International Journal on Digital Libraries (2022) 23:289–301
https://doi.org/10.1007/s00799-022-00329-y

ScienceQA: a novel resource for question answering on scholarly articles

Tanik Saikh[1] · Tirthankar Ghosal[2] · Amish Mittal[1] · Asif Ekbal[1] · Pushpak Bhattacharyya[3]

Received: 14 May 2021 / Revised: 9 June 2022 / Accepted: 15 June 2022 / Published online: 20 July 2022

Abstract

Machine Reading Comprehension (MRC) of a document is a challenging problem that requires discourse-level understanding. Information extraction from scholarly articles nowadays is a critical use case for researchers to understand the underlying research quickly and move forward, especially in this age of infodemic. MRC on research articles can also provide helpful information to the reviewers and editors. However, the main bottleneck in building such models is the availability of human-annotated data. In this paper, firstly, we introduce a dataset to facilitate question answering (QA) on scientific articles. We prepare the dataset in a semi-automated fashion having more than 100k human-annotated context–question–answer triples. Secondly, we implement one baseline QA model based on Bidirectional Encoder Representations from Transformers (BERT). Additionally, we implement two models: the first one is based on Science BERT (SciBERT), and the second is the combination of SciBERT and Bi-Directional Attention Flow (Bi-DAF). The best model (i.e., SciBERT) obtains an F1 score of 75.46%. Our dataset is novel, and our work opens up a new avenue for scholarly document processing research by providing a benchmark QA dataset and standard baseline. We make our dataset and codes available here at https://github.com/TanikSaikh/Scientific-Question-Answering.

Keywords Machine reading comprehension · Scholarly articles · BERT · SciBERT · BiDAF · Question answering · Automatic article review system

1 Introduction

With the deluge of research articles, it is increasingly getting difficult for researchers to stay abreast with the latest development in science and prior works. It is arduous for researchers to read all the papers, even in their specific domain of interest. The problem becomes more pressing for medical practitioners who want to consume the latest scientific information in medicine and biology but struggle with time from their critical daily duties, even for senior academics who are always loaded with diverse commitments and responsibilities. Extracting desired and meaningful information from scientific articles is an essential yet time-consuming practice for all researchers. With the recent progress in machine learning (ML) and natural language processing (NLP) for scholarly document processing [2], researchers have started leveraging state-of-the-art advancements to mine and process scholarly documents to derive actionable insights. With the exponential rise in research articles nowadays, it seems a natural direction to employ state-of-the-art NLP/ML techniques to help researchers counter the scholarly information over-

✉ Tanik Saikh
1821cs08@iitp.ac.in

✉ Asif Ekbal
asif@iitp.ac.in

Tirthankar Ghosal
ghosal@ufal.mff.cuni.cz

Amish Mittal
1801cs07@iitp.ac.in

Pushpak Bhattacharyya
pb@cse.iitb.ac.in

[1] Department of Computer Science and Engineering, Indian Institute of Technology Patna, Bihta, Patna, India

[2] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Malostranské náměstí 25, 118 00 Praha, Czech Republic

[3] Department of Computer Science and Engineering, Indian Institute of Technology Bombay, Powai, Mumbai, Maharashtra, India

load by assisting them to understand the scientific discourse quickly. However, automatically comprehending scholarly discourse is not straightforward, primarily due to the embedded background knowledge in a scientific article. MRC is one such NLP task that aims to make machines understand a discourse to perform downstream problems. Understanding a text document depends on how much one can efficiently answer questions based on that text document. There has been ongoing research on MRC on various textual documents like Wikipedia articles [37,59], news articles [7,47], etc. However, there has not been a significant attempt for QA in scientific articles except a few, such as the one proposed for biomedical articles [34]. One reason is the lack of high-quality datasets to perform MRC on scholarly papers. To address this issue, in our current work, we introduce *ScienceQA*, a dataset for QA on scholarly articles, which is the first attempt in this direction to the best of our knowledge.

1.1 Problem statement

The Reading Comprehension (RC) task is similar to Question Answering (QA). RC could be thought of as an instance of QA as it is essentially a QA problem over a short passage of text. We view MRC as a supervised learning problem in a machine learning platform: given a collection of training examples $\{(d_i, q_i, a_i)\}_{i=1}^{n}$, the aim is to learn a predictor function f that takes a document of text d and an associated question q as inputs and yields the answer a as output:

$$f : (d, q) \longrightarrow a \quad (1)$$

In this article, we consider the abstracts only, which we treat as a document, a kind of proxy for the full-text paper. The above learning prediction is for a single document. We could extend this task to a multi-document setting where d is a set of documents or break it into smaller linguistic units such as sentences [10]. We show one example from an article and its possible question–answer pairs in Table 1.

This task on scholarly articles is more challenging compared to general domain texts. It is even difficult for human experts to understand a research paper after one reading. On the other hand, MRC on scientific articles can help researchers quickly comprehend the research presented and pull up the desired information according to their needs.

As the number of such articles increases rapidly, collecting and assembling information from these articles become more and more difficult. Information retrieval or extraction, text mining, and automatic QA solutions can help mitigate this problem.

We could find only two prior research in this domain. The first one is by Kim et al. [26] on bio-medical texts, whereas the second one [35] is concentrated on scientific articles. Both of these systems are on *cloze style (Fill-in the blank)* type MRC model. According to Chen et al. [6], depending on the type of answers, MRC models could be divided into four major types, *viz.,* (i). cloze style, (ii). multiple-choice, (iii). span prediction, and (iv). free-form answer. In cloze style, the questions have a placeholder; systems have to identify a word or entity that would be the best suitable to complete the sentence or question. The answer is chosen from the predefined choices or the whole vocabulary. A few notable systems in this type of task are Hermann et al. [20], Chen et al. [7], Šuster et al. [45], and Dhingra et al. [16]. In multiple-choice MRC, systems have to choose a correct answer from a set of hypothesized answers. Many models have been built for this type of task, to name a few: Huang et al. [23], Lai et al. [29], and Richardson et al. [39]. Span prediction-based QA systems [20,21,25,37] require extract a segment of text from the corresponding supporting passage, as it is a kind of extractive QA, where systems need to identify start and end indices of the reading passage.

Free-form QA [27] finds an answer that could be any free form text (i.e., a word sequence of any arbitrary length). The current work focuses on the extractive span-of-texts-based QA system and the dataset is also prepared. Recently, Saikh et al. [41] have created 10K *context–question–answer* triples on scholarly articles for this kind of task and dubbed it as *ScholarlyRead*. *ScholarlyRead* was created in a semi-automated way, and it is a benchmark resource for QA on scholarly articles. *ScholarlyRead* is small in size and lies in a close domain. In contrast, *ScienceQA* is a large-scale, open-domain dataset.

