PubMedQA: A Dataset for Biomedical Research Question Answering

Qiao Jin
University of Pittsburgh
qiao.jin@pitt.edu

Bhuwan Dhingra
Carnegie Mellon University
bdhingra@cs.cmu.edu

Zhengping Liu
University of Pittsburgh
zliu@pitt.edu

William W. Cohen
Google AI
wcohen@google.com

Xinghua Lu
University of Pittsburgh
xinghua@pitt.edu

Abstract

We introduce PubMedQA, a novel biomedical question answering (QA) dataset collected from PubMed abstracts. The task of PubMedQA is to answer research questions with yes/no/maybe (e.g.: *Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?*) using the corresponding abstracts. PubMedQA has 1k expert-annotated, 61.2k unlabeled and 211.3k artificially generated QA instances. Each PubMedQA instance is composed of (1) a question which is either an existing research article title or derived from one, (2) a context which is the corresponding abstract without its conclusion, (3) a long answer, which is the conclusion of the abstract and, presumably, answers the research question, and (4) a yes/no/maybe answer which summarizes the conclusion. PubMedQA is the first QA dataset where reasoning over biomedical research texts, especially their quantitative contents, is required to answer the questions. Our best performing model, multi-phase fine-tuning of BioBERT with long answer bag-of-word statistics as additional supervision, achieves 68.1% accuracy, compared to single human performance of 78.0% accuracy and majority-baseline of 55.2% accuracy, leaving much room for improvement. PubMedQA is publicly available at https://pubmedqa.github.io.

Question:
Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?
Context:
(Objective) Recent studies have demonstrated that statins have pleiotropic effects, including anti-inflammatory effects and atrial fibrillation (AF) preventive effects [...]
(Methods) 221 patients underwent CABG in our hospital from 2004 to 2007. 14 patients with preoperative AF and 4 patients with concomitant valve surgery [...]
(Results) The overall incidence of postoperative AF was 26%. ***Postoperative AF was significantly lower in the Statin group compared with the Non-statin group (16% versus 33%, p=0.005).*** Multivariate analysis demonstrated that independent predictors of AF [...]
Long Answer:
(Conclusion) Our study indicated that preoperative statin therapy seems to reduce AF development after CABG.
Answer: yes

Figure 1: An instance (Sakamoto et al., 2011) of PubMedQA dataset: Question is the original question title; Context includes the structured abstract except its conclusive part, which serves as the Long Answer; Human experts annotated the Answer yes. Supporting fact for the answer is ***highlighted***.

1 Introduction

A long-term goal of natural language understanding is to build intelligent systems that can reason and infer over natural language. The question answering (QA) task, in which models learn how to answer questions, is often used as a benchmark for quantitatively measuring the reasoning and inferring abilities of such intelligent systems.

While many large-scale annotated general domain QA datasets have been introduced (Rajpurkar et al., 2016; Lai et al., 2017; Kočiskỳ et al., 2018; Yang et al., 2018; Kwiatkowski et al., 2019), the largest annotated biomedical QA dataset, BioASQ (Tsatsaronis et al., 2015) has less than 3k training instances, most of which are simple factual questions. Some works proposed automatically constructed biomedical QA datasets (Pampari et al., 2018; Pappas et al., 2018; Kim et al., 2018), which have much larger sizes. However, questions of these datasets are mostly factoid, whose answers can be extracted in the contexts without much reasoning.

In this paper, we aim at building a biomedical QA dataset which (1) has substantial instances with some expert annotations and (2) requires reasoning over the contexts to answer the questions. For this, we turn to the PubMed[1], a search engine providing access to over 25 million references of

[1] https://www.ncbi.nlm.nih.gov/pubmed/

biomedical articles. We found that around 760k articles in PubMed use questions as their titles. Among them, the abstracts of about 120k articles are written in a structured style – meaning they have subsections of "Introduction", "Results" etc. Conclusive parts of the abstracts, often in "Conclusions", are the authors' answers to the question title. Other abstract parts can be viewed as the contexts for giving such answers. This pattern perfectly fits the scheme of QA, but modeling it as abstractive QA, where models learn to generate the conclusions, will result in an extremely hard task due to the variability of writing styles.

