering with large-scale reading comprehension. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

23. Yu, W.; Jiang, Z.H.; Dong, Y.; Feng, J. ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. *arXiv* **2020**, arXiv:2002.04326.

24. Abacha, A.B.; Agichtein, E.; Pinter, Y.; Demner-Fushman, D. *Overview of the Medical Question Answering Task at TREC 2017 LiveQA*; TREC: Gaithersburg, MD, USA, 2017.

25. Abacha, A.B.; Mrabet, Y.; Sharp, M.; Goodwin, T.R.; Shooshan, S.E.; Demner-Fushman, D. *Bridging the Gap between Consumers' Medication Questions and Trusted Answers*; MedInfo: Westminster, CO, USA, 2019; pp. 25–29.

26. Ben Abacha, A.; Shivade, C.; Demner-Fushman, D. 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, Florence, Italy, 1 August 2019; pp. 370–379. [CrossRef]

27. Abacha, A.B.; Demner-Fushman, D. A question-entailment approach to question answering. *BMC Bioinf.* **2019**, *20*, 511.

28. Demner-Fushman, D.; Mrabet, Y.; Ben Abacha, A. Consumer health information and question answering: Helping consumers find answers to their health-related information needs. *J. Am. Med. Inf. Assoc.* **2020**, *27*, 194–201. [CrossRef]

29. Nentidis, A.; Krithara, A.; Bougiatiotis, K.; Krallinger, M.; Rodriguez-Penagos, C.; Villegas, M.; Paliouras, G. Overview of BioASQ 2020: The Eighth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. In Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages, Thessaloniki, Greece, 22–25 September 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 194–214.

30. Pampari, A.; Raghavan, P.; Liang, J.; Peng, J. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 31 October–4 November 2018; pp. 2357–2368. [CrossRef]

31. Ha, L.A.; Yaneva, V. Automatic Question Answering for Medical MCQs: Can It go Further than Information Retrieval? In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), Varna, Bulgaria, 2–4 September 2019; INCOMA Ltd.: Varna, Bulgaria, 2019; pp. 418–422. [CrossRef]

32. Talmor, A.; Elazar, Y.; Goldberg, Y.; Berant, J. oLMpics—On what Language Model Pre-training Captures. *Trans. Assoc. Comput. Linguist.* **2019**, *8*, 743–758. [CrossRef]

33. Petroni, F.; Rocktäschel, T.; Lewis, P.; Bakhtin, A.; Wu, Y.; Miller, A.H.; Riedel, S. Language Models as Knowledge Bases? *arXiv* **2019**, arXiv:1909.01066.

34. Clark, P.; Etzioni, O.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P.; Khashabi, D. Combining retrieval, statistics, and inference to answer elementary science questions. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.

35. Jin, D.; Gao, S.; Kao, J.Y.; Chung, T.; Hakkani-Tür, D.Z. MMM: Multi-stage Multi-task Learning for Multi-choice Reading Comprehension. *arXiv* **2019**, arXiv:1910.00458.

36. Yan, M.; Zhang, H.; Jin, D.; Zhou, J.T. Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 5–10 July 2020; pp. 7331–7341. [CrossRef]

37. Moen, S.; Ananiadou, T.S.S. Distributional semantics resources for biomedical text processing. *Proc. LBM* **2013**, 39–44.

38. Bojanowski, P.; Grave, E.; Joulin, A.; Mikolov, T. Enriching Word Vectors with Subword Information. *Trans. Assoc. Comput. Linguist.* **2017**, *5*, 135–146. [CrossRef]

39. Cui, Y.; Che, W.; Liu, T.; Qin, B.; Yang, Z.; Wang, S.; Hu, G. Pre-Training with Whole Word Masking for Chinese BERT. *arXiv* **2019**, arXiv:1906.08101.

40. Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C.H.; Kang, J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **2019**. [CrossRef]

41. Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.H.; Jin, D.; Naumann, T.; McDermott, M. Publicly Available Clinical BERT Embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, MN, USA, 7 June 2019; pp. 72–78. [CrossRef]

42. Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; Smith, N.A. Do not Stop Pretraining: Adapt Language Models to Domains and Tasks. *arXiv* **2020**, arXiv:2004.10964.