1.2 Motivation

The final goal of our continuing research is to build an AI-assisted peer review system consisting of several modules, use-cases for the editors, reviewers, and authors. A QA tool built on our current research can be part of this AI-assisted peer review pipeline. Both editors and reviewers can benefit from this proposed system as they can query the literature under review to retrieve the relevant information. Also an inquisitive reader would be able to ask questions before going into the details of the paper. Some very obvious questions could be:

- *What problem does the article strive to solve?*
  e.g., in Table 1, the authors present two algorithms for solving CUMULATIVE constraint based on a new energetic relaxation.
- *What is the approach adopted?*
  e.g., in the abstract contained in Table 1, the authors took two novel filtering algorithms to solve the problem.
- *What motivates the authors?*
  For this question, the system has to consider the Introduction Section.

Table 1 Question–Answer pairs for a sample abstract in the ScienceQA dataset. Each of the answers is a span-of-texts from the abstract

Abstract: We present *two novel filtering algorithms* for the CUMULATIVE constraint based on a new energetic relaxation. We introduce a *a generalization of the Overload Check and Edge-Finder rules based on a function computing the earliest completion time for a set of tasks.* Depending on the relaxation used to compute this function, one obtains different levels of filtering. We present *two algorithms that enforce these rules*. The algorithms utilize a novel data structure that we call Profile and that encodes the resource utilization over time. Experiments show that these algorithms are competitive with the state-of-the-art algorithms, by doing a greater filtering and having a faster runtime.
Q1: What did authors present for the CUMULATIVE constraint based on a new energetic relaxation?
Q2: What does the paper introduce?
Q3: What does the paper present?

- *What is (are) the benchmark dataset(s) on which the evaluation is performed?*
  e.g., from Table 1, we can say that the paper makes use of the *Profile dataset* for evaluating the system.
- *Do the results outperform the existing state-of-the-arts?*
  The answer to this question lies in the last sentence of the abstract.
- *Are the results obtained by the proposed system/s noteworthy?*
  The answer to this question lies in the last sentence of the abstract too.

Based on the answers to these crucial questions, the editor would get an idea about the research article, enabling them to make an appropriate decision, i.e., either desk-rejecting it or forwarding it to the next level and assigning it the appropriate reviewers. This way, editors can assign reviewers better than the existing approach, where the reviewers are assigned based on the authors' given keywords.

1.3 Contributions

The key contributions of this article are *threefold* as follows:

- *We have created more than 100k data points, i.e., context–question–answer triplets for training data-hungry deep neural network-based MRC models on scholarly articles. This dataset will serve as the benchmark evaluation of Question Answering (QA) and Question Generation (QG) on scholarly articles. We are unaware of any such dataset in such a large volume.*
- *We fine-tune the answers of the dataset by the outputs of our implemented SciBERT model. It seems that the answers produced by this model are pretty promising.*
- *We have proposed several models based on BERT, SciBERT, and SciBERT combined with BiDAF.*

The rest of the paper is organized as follows. Section 2 describes a literature survey (i.e., existing datasets and models) for MRC. We describe our dataset preparation process, evaluation scheme, and data annotation guidelines in Sect. 3. Section 4 describes our baseline and proposed models. This section is followed by Sect. 5, which contains the details of the experiments carried out, results obtained, comparison and discussion. Section 5.1 shows the error analysis. We conclude the paper in Sect. 6 and point to some directions to future works.

2 Related work

The deep learning era has attracted lots of attention in this field of MRC. *Large-scale datasets* and *end-to-end neural RC models* have been the key attributes in the success of this problem. On the one hand, creating these large-scale datasets have made it possible to train data-hungry deep neural models. On the other hand, a clear understanding of existing models helps us identify these models' limitations and motivates us to develop further. Recent interest in MRC/QA has introduced several datasets as shown below:

Deep Read: Hirschman et al. [22] had curated a corpus consisting of 60 stories each for the development and test set of 3rd- to 6th-grade materials. They provided *Deep Read*. This automated MRC system uses pattern-matching algorithms that rely on shallow linguistic processing (i.e., stemming, named entity, semantic class identification, pronoun resolution, etc.).

MCTest: MCTest [39] is an open-domain MCQ type RC tasks' dataset that aims at the 7/8 years old RC level. The documents are in the form of fictional stories. They defined two baseline models based on lexical level features.

CNN/Daily Mail: Chen et al. [7] thoroughly examined the CNN Daily mail dataset [20], consisting of over a million examples. This dataset was created by parsing the CNN/Daily mail's news contents for the cloze style MRC systems, where the questions are constructed from bullet point summaries by blanking out a word or phrase. The model they developed is based on the Attentive Reader [20].

Children Book Test: Children Book Test (CBT) [21] was constructed in a similar spirit to *CNN/Daily Mail*. The document was created taking the first 20 sentences of a children's book, aiming to infer the missing word in the 21st sentence. They also categorized the questions based on the type (i.e., named entity, common noun, preposition, or verb) of missing words. Their model is based on memory networks.

bAbI: Weston et al. [53] proposed this artificial dataset, comprising 20 different reasoning types. The bAbI motivates building the model to capture multi-sentence reasoning to answer a particular question. They proposed a framework and a set of synthetic tasks for text understanding and reasoning.

SQuAD: Stanford Question Answering Dataset (SQuAD) [37] is a widely used dataset for extractive span-of-texts-based MRC task, having more than 100k context–question–answer triples created by Crowdworkers from Wikipedia. The questions are *Wh*-questions with guaranteed answers. The authors provided a logistic regression-based method. Later, Rajpurkar et al. [36] added diversity in SQuAD by adding 53,775 unanswerable / impossible questions to test the MRC systems' robustness and named it *SQuAD 2.0*. They evaluated three existing models, namely BiDAF-No-Answer (BNA) and two variants (i.e., versions with and without ELMo) of DocumentQA No-Answer (DocQA).

MS-MARCO: MAchine Reading COmprehension dataset (MS-MARCO) [32] comprises 1,010,916 anonymized questions from Bing's search query logs. The contexts are relevant to web articles indexed by Bing. The answers to these questions are human-generated. This dataset contains 8,841,823 passages–extracted from 3,563,535 web documents retrieved by Bing. This article proposed three different sub-tasks: (i). given a set of passages, determine if a question is answerable, then extract and synthesize the answer as human; (ii). generate a well-formed answer that is understandable with the question and supporting context; and (iii). rank a set of retrieved passages conditioned on a question.

NewsQA: Trischler et al. [47] presented NewsQA containing 100,000 human-generated question–answer pairs for span of texts based QA. The authors provided four models for benchmarking: (i). a heuristic sentence-level baseline, (ii). two neural models, viz., one is based on match-LSTM, and the other is FastQA, and (iii). model based on human data analysis.