Interestingly, more than half of the question titles of PubMed articles can be briefly answered by yes/no/maybe, which is significantly higher than the proportions of such questions in other datasets, e.g.: just 1% in Natural Questions (Kwiatkowski et al., 2019) and 6% in HotpotQA (Yang et al., 2018). Instead of using conclusions to answer the questions, we explore answering them with yes/no/maybe and treat the conclusions as a long answer for additional supervision.

To this end, we present PubMedQA, a biomedical QA dataset for answering research questions using yes/no/maybe. We collected all PubMed articles with question titles, and manually labeled 1k of them for cross-validation and testing. An example is shown in Fig. 1. The rest of yes/no/answerable QA instances compose of the unlabeled subset which can be used for semi-supervised learning. Further, we automatically convert statement titles of 211.3k PubMed articles to questions and label them with yes/no answers using a simple heuristic. These artificially generated instances can be used for pre-training. Unlike other QA datasets in which questions are asked by crowd-workers for existing contexts (Rajpurkar et al., 2016; Yang et al., 2018; Kočiskỳ et al., 2018), in PubMedQA contexts are generated to answer the questions and both are written by the same authors. This consistency assures that contexts are perfectly related to the questions, thus making PubMedQA an ideal benchmark for testing scientific reasoning abilities.

As an attempt to solve PubMedQA and provide a strong baseline, we fine-tune BioBERT (Lee et al., 2019) on different subsets in a multi-phase style with additional supervision of long answers. Though this model generates decent results and vastly outperforms other baselines, it's still much worse than the single-human performance, leaving significant room for future improvements.

2 Related Works

Biomedical QA: Expert-annotated biomedical QA datasets are limited by scale due to the difficulty of annotations. In 2006 and 2007, TREC[2] held QA challenges on genomics corpus (Hersh et al., 2006, 2007), where the task is to retrieve relevant documents for 36 and 38 topic questions, respectively. QA4MRE (Peñas et al., 2013) included a QA task about Alzheimer's disease (Morante et al., 2012). This dataset has 40 QA instances and the task is to answer a question related to a given document using one of five answer choices. The QA task of BioASQ (Tsatsaronis et al., 2015) has phases of (a) retrieve question-related documents and (b) using related documents as contexts to answer yes/no, factoid, list or summary questions. BioASQ 2019 has a training set of 2,747 QA instances and a test set of 500 instances.

Several large-scale automatically collected biomedical QA datasets have been introduced: emrQA (Pampari et al., 2018) is an extractive QA dataset for electronic medical records (EHR) built by re-purposing existing annotations on EHR corpora. BioRead (Pappas et al., 2018) and BMKC (Kim et al., 2018) both collect cloze-style QA instances by masking biomedical named entities in sentences of research articles and using other parts of the same article as context.

Yes/No QA: Datasets such as HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), ShARC (Saeidi et al., 2018) and BioASQ (Tsatsaronis et al., 2015) contain yes/no questions as well as other types of questions. BoolQ (Clark et al., 2019) specifically focuses on naturally occurring yes/no questions, and those questions are shown to be surprisingly difficult to answer. We add a "maybe" choice in PubMedQA to cover uncertain instances.

Typical neural approaches to answering yes/no questions involve encoding both the question and context, and decoding the encoding to a class output, which is similar to the well-studied natural language inference (NLI) task. Recent breakthroughs of pre-trained language models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) show significant performance im-

[2] https://trec.nist.gov/

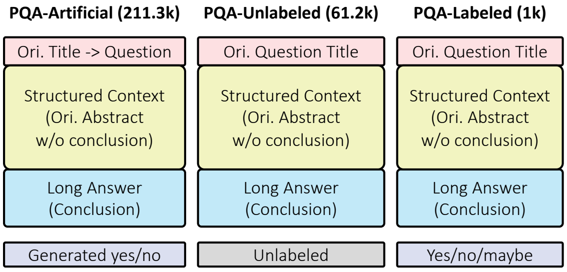

Figure 2: Architecture of PubMedQA dataset. PubMedQA is split into three subsets, PQA-A(rtificial), PQA-U(nlabeled) and PQA-L(abeled).

provements on NLI tasks. In this work, we use domain specific versions of them to set baseline performance on PubMedQA.