TriviaQA: Joshi et al. [25] introduced TriviaQA, which contains over 650K question-answer-document triples prepared from Wikipedia, web articles, and trivia websites. The paper used a Bi-DAF-based random classifier.

SearchQA: Dunn et al. [17] proposed SearchQA, containing question–answer pairs, accompanied by more than one document assumed to be the context. The challenges here are to handle multiple documents. Since the supporting documents are collected after question–answer pairs with information retrieval, the questions are not guaranteed to involve interesting reasoning between multiple documents. Their methods are based on term frequency——inverse document frequency (TF-IDF) (selecting word with maximum TF-IDF score as the answer) and Attention Sum Reader.

FreebaseQA: FreebaseQA [24] is dedicated to open-domain factoid QA over a structured knowledge base. The authors constructed this dataset by matching trivia-type question–answer pairs with subject–predicate–object triples in the knowledge base (i.e., freebase). It has 28k unique questions. They used the Fixed-size Ordinally Forgetting Encoding (FOFE)-neural network-based model to build a Knowledge Base Question Answering (KBQA) system.

NarrativeQA: NarrativeQA [27] is based on summaries of movie scripts and books, addressing the limitations of existing datasets and tasks. The article presents a simple Bi-LSTM-based sequence to sequence (Seq2Seq) model.

QAngaroo: Most existing methods rely on a single sentence or document to answer a query. Enabling models to combine multiple pieces of textual evidences from several documents would extend the scope of RC models. This QAngaroo resource [52] serves this purpose. The proposed models are based on TF-IDF, FastQA, Bi-DAF, etc.

DuReader: DuReader [19] is an open-domain, large-scale Chinese MRC dataset specifically designed to address real-world MRC. It contains 200K questions, 420K answers, and 1M documents from Baidu Search and Baidu Zhidao. The models are based on Match-LSTM and BiDAF.

RACE: RACE [29] consists of 28,000 passages and 100,000 questions generated by human experts that cover a variety of topics of English examinations for middle and high school level Chinese students to test their ability of understanding and reasoning. They compared the performance of several state-of-the-art reading comprehension models like the sliding window algorithm, Stanford Attentive Reader, and Gated-Attention Reader.

ARC Dataset: AI2 Reasoning Challenge (ARC) dataset [11] has 7787 science questions of all non-diagram, multiple-choice (4-way) QA types. This dataset encourages building the QA models that require reasoning to answer a question rather than just surface-level cues to find answers, as most of the datasets have followed this path. Their significant models are based on Bi-DAF and a few neural entailment models like *DecompAttn*, *DGEM*, and *DGEM-OpenIE*.

HotpotQA: Existing MRC datasets cannot perform training of MRC systems that do complex reasoning and provide explanations for the answer. It is obvious for some questions where systems have to travel or reason over multiple sentences and/or passages to find the answer. *HotpotQA* [60] is dedicated for testing those kinds of systems, comprising 113k question–answer pairs from Wikipedia. The paper re-implemented the method as described in Clark et al. [10], subsuming the latest techniques of QA, namely character-level models, self-attention, and bi-attention.

CommonsenseQA: Talmor et al. [46] presented CommonsenseQA, containing 12,247 examples for testing commonsense knowledge. They offered BERT-based baseline model.

Table 2 A comparison of existing MRC and QA datasets. Here, *ScienceQA* is different from other datasets in terms of domain and volume

Dataset	Question source	Answer	Size	Domain
ScentificQA (proposed)	Semi-automatic	Span-of-words	100K+	Scholarly articles (open domain)
ScholarlyRead [41]	Semi-Automatic	span-of-words	10K	Scholarly Articles (Close Domain)
BioRead [34]	Cloze	Fill in single word	16.4 million	Bio-Medical Literature
TweetQA [57]	Crowd-sourced	Generative	14K	News, Twitter
SubjQA [4]	Reviews	Span of words	10K	Reviews: Movies, Restaurants
SQuAD [37]	Crowd-sourced	Span of words	100K	Wikipedia
TREC-QA [51]	Query Logs	IR, Free Form	1479	Short answer questions from any domain
WikiQA [59]	Bing Query Logs	IR, Sentence selection	3047	Wikipedia
Algebra [28]	Standardized tests	Computation	514	Algebra word problems
Science [12]	Standardized tests	multiple choice	855	Math. and Science Test
NewsQA [47]	Crowd-sourced	Span of Words	100k	News
DuReader [19]	Crowd-sourced	Human Generated	200K	Chinese Document.
Narrative QA [27]	Crowd-sourced	Human Generated	46,765	books and movie scripts
MC Test [39]	Crowd-sourced	Multiple choice	2640	Fictional story
CNN/Daily Mail [7]	Cloze+Summary	Fill in single word	1.4M	News Articles
CBT [21]	Cloze	Fill in single word	688k	freely available cultural eBooks

MathQA: MathQA [1] is a large-scale dataset for testing math word and interpretable neural math problems. They developed a neural encoder–decoder model.

CliCR: CliCR [45] comprises around 100k data points constructed from clinical case reports for cloze-style QA models. Their neural models are based on Stanford Attentive Reader and Gated-Attention (GA) Reader.

CODAH: COmmonsense Dataset Adversarially authored by Humans (CODAH) [8] is an adversarially created 2.8K questions for testing commonsense. They proposed a BERT and Generative Pre-trained Transformer (GPT)-1 based models.

CoQA: Conversational Question Answering (CoQA) [38] is for building conversational QA systems, comprising 127,000+ questions with answers collected from 8000+ conversations that are created using the conversation's history between two Crowdworkers in the form of QA. They used a Seq2Seq with an attention model and DrQA model for QA.

RecipeQA: RecipeQA [58] is a multimodal RC dataset in the recipe domain consisting of 36K question–answer pairs from 20K cooking recipes. The proposed model is based on the Impatient Reader.

BioRead: BioRead [34] comprises 16.4 million cloze-style QA examples in the biomedical domain, created in the same spirit to Children's Book Tests. AS-READER, AOA-READER-based baseline models are being used in this article.

DREAM: Dialogue-based REAding Comprehension Examination (DREAM) [44] is a multiple-choice RC dataset containing 10,197 multiple-choice questions for 6,444 dialogues. In contrast to existing datasets, this one is the first that focuses on in-depth multi-party dialogue understanding. They proposed a generative pre-trained language model (LM) following the framework of fine-tuned transformer LM.