3 PubMedQA Dataset

3.1 Data Collection

PubMedQA is split into three subsets: labeled, unlabeled and artificially generated. They are denoted as PQA-L(abeled), PQA-U(nlabeled) and PQA-A(rtificial), respectively. We show the architecture of PubMedQA dataset in Fig. 2.

Statistic	PQA-L	PQA-U	PQA-A
Number of QA pairs	1.0k	61.2k	211.3k
Prop. of yes (%)	55.2	–	92.8
Prop. of no (%)	33.8	–	7.2
Prop. of maybe (%)	11.0	–	0.0
Avg. question length	14.4	15.0	16.3
Avg. context length	238.9	237.3	238.0
Avg. long answer length	43.2	45.9	41.0

Table 1: PubMedQA dataset statistics.

Collection of PQA-L and PQA-U: PubMed articles which have i) a question mark in the titles and ii) a structured abstract with conclusive part are collected and denoted as pre-PQA-U. Now each instance has 1) a question which is the original title 2) a context which is the structured abstract without the conclusive part and 3) a long answer which is the conclusive part of the abstract.

Two annotators[3] labeled 1k instances from pre-PQA-U with yes/no/maybe to build PQA-L using Algorithm 1. The annotator 1 doesn't need to do much reasoning to annotate since the long answer is available. We denote this reasoning-free setting. However, the annotator 2 cannot use the long answer, so reasoning over the context is required for annotation. We denote such setting as reasoning-required setting. Note that the annotation process might assign wrong labels when both annotator 1 and annotator 2 make a same mistake, but considering human performance in §5.1, such error rate could be as low as 1%[4]. 500 randomly sampled PQA-L instances are used for 10-fold cross validation and the rest 500 instances consist of PubMedQA test set.

Algorithm 1 PQA-L data collection procedure

```
Input: pre-PQA-U
ReasoningFreeAnnotation $\leftarrow \{\}$
ReasoningRequiredAnnotation $\leftarrow \{\}$
GroundTruthLabel $\leftarrow \{\}$
while not finished do
    Randomly sample an instance $inst$ from pre-PQA-U
    if $inst$ is not yes/no/maybe answerable then
        Remove $inst$ and continue to next iteration
    end if
    Annotator 1 annotates $inst$ with $l_1 \in \{\text{yes}, \text{no}, \text{maybe}\}$ using question, context and long answer
    Annotator 2 annotates $inst$ with $l_2 \in \{\text{yes}, \text{no}, \text{maybe}\}$ using question and context
    if $l_1 = l_2$ then
        $l_a \leftarrow l_1$
    else
        Annotator 1 and Annotator 2 discuss for an agreement annotation $l_a$
        if not $\exists l_a$ then
            Remove $inst$ and continue to next iteration
        end if
    end if
    ReasoningFreeAnnotation$[inst] \leftarrow l_1$
    ReasoningRequiredAnnotation$[inst] \leftarrow l_2$
    GroundTruthLabel$[inst] \leftarrow l_a$
end while
```

Further, we include the unlabeled instances in pre-PQA-U with yes/no/maybe answerable questions to build PQA-U. For this, we use a simple rule-based method which removes all questions started with interrogative words (i.e. wh-words) or involving selections from multiple entities. This results in over 93% agreement with annotator 1 in identifying the questions that can be answered by yes/no/maybe.

Collection of PQA-A: Motivated by the recent successes of large-scale pre-training from ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), we use a simple heuristic to collect many noisily-labeled instances to build PQA-A for pre-training. Towards this end, we use PubMed articles with 1) a statement title which has POS tagging structures of NP-(VBP/VBZ)[5] and 2) a structured abstract including a conclusive part. The

[3] Both are qualified M.D. candidates.

[4] Roughly half of the products of two annotator error rates.

[5] Using Stanford CoreNLP parser (Manning et al., 2014).