There has been an interest in building NLP systems, including QA on COVID-19. A workshop named *NLP for COVID-19* was held as a part of ACL 2020 [50]. As a part of this workshop, Das et al. [13] presented an information retrieval system on scientific articles related to COVID-19. Their method extracts the relevant articles and sections based on a given query. A competition on *Biomedical Semantic Indexing and Question Answering (BioASQ)* [48] has provided a dataset on QA in the biomedical domain. It comprises questions (Q), human-annotated answers (A), and the relevant contexts (C). This challenge aims to develop systems that will be able to semantically index huge numbers of biomedical scientific articles and return good quality answers given a question. The systems make use of information from biomedical articles and ontologies. In contrast to these datasets, our dataset is on scholarly articles from the scientific domain with 100k QA pairs. Some of these datasets and comparisons in multiple levels are shown in Table 2.

3 Dataset creation

We crawl three years' accepted research articles from the *International Joint Conferences on Artificial Intelligence (IJCAI)* conference. We collect 1825 such articles. IJCAI articles are openly accessible; hence, we use those to develop our dataset. The articles are in portable document format (PDF). At first, we convert such PDF articles to *JavaScript Object*

Table 3 Examples of a few phrases extracted by the Stanford constituency parser that cannot be the plausible answers

We/we	Some	An example
Our method/s	Them	Best
This	It/it	That
These/these	Our/our	All
These rules	They	a
These algorithms	Many	Ad
That	Much	Each
Those	Past	I
The algorithms	Other	T
The models	Me	q
The paper	Use	Their/their
Such problems	Both	He
These conditions	The	And many more

Notation (JSON) encoded files using the Science Parse library[1]. We extract the abstracts only from these JSON formatted articles, considering them as the context/document/paragraph/passage in our experiment, and use these terms interchangeably throughout this paper. The average length of abstracts remains within 260–300 words. We split these contexts into sentences using NLTKs' Punkt Sentence Tokenizer[2]. This tokenizer splits a context into a list of sentences. These extracted list of sentences are passed through a constituency parser [63]. We use the Stanford constituency parser (SCP) for this purpose, which essentially divides the given sentences into nouns and verb phrases. We consider the noun phrases. The studies of Rajpurkar et al. [37] and Trischler el al. [47] have suggested that the noun phrases of a particular passage are the plausible answers for that very passage. Particularly, the study of Rajpurkar et al. [37] has explored the diversity of the answer types of SQuAD. They parsed answers using the constituency parser and PoS tagger contained in the Stanford CoreNLP tool. Their analysis revealed that 32.6% and 31.8% of the answers are proper nouns and common nouns, respectively. The article by Trischler el al. [47] further showed that most of answers (i.e., 22.2%) are common noun phrases. In line with these findings, we extract and manually evaluate all the noun phrases for each abstract. It is found that there are many such phrases (shown in Table 3) that cannot be the plausible answers. We discard those phrases from the list of answers and consider the remaining phrases as plausible answers for a particular abstract.

We make a pairing of the plausible answers with its abstract. We feed those paired document–answer as inputs to an answer aware question generator (QG) [62] model to obtain the accurate questions of those answers. We train that model with combination of SQuAD [37] and ScholarlyRead [41] datasets. The SQuAD is a widely used benchmark dataset on Wikipedia articles, and ScholarlyRead is a recently proposed benchmark dataset on scholarly articles for MRC. The questions yielded by the QG model are manually checked by human annotators. The diagram of the QG model from research articles is shown in Fig. 1. This way, we create the context–answer–question triples of more than 100k. We provide span indices (i.e., start and end index) of the answer in the context of the training/dev/test examples, as it is for extractive QA (i.e., the answer to a question should contain in the supporting passage). We compute the indices in this way: Finding the indices is straightforward if the answer phrase occurs once throughout the passage. We face the challenge when the targeting answer phrase appears in multiple sentences in the supporting passage. To overcome this, we must find out the answer containing exact sentence. To do this, we take the Levenshtein distance ratio between the question and every sentence of the document. Levenshtein distance between two pieces of texts provides the distance between them, whereas Levenshtein distance ratio provides similarity between two comparing sentences. We pick up the sentence, which has the maximum ratio. This sentence ultimately corresponds to the answer containing sentence. Then, we compute the start and end indices of the answer phrase from the sentence. We coin our dataset as *ScienceQA*[3]. We split the whole dataset into training, development, and testing sets with 82415, 10000, and 10000 number of instances, respectively.

3.1 Evaluation of generated questions

For evaluating the QG system, we use the metrics that are widely used for Machine Translation (MT) and Summarization tasks, such as BLEU [33], METEOR [14], and ROUGE [31]. We apply these metrics to a sample of 2000 outputs for evaluation. Our annotators generate questions for these 2000 examples. These metrics are the n-gram-based metrics, where lexical matching is performed. We also apply consensus-based image description evaluation (CIDEr) [49], which is a popular evaluation metric for evaluating various tasks in computer vision. Evaluation results yield the BLEU, METEOR, ROUGE, and CIDEr scores of 0.12, 0.098, 0.117, and 0.244, respectively. We also define an entailment-based metric to determine the entailment relation between the machine-generated questions and reference questions. For this purpose, we use a state-of-the-art entailment model, equipped with external knowledge [9]. The model is trained with the combination of SNLI [5] and Multi-NLI [54] corpus. We evaluate the trained natural language inference

[1] https://github.com/allenai/science-parse

[2] https://www.nltk.org/api/nltk.tokenize.html

[3] We make *ScienceQA* publicly available at: https://github.com/TanikSaikh/Scientific-Question-Answering.

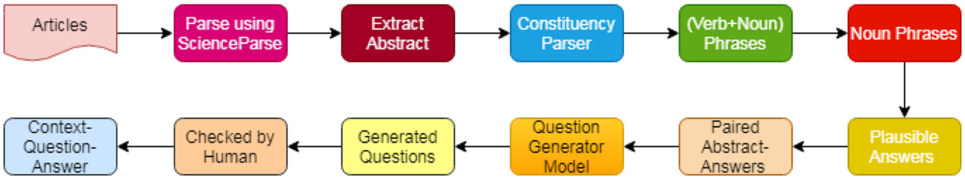

Fig. 1 Proposed system for QG from research articles

(NLI) system on our questions pairs (i.e., system-generated and human-generated questions as premise and hypothesis, respectively) to predict the entailment relation between them. We asked the annotators to provide entailment labels (i.e., entailment, contradiction, and neutral) to each reference question given the corresponding system-generated questions. We compare these labels with the systems' predicted labels. Further, we obtain an accuracy of 65%. Entailment relation between these two questions indicates that generated questions are logically accurate and correct, i.e., much closer to the question as humans generally used to ask. This entailment-based metric could be a potential evaluation metric for NLG tasks.