Original Statement Title	Converted Question	Label	%
Spontaneous electrocardiogram alterations ***predict*** ventricular fibrillation in Brugada syndrome.	***Do*** spontaneous electrocardiogram alterations ***predict*** ventricular fibrillation in Brugada syndrome?	***yes***	92.8
Liver grafts from selected older donors ***do not have*** significantly more ischaemia reperfusion injury.	***Do*** liver grafts from selected older donors ***have*** significantly more ischaemia reperfusion injury?	***no***	7.2

Table 2: Examples of automatically generated instances for PQA-A. Original statement titles are converted to questions and answers are automatically generated according to the negation status.

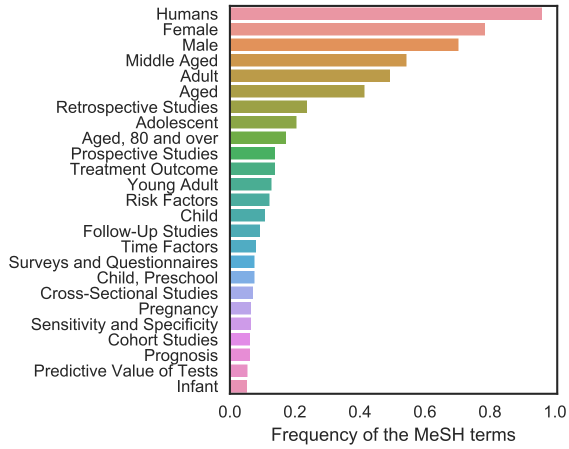

Figure 3: MeSH topic distribution of PubMedQA.

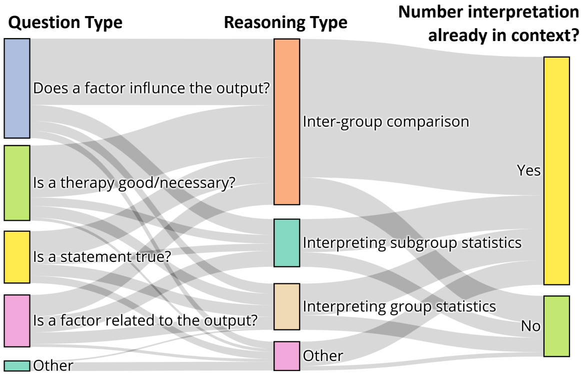

Figure 4: Proportional relationships between corresponded question types, reasoning types, and whether the text interpretations of numbers exist in contexts.

statement titles are converted to questions by simply moving or adding copulas ("is", "are") or auxiliary verbs ("does", "do") in the front and further revising for coherence (e.g.: adding a question mark). We generate the yes/no answer according to negation status of the VB. Several examples are shown in Table 2. We collected 211.3k instances for PQA-A, of which 200k randomly sampled instances are for training and the rest 11.3k instances are for validation.

3.2 Characteristics

We show the basic statistics of three PubMedQA subsets in Table 1.

Instance Topics: PubMed abstracts are manually annotated by medical librarians with Medical Subject Headings (MeSH)[6], which is a controlled vocabulary designed to describe the topics of biomedical texts. We use MeSH terms to represent abstract topics, and visualize their distribution in Fig. 3. Nearly all instances are human studies and they cover a wide variety of topics, including retrospective, prospective, and cohort studies, different age groups, and healthcare-related subjects like treatment outcome, prognosis and risk factors of diseases.

Question and Reasoning Types: We sampled 200 examples from PQA-L and analyzed the types of questions and types of reasoning required to answer them, which is summarized in Table 3. Various types of questions have been asked, including causal effects, evaluations of therapies, relatedness, and whether a statement is true. Besides, PubMedQA also covers several different reasoning types: most (57.5%) involve comparing multiple groups (e.g.: experiment and control), and others require interpreting statistics of a single group or its subgroups. Reasoning over quantitative contents is required in nearly all (96.5%) of them, which is expected due to the nature of biomedical research. 75.5% of contexts have text descriptions of the statistics while 21.0% only have the numbers. We use a Sankey diagram to show the proportional relationships between corresponded question type and reasoning type, as well as corresponded reasoning type and whether there are text interpretations of numbers in Fig. 4.