3.2 Annotation guidelines

We employ two annotators to check the quality of the system-generated questions for more than one and a half years. Both the annotators, aged 36 and 40, are post-graduates in science with prior experience for the similar task. We instructed the annotators as follows:

- The generated questions should be grammatically correct and spelled correctly; proper punctuation should be there; questions words and proper nouns should begin with a capital letter.
- The question should be relevant to its answer and understandable to anyone, even those unaware of the context.
- The answer should be unique and factual for all the factoid questions.

We randomly chose 1000 samples. We employ two annotators to judge the naturalness (i.e., verify how the generated questions are grammatically correct and fluent). They were asked to give scores (between 0-4) to each question based on the above two parameters. We compute the inter-annotator agreement ratio in terms of the kappa coefficient [18]. It is obtained as 0.81, which is considered good according to Landis et al. [30].

3.3 Dataset analysis

We compute the average length of context, answer, and the question in the *ScienceQA* obtained as 121.18, 10.21, and 3.46, respectively. To understand the properties of the *ScienceQA*, we analyze the questions and answers in the training or development set.

Diversity in answer: As we assumed that the noun phrases are the answers for a particular document, our answers' types are noun phrases.

Diversity in question: Simple QA systems mainly deal with factoid questions. To make our QA system simple, we generate all the questions as a factoid. Among them, the maximum number of questions is of *What* type; other types include *Which, How*, and *Why*, etc.

Reasoning required to answer questions: We randomly picked up triples to understand the reasoning required to answer the questions. We analyze the triples, denoting what is required to answer the questions, and like many standard RC datasets, we manually label the examples with the categories shown in Table 4. From Table 4, it is evident that some percentage of questions in *ScienceQA* are not straightforward to answer, whereas many are so easy to answer. It is also observed maximum examples have some lexical divergence between the question and the answer in the passage.

4 Methods

We implement the models based on BERT, Science BERT (SciBERT) and a combination of SciBERT and Bi-DAF. We describe each of these in the following sections.

BERT: We develop a baseline model using BERT [15]. The model is fine-tuned on our dataset combined with the SQuAD dataset. An architectural diagram is shown in Fig. 2. BERT requires highlighting the span of the texts containing the answer, which implies simply predicting the start and end indices of the answer. For this, we apply two classifiers for predicting both indices. We feed its final embedding into the start token classifier for each word in the passage. This classifier has a single set of weights applied to every word. After taking the dot product between the output embeddings and the start weights, we apply the softmax activation to produce a probability distribution over all of the words. We pick the word with the highest probability as a possible candidate for the start token. The process is repeated for the end index prediction using the end token classifier, which has a separate weight vector.

Table 4 Manually labeled 200 examples into various categories

Reasoning	Description	Examples	%
Lexical Divergence	Major correspondences between the question and the answer sentence are synonyms.	Q: What did authors introduce for the CUMULATIVE constraint based on a new energetic relaxation? Sentence: We present two novel filtering algorithms for the CUMULATIVE constraint based on a new energetic relaxation.	65
Multiple sentence reasoning	There is anaphora, or higher-level fusion of multiple sentences is required.	Q: Using which framework the existing TransH model is generalized to a new model, m-TransH? Sentences: We advocate a novel modeling framework, which models multi-fold relations directly using this canonical representation. Using this framework, the existing TransH model is generalized to a new model, m-TransH. We demonstrate experimentally that m-TransH outperforms TransH by a large margin, thereby establishing a new state of the art.	12
Lexical level Variation (world Knowledge)	Major correspondences between the question and the answer sentence require extra knowledge to resolve.	Q: What is the purpose of proposing two novel filtering algorithms? Sentence: We present two novel filtering algorithms for the CUMULATIVE constraint based on a new energetic relaxation.	16

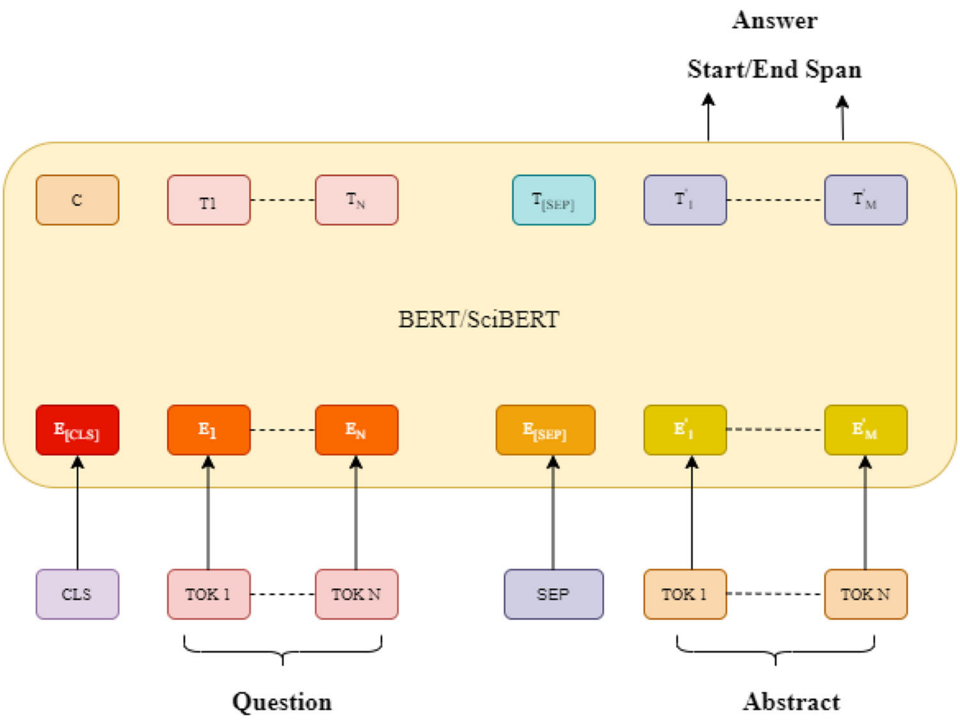

Fig. 2 A BERT/SciBERT-based QA Model. Image courtesy: ([15]; https://blog.scaleway.com/2019/understanding-text-with-bert/) with minimal modifications

Science-BERT: This paper deals with scientific texts, so we use a variant of BERT trained on scientific texts to represent the input texts better. Specifically, we use *SciBERT* [3] that is trained on 1.14M full-text papers with 3.1B tokens from the semantic scholar. SciBERT has its vocabulary built so that it best fits the training corpus. We use this pre-trained model, to further fine-tune on our dataset combined with the SQuAD dataset. The architecture diagram (i.e., Fig. 2) remains the same as the previous one.