3.3 Evaluation Settings

The main metrics of PubMedQA are accuracy and macro-F1 on PQA-L test set using question and context as input. We denote prediction using question and context as a **reasoning-required** setting, because under this setting answers are not directly

[6] https://www.nlm.nih.gov/mesh

Question Type	**%**	**Example Questions**
Does a factor ***influence*** the output?	36.5	Does reducing spasticity ***translate into*** functional benefit? Does ibuprofen ***increase*** perioperative blood loss during hip arthroplasty?
Is a therapy ***good/necessary***?	26.0	***Should*** circumcision be performed in childhood? Is external palliative radiotherapy for gallbladder carcinoma ***effective***?
Is a ***statement*** true?	18.0	Sternal fracture in growing children: ***A rare and often overlooked fracture?*** Xanthogranulomatous cholecystitis: ***a premalignant condition***?
Is a factor ***related to*** the output?	18.0	Can PRISM ***predict*** length of PICU stay? Is trabecular bone ***related to*** primary stability of miniscrews?
Reasoning Type	**%**	**Example Snippet in Context**
Inter-group comparison	57.5	[...] Postoperative AF was significantly lower in the ***Statin group*** compared with the ***Non-statin group*** (16% versus 33%, p=0.005). [...]
Interpreting ***subgroup*** statistics	16.5	[...] 57% of patients were ***of lower socioeconomic status*** and they had more health problems, less functioning, and more symptoms [...]
Interpreting ***(single) group*** statistics	16.0	[...] ***A total of 4 children*** aged 5-14 years with a sternal fracture were treated in 2 years, 2 children were hospitalized for pain management and [...]
Text Interpretations of Numbers	**%**	**Example Snippet in Context**
Existing ***interpretations*** of ***numbers***	75.5	[...] Postoperative AF was ***significantly lower*** in the Statin group compared with the Non-statin group (***16% versus 33%, p=0.005***). [...]
No interpretations (***numbers only***)	21.0	[...] 30-day mortality was ***12.4%*** in those aged<70 years and ***22%*** in those>70 years (***p<0.001***). [...]
No numbers (***texts only***)	3.5	[...] The halofantrine therapeutic dose group showed ***loss and distortion of inner hair cells and inner phalangeal cells*** [...]

Table 3: Summary of PubMedQA question types, reasoning types and whether there are text descriptions of the statistics in context. Colored texts are matched key phrases (sentences) between types and examples.

expressed in the input and reasoning over the contexts is required to answer the question. Additionally, long answers are available at training time, so generation or prediction of them can be used as an auxiliary task in this setting.

A parallel setting, where models can use question and long answer to predict yes/no/maybe answer, is denoted as **reasoning-free** setting since yes/no/maybe are usually explicitly expressed in the long answers (i.e.: conclusions of the abstracts). Obviously, it's a much easier setting which can be exploited for bootstrapping PQA-U.

4 Methods

4.1 Fine-tuning BioBERT

We fine-tune BioBERT (Lee et al., 2019) on PubMedQA as a baseline. BioBERT is initialized with BERT (Devlin et al., 2018) and further pre-trained on PubMed abstracts and PMC[7] articles. Expectedly, it vastly outperforms BERT in various biomedical NLP tasks. We denote the original transformer weights of BioBERT as θ_0.

While fine-tuning, we feed PubMedQA questions and contexts (or long answers), separated by the special `[SEP]` token, to BioBERT. The yes/no/maybe labels are predicted using the special `[CLS]` embedding using a softmax function. Cross-entropy loss of predicted and true label distribution is denoted as $\mathcal{L}_{\text{QA}}$.

4.2 Long Answer as Additional Supervision

Under reasoning-required setting, long answers are available in training but not inference phase. We use them as an additional signal for training: similar to Ma et al. (2018) regularizing neural machine translation models with binary bag-of-word (BoW) statistics, we fine-tune BioBERT with an auxiliary task of predicting the binary BoW statistics of the long answers, also using the special `[CLS]` embedding. We minimize binary cross-entropy loss of this auxiliary task:

$$\mathcal{L}_{\text{BoW}} = -\frac{1}{N}\sum_i b_i \text{log}\hat{b_i} + (1-b_i)\text{log}(1-\hat{b_i})$$

where b_i and $\hat{b_i}$ are ground-truth and predicted probability of whether token i is in the long answers (i.e.: $b_i \in \{0,1\}$ and $\hat{b_i} \in [0,1]$), and N is the BoW vocabulary size. The total loss is:

$$\mathcal{L} = \mathcal{L}_{\text{QA}} + \beta\mathcal{L}_{\text{BoW}}$$

[7] https://www.ncbi.nlm.nih.gov/pmc/

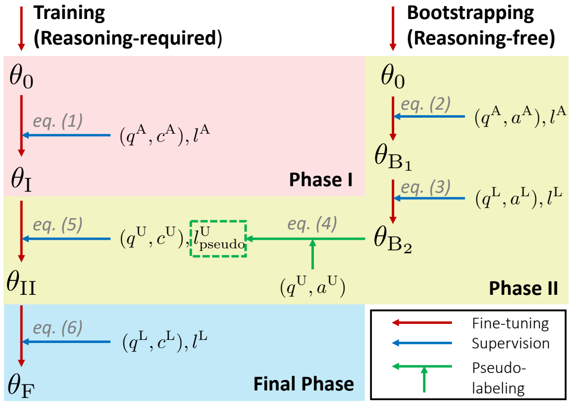

Figure 5: Multi-phase fine-tuning architecture. Notations and equations are described in §4.3.

In reasoning-free setting which we use for bootstrapping, the regularization coefficient β is set to 0 because long answers are directly used as input.

4.3 Multi-phase Fine-tuning Schedule

Since PQA-A and PQA-U have different properties from the ultimate test set of PQA-L, BioBERT is fine-tuned in a multi-phase style on different subsets. Fig. 5 shows the architecture of this training schedule. We use q, c, a, l to denote question, context, long answer and yes/no/maybe label of instances, respectively. Their source subsets are indexed by the superscripts of A for PQA-A, U for PQA-U and L for PQA-L.

Phase I Fine-tuning on PQA-A: PQA-A is automatically collected whose questions and labels are artificially generated. As a result, questions of PQA-A might differ a lot from those of PQA-U and PQA-L, and it only has yes/no labels with a very imbalanced distribution (92.8% yes v.s. 7.2% no). Despite these drawbacks, PQA-A has substantial training instances so models could still benefit from it as a pre-training step.

Thus, in Phase I of multi-phase fine-tuning, we initialize BioBERT with θ_0, and fine-tune it on PQA-A using question and context as input:

$$\theta_I \leftarrow \text{argmin}_\theta \, \mathcal{L}(\text{BioBERT}_\theta(q^A, c^A), l^A) \quad (1)$$

Phase II Fine-tuning on Bootstrapped PQA-U: To fully utilize the unlabeled instances in PQA-U, we exploit the easiness of reasoning-free setting to pseudo-label these instances with a bootstrapping strategy: first, we initialize BioBERT with θ_0, and fine-tune it on PQA-A using question and long answer (reasoning-free),

$$\theta_{B_1} \leftarrow \text{argmin}_\theta \, \mathcal{L}(\text{BioBERT}_\theta(q^A, a^A), l^A) \quad (2)$$

then we further fine-tune BioBERT$_{\theta_{B_1}}$ on PQA-L, also under the reasoning-free setting:

$$\theta_{B_2} \leftarrow \text{argmin}_\theta \, \mathcal{L}(\text{BioBERT}_\theta(q^L, a^L), l^L) \quad (3)$$

We pseudo-label PQA-U instances using the most confident predictions of BioBERT$_{\theta_{B_2}}$ for each class. Confidence is simply defined by the corresponding softmax probability and then we label a subset which has the same proportions of yes/no/maybe labels as those in the PQA-L:

$$l^U_{\text{pseudo}} \leftarrow \text{BioBERT}_{\theta_{B_2}}(q^U, a^U) \quad (4)$$