It is clear from Fig. 2 that these two models take input as the concatenation of abstract and question separated by [SEP] token. We have vector representations of both the documents (abstracts) with l_d tokens $d_1[d_2, d_3......d_{l_d} \in \mathbb{R}^h]$ and the question $[q \in \mathbb{R}^h]$. The aim is to predict the span that is most likely the accurate answer. There are two classifiers on top of this with a start vector S and end vector E to produce start and end indices of the predicted answer. More explicitly, we use a bilinear product to capture the similarity between d_i

and q.

$$P^{(start)}(i) = \frac{exp(d_i W^{(start)} q)}{\sum_{i'} exp(d_{i'} W^{(start)} q)} \tag{2}$$

$$P^{(end)}(i) = \frac{exp(d_i W^{(end)} q)}{\sum_{i'} exp(d_{i'} W^{(end)} q)} \tag{3}$$

$W^{(start)}$, $W^{(end)}$ $\in \mathbb{R}^{h \times h}$ are additional parameters to be learned.

Combination of SciBERT and Bi-DAF: This model combines both SciBERT and Bi-DAF approaches. This model builds on top of a fixed word embedding pre-trained from the unlabeled text while all the remaining parameters need to be learned from the limited training data using the Bi-DAF architecture. Bi-DAF is one of the promising models for span-of-texts-based QA systems, which have achieved state-of-the-art results on many standard datasets. We want to stress that this is different from an ensemble model, but instead is a single model that uses SciBERT's contextualized and combined embedding of both the question and the context to train the Bi-DAF model and then predict. We propose this method to leverage the following: *viz.,* (i). representation of inputs from a powerful pre-trained language model (BERT) and (ii). question-aware passage representation that does not include early summarization by bi-directional attention flow mechanism (Bi-DAF). We use BERT-as-service [56] to generate SciBERT embedding quickly and efficiently. Let $\{q_1, q_2...q_n\}$ and $\{c_1, c_2...c_m\}$ denote the tokens in question and context, respectively. Then **F** is as follows: $\{q_1...q_n,'' \;|||\; '', c_1...c_m\}$.

$$SciBERT\_Embedding(F) = H$$

where H is the contextualized combined embedding. H is then split, respectively, into Q and C, where Q and C denote contextualized embeddings of question and context that are matrices of dimension $(token\_length\_question, 768)$ and $(token\_length\_context, 768)$, respectively. These questions' and contexts' contextual embeddings are then sent to the Bi-DAF architecture attention layer.

5 Experiments, results, and discussion

We run the above-proposed baseline and models on our ScienceQA dataset and report our results. We use the following data augmentation technique to increase the training examples: combine the following three datasets, *viz.,* (a). *ScholarlyRead*, (b). *SQuAD v1.1*, and (c). *ScienceQA* to train our models and test on an unseen set from ScienceQA. We then split the combined dataset into the following training / development / testing instances: 159760/299955/9986.

Table 5 Proposed methods' results and comparison with baseline model and previous system

Models	Results (%) Exact Match	F1
Best performing systems		
SciBERT+Bi-DAF	48.74	65.5
SciBERT	63.8	75.46
Baseline system		
BERT	41.2	65.73
Comparing system		
ScholarlyRead [41]	20.6	37.31

We use two standard evaluation metrics (i.e., Exact Match and F1) that are widely used for evaluating span-of-texts-based QA systems [25,29,37]. The metrics do not consider punctuation and articles (e.g., a, an, the, etc.). We have only one reference answer for testing.

Exact Match: This metric computes the number of matching predicted answers with the ground truth answer character by character.

F1 score: This is a macro-averaged F1 score. It converts the predicted and ground-truth answers as bag-of-words. The average overlap between the predicted and gold standard answers is then predicted. Further, we compute their F1 and then take the average over all the instances.

We show the results in Table 5. It is evident that the models we apply here perform way better than ScholarlyRead [41]. We use a pre-trained (trained on the combination of Wikipedia and book corpus) vanilla BERT model fine-tuned on our dataset as our baseline. Surprisingly, the vanilla model even performs better than our comparing system. The SciBERT only model performs the best among all. The other model which is a combination of SciBERT and Bi-DAF model although performs way better than our comparing systems and baseline, but falls short of the only SciBERT model. It shows inferior performance than the SciBERT (only) model in *Exact Match* and *F1* and vanilla BERT model in terms of the only *F1*. BERT is too harsh to tackle the instances that require complex reasoning [40]. Also, many studies [10,42,43,61] suggest that Bi-DAF is also not very efficient in capturing these instances. The BERT implicitly uses self-attention that learns global interaction between each pair of words. On the contrary, Bi-DAF utilizes attention and Bi-LSTM at the end. It seems counter-intuitive, and this could be one of the possible reasons for the performance drop. Another reason could be different tokenization methods used in SciBERT and Bi-DAF architecture. We reserve this as a future investigation.

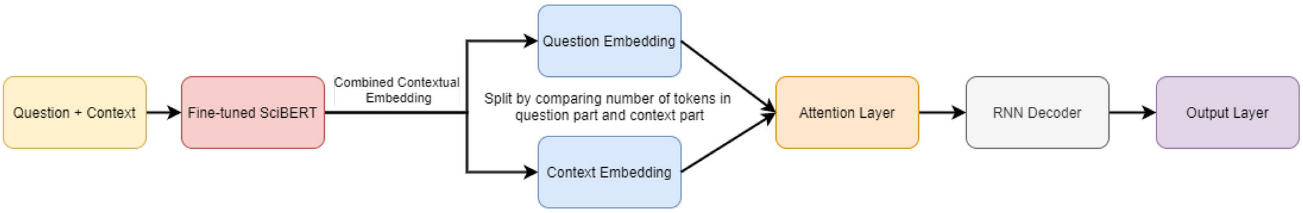

Fig. 3 BiDAF + SciBERT MRC model architecture

5.1 Error analysis

We perform qualitative and quantitative error analyses. We extract wrongly classified instances from our best-performing model for qualitative error analysis. We examine those examples and try to find some patterns. The following are our observations:

- The combined model (i.e., SciBERT + Bi-DAF) works well when the answers combine one or two words. It fails to predict correctly in case of answers with longer sequence of tokens.
- Abstracts are usually 260-300 words, which can be considered a long document compared to the instances in the SQuAD dataset. Such long documents make the QA task difficult.

5.2 Challenges of using SciBERT pre-trained embeddings

The first challenge that we faced was due to the different tokenization methods of SciBERT as compared to the Bi-DAF model. SciBERT uses WordPiece [55] tokenization method, often splitting individual words into multiple tokens. For example, if the input *”John Johanson’s house”* has the answer start and end index as 0 and 1, respectively, for word-level tokenization method, then due to WordPiece tokenization, it changes to *”john johan ##son ’ s house”*. Now, the answer start and end indices could be (0, 1) or (0, 2). This creates a lot of difference in a model in which accurate start and end spans are a necessity to achieve high exact match (EM) scores. The exact end token is often unavailable, so the nearest neighbor needs to be chosen, which adds to the uncertainty in the dataset. This also leads to a significant difference between the EM and F1 scores achieved by this model.

The second challenge is due to the maximum sequence length limitation. We tried to maximize that length to allow for more number tokens at once, which would prevent truncation of contexts, but there was a considerable trade-off in the batch size we could choose due to the memory limitation of GPUs. We trained on multiple NVIDIA RTX 2080 Ti 11GB and finally had to settle with a max_sequence_length of 384 and a batch size of 8. As hypothesized, a larger value of the above variables leads to higher model accuracy.

6 Conclusion and future work

In this paper, we present ScienceQA, a novel dataset for benchmark evaluation of methods in the MRC (QA and QG in particular) task on scholarly articles. The dataset is created semi-automatically, consisting of over 100k triples of context–question–answer. The developed QA system could provide valuable evidence in managing the vast number of scholarly submissions. We offer a baseline and two more models, *viz.,* (i). Vanilla BERT, (ii). Science BERT (i.e., SciBERT), and (iii). Combination of BERT and Bi-DAF. Our proposed models are competitive compared to the existing state-of-the-art models. Our future works would include:

1. Extension of this task considering the full-text articles instead of only abstracts. Abstracts are not enough to answer the intricate details of the paper.
2. Advancement of the Bi-DAF model by incorporating multi-hop attention.
3. Enrichment in size of the dataset up to 500k, and also multi-hop version of ScienceQA (like HotpotQA).
4. Building visual question answering (VQA) models utilizing images and tables available in the full-text articles.
5. Model based on Generative Pre-trained Transformer (GPT) - 3.

We make our code and the *ScienceQA* dataset available at https://www.iitp.ac.in/~ai-nlp-ml/resources.html#ScienceQA also to further research in QA and QG on scholarly articles.

Acknowledgements On behalf of all the co-authors, Mr. Tanik Saikh would like to acknowledge ”Elsevier Centre of Excellence for Natural Language Processing” at the Department of Computer Science and Engineering at Indian Institute of Technology Patna for supporting research work carried out in this paper.

References

1. Amini, A., Gabriel, S., Lin, S., Koncel-Kedziorski, R., Choi, Y., Hajishirzi, H.: MathQA: towards interpretable math word problem solving with operation-based formalisms. 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), pp. 2357–2367. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1245
2. Beltagy, I., Cohan, A., Feigenblat, G., Freitag, D., Ghosal, T., Hall, K., Herrmannova, D., Knoth, P., Lo, K., Mayr, P., Patton, R., Shmueli-Scheuer, M., de Waard, A., Wang, K., Wang, L.: Overview of the second workshop on scholarly document processing. In: Proceedings of the Second Workshop on Scholarly Document Processing, pp. 159–165. Association for Computational Linguistics, Online (2021). https://aclanthology.org/2021.sdp-1.22
3. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1371
4. Bjerva, J., Bhutani, N., Golahn, B., Tan, W.C., Augenstein, I.: SubjQA: a dataset for subjectivity and review comprehension. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (2020)
5. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 632–642. Association for Computational Linguistics, Lisbon, Portugal (2015). https://doi.org/10.18653/v1/D15-1075
6. Chen, D.: Neural Reading Comprehension and Beyond. Ph.D. thesis, Stanford University (2018)
7. Chen, D., Bolton, J., Manning, C.D.: A Thorough examination of the cnn/daily mail reading comprehension task. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2358–2367. Association for Computational Linguistics, Berlin, Germany (2016). https://doi.org/10.18653/v1/P16-1223
8. Chen, M., D'Arcy, M., Liu, A., Fernandez, J., Downey, D.: CODAH: An adversarially-authored question answering dataset for common sense. In: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, pp. 63–69 (2019)
9. Chen, Q., Zhu, X., Ling, Z.H., Inkpen, D., Wei, S.: Neural natural language inference models enhanced with external knowledge. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2406–2417. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1224
10. 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), pp. 845–855. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-1078
11. 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. (2018) CoRR **abs/1803.05457**http://arxiv.org/abs/1803.05457
12. Clark, P., Etzioni, O.: My computer is an honor student–but how intelligent is it? Standardized tests as a measure of AI. AI Mag. **37**(1), 5–12 (2016)
13. Das, D., Katyal, Y., Verma, J., Dubey, S., Singh, A., Agarwal, K., Bhaduri, S., Ranjan, R.: Information retrieval and extraction on covid-19 clinical articles using graph community detection and bio-BERT embeddings. In: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics, Online (2020). https://aclanthology.org/2020.nlpcovid19-acl.7
14. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the EACL 2014 Workshop on Statistical Machine Translation (2014)
15. 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), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423. https://www.aclweb.org/anthology/N19-1423
16. Dhingra, B., Liu, H., Yang, Z., Cohen, W., Salakhutdinov, R.: Gated-attention readers for text comprehension. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1832–1846. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/P17-1168
17. Dunn, M., Sagun, L., Higgins, M., Güney, V.U., Cirik, V., Cho, K.: SearchQA: a new Q&A dataset augmented with context from a search engine. CoRR **abs/1704.05179** http://arxiv.org/abs/1704.05179 (2017)
18. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychol. Bull. **76**(5), 378 (1971)
19. He, W., Liu, K., Liu, J., Lyu, Y., Zhao, S., Xiao, X., Liu, Y., Wang, Y., Wu, H., She, Q., Liu, X., Wu, T., Wang, H.: DuReader: A chinese machine reading comprehension dataset from real-world applications. In: Proceedings of the Workshop on Machine Reading for Question Answering, pp. 37–46. Association for Computational Linguistics, Melbourne, Australia (2018)
20. Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in neural information processing systems, pp. 1693–1701 (2015)
21. Hill, F., Bordes, A., Chopra, S., Weston, J.: The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. arXiv preprint arXiv:1511.02301 (2015)
22. Hirschman, L., Light, M., Breck, E., Burger, J.D.: Deep read: a reading comprehension system. In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, pp. 325–332. Association for Computational Linguistics, College Park, Maryland, USA (1999). https://doi.org/10.3115/1034678.1034731
23. Huang, L., Le Bras, R., Bhagavatula, C., Choi, Y.: Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2391–2401. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1243
24. Jiang, K., Wu, D., Jiang, H.: FreebaseQA: a new factoid QA data set matching trivia-style question-answer pairs with freebase. 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), pp.

318–323. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1028

25. Joshi, M., Choi, E., Weld, D.S., 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. Association for Computational Linguistics, Vancouver, Canada (2017)
26. Kim, S., Park, D., Choi, Y., Lee, K., Kim, B., Jeon, M., Kim, J., Tan, A.C., Kang, J.: A pilot study of biomedical text comprehension using an attention-based deep neural reader: design and experimental analysis. JMIR Med. Inf. **6**(1), e2 (2018)
27. Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The NarrativeQA reading comprehension challenge. Trans. Assoc. Comput. Linguist. **6**, 317–328 (2018)
28. Kushman, N., Artzi, Y., Zettlemoyer, L., Barzilay, R.: Learning to automatically solve algebra word problems. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 271–281 (2014)
29. 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, pp. 785–794. Association for Computational Linguistics, Copenhagen, Denmark (2017). https://doi.org/10.18653/v1/D17-1082
30. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics pp. 159–174 (1977)
31. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, Spain (2004). https://aclanthology.org/W04-1013
32. Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: MS MARCO: A human generated machine reading comprehension dataset. In: Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016 (2016)
33. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). https://doi.org/10.3115/1073083.1073135
34. Pappas, D., Androutsopoulos, I., Papageorgiou, H.: BioRead: a new dataset for biomedical reading comprehension. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018). https://www.aclweb.org/anthology/L18-1439
35. Park, D., Choi, Y., Kim, D., Yu, M., Kim, S., Kang, J.: Can machines learn to comprehend scientific literature? IEEE Access **7**, 16246–16256 (2019)
36. 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), pp. 784–789. Association for Computational Linguistics, Melbourne, Australia (2018). https://doi.org/10.18653/v1/P18-2124
37. 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, pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (2016). https://doi.org/10.18653/v1/D16-1264
38. Reddy, S., Chen, D., Manning, C.D.: CoQA: a conversational question answering challenge. Trans. Assoc. Comput. Linguist. **7**, 249–266 (2019)
39. Richardson, M., Burges, C.J., Renshaw, E.: MCTest: a challenge dataset for the open-domain machine comprehension of text. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 193–203. Association for Computational Linguistics, Seattle, Washington, USA (2013). https://www.aclweb.org/anthology/D13-1020
40. Rogers, A., Kovaleva, O., Rumshisky, A.: A primer in BERTology: what we know about how BERT works. Trans. Assoc. Comput. Linguist. **8**, 842–866 (2020)
41. Saikh, T., Ekbal, A., Bhattacharyya, P.: ScholarlyRead: a new dataset for scientific article reading comprehension. In: Proceedings of The 12th Language Resources and Evaluation Conference, pp. 5498–5504. European Language Resources Association, Marseille, France (2020). https://www.aclweb.org/anthology/2020.lrec-1.675
42. Sarkar, S.: Effectiveness of deep networks in NLP using BiDAF as an example architecture. arXiv preprint arXiv:2109.00074 (2021)
43. Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bi-directional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)
44. Sun, K., Yu, D., Chen, J., Yu, D., Choi, Y., Cardie, C.: DREAM: a challenge data set and models for dialogue-based reading comprehension. Trans. Assoc. Comput. Linguist. **7**, 217–231 (2019)
45. Šuster, S., Daelemans, W.: CliCR: a dataset of clinical case reports for machine reading comprehension. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (2018)
46. Talmor, A., Herzig, J., Lourie, N., Berant, J.: CommonsenseQA: a question answering challenge targeting commonsense knowledge. 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), pp. 4149–4158. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1421
47. Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., Suleman, K.: NewsQA: a machine comprehension dataset. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, pp. 191–200. Association for Computational Linguistics, Vancouver, Canada (2017). https://doi.org/10.18653/v1/W17-2623
48. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinf. **16**, 138 (2015). https://doi.org/10.1186/s12859-015-0564-6
49. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: consensus-based image description evaluation. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4566–4575 (2015). https://doi.org/10.1109/CVPR.2015.7299087
50. Verspoor, K., Cohen, K.B., Dredze, M., Ferrara, E., May, J., Munro, R., Paris, C., Wallace, B. (eds.): Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020. Association for Computational Linguistics, Online (2020). https://aclanthology.org/2020.nlpcovid19-acl.0
51. Voorhees, E.M., Tice, D.M.: Building a question answering test collection. In: Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '00, pp. 200–207. ACM, New York, NY,

USA (2000). https://doi.org/10.1145/345508.345577. http://doi.acm.org/10.1145/345508.345577

52. Welbl, J., Stenetorp, P., Riedel, S.: Constructing datasets for multi-hop reading comprehension across documents. Trans. Assoc. Comput. Linguist. **6**, 287–302 (2018)
53. Weston, J., Bordes, A., Chopra, S., Rush, A.M., van Merriënboer, B., Joulin, A., Mikolov, T.: Towards AI-complete question answering: a set of prerequisite toy tasks. arXiv preprint arXiv:1502.05698 (2015)
54. Williams, A., Nangia, N., Bowman, S.: A broad-coverage challenge corpus for sentence understanding through inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pp. 1112–1122. Association for Computational Linguistics (2018)
55. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation (2016)
56. Xiao, H.: Bert-as-service. https://github.com/hanxiao/bert-as-service (2018)
57. Xiong, W., Wu, J., Wang, H., Kulkarni, V., Yu, M., Chang, S., Guo, X., Wang, W.Y.: TWEETQA: a social media focused question answering dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5020–5031. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1496
58. Yagcioglu, S., Erdem, A., Erdem, E., Ikizler-Cinbis, N.: RecipeQA: a challenge dataset for multimodal comprehension of cooking recipes. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 1358–1368. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1166
59. Yang, Y., Yih, W.t., Meek, C.: WikiQA: a challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2013–2018. Association for Computational Linguistics, Lisbon, Portugal (2015). https://doi.org/10.18653/v1/D15-1237
60. 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, pp. 2369–2380. Association for Computational Linguistics, Brussels, Belgium (2018). https://doi.org/10.18653/v1/D18-1259
61. Yu, A.W., Dohan, D., Le, Q., Luong, T., Zhao, R., Chen, K.: Fast and accurate reading comprehension by combining self-attention and convolution. In: International Conference on Learning Representations (2018). https://openreview.net/forum?id=B14TlG-RW
62. Yuan, X., Wang, T., Gülçehre, Ç., Sordoni, A., Bachman, P., Zhang, S., Subramanian, S., Trischler, A.: Machine comprehension by text-to-text neural question generation. In: Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017, Vancouver, Canada, August 3, 2017, pp. 15–25 (2017)
63. Zhu, M., Zhang, Y., Chen, W., Zhang, M., Zhu, J.: Fast and accurate shift-reduce constituent parsing. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 434–443. Association for Computational Linguistics, Sofia, Bulgaria (2013). https://www.aclweb.org/anthology/P13-1043