In phase II, we fine-tune BioBERT$_{\theta_I}$ on the bootstrapped PQA-U using question and context (under reasoning-required setting):

$$\theta_{II} \leftarrow \text{argmin}_\theta \, \mathcal{L}(\text{BioBERT}_\theta(q^U, c^U), l^U_{\text{pseudo}}) \quad (5)$$

Final Phase Fine-tuning on PQA-L: In the final phase, we fine-tune BioBERT$_{\theta_{II}}$ on PQA-L:

$$\theta_F \leftarrow \text{argmin}_\theta \, \mathcal{L}(\text{BioBERT}_\theta(q^L, c^L), l^L) \quad (6)$$

Final predictions on instances of PQA-L validation and test sets are made using BioBERT$_{\theta_F}$:

$$l_{\text{pred}} = \text{BioBERT}_{\theta_F}(q^L, c^L)$$

4.4 Compared Models

Majority: The majority (about 55%) of the instances have the label “yes”. We use a trivial baseline denoted as Majority where we simply predict “yes” for all instances, regardless of the question and context.

Shallow Features: For each instance, we include the following shallow features: 1) TF-IDF statistics of the question 2) TF-IDF statistics of the context/long answer and 3) sum of IDF of the overlapping non-stop words between the question and the context/long answer. To allow multi-phase fine-tuning, we apply a feed-forward neural network on the shallow features instead of using a logistic classifier.

BiLSTM: We simply concatenate the question and context/long answer with learnable segment embeddings appended to the biomedical word2vec embeddings (Pyysalo et al., 2013) of each token. The concatenated sentence is then fed to a biLSTM, and the final hidden states of the forward and backward network are used for classifying the yes/no/maybe label.

ESIM with BioELMo: Following the state-of-the-art recurrent architecture of NLI (Peters et al., 2018), we use pre-trained biomedical contextualized embeddings BioELMo (Jin et al., 2019) for word representations. Then we apply the ESIM model (Chen et al., 2016), where a biLSTM is used to encode the question and context/long answer, followed by an attentional local inference layer and a biLSTM inference composition layer. After pooling, a softmax output unit is applied for predicting the yes/no/maybe label.

4.5 Compared Training Schedules

Final Phase Only: Under this setting, we train models only on PQA-L. It's an extremely low resources setting where there are only 450 training instances in each fold of cross-validation.

Phase I + Final Phase: Under this setting, we skip the training on bootstrapped PQA-U. Models are first fine-tuned on PQA-A, and then fine-tuned on PQA-L.

Phase II + Final Phase: Under this setting, we skip the training on PQA-A. Models are first fine-tuned on bootstrapped PQA-U, and then fine-tuned on PQA-L.

Single-phase Training: Instead of training a model sequentially on different splits, under single-phase training setting we train the model on the combined training set of all PQA splits: PQA-A, bootstrapped PQA-U and PQA-L.

5 Experiments

5.1 Human Performance

Human performance is measured during the annotation: As shown in Algorithm 1, annotations of annotator 1 and annotator 2 are used to calculate reasoning-free and reasoning-required human performance, respectively, against the discussed ground truth labels. Human performance on the test set of PQA-L is shown in Table ??. We only test single-annotator performance due to limited resources. Kwiatkowski et al. (2019) show that an ensemble of annotators perform significantly better than single-annotator, so the results reported in Table ?? are the lower bounds of human performance. Under reasoning-free setting where the annotator can see the conclusions, a single human achieves 90.4% accuracy and 84.2% macro-F1. Under reasoning-required setting, the task becomes much harder, but it's still possible for humans to solve: a single annotator can get 78.0% accuracy and 72.2% macro-F1.

Setting	**Accuracy (%)**	**Macro-F1 (%)**
Reasoning-Free	90.40	84.18
Reasoning-Required	78.00	72.19

Table 4: Human performance (single-annotator).

5.2 Main Results

We report the test set performance of different models and training schedules in Table 5. In general, multi-phase fine-tuning of BioBERT with additional supervision outperforms other baselines by large margins, but the results are still much worse than just single-human performance.

Comparison of Models: A trend of BioBERT > ESIM w/ BioELMo > BiLSTM > shallow features > majority, conserves across different training schedules on both accuracy and macro-F1. Fine-tuned BioBERT is better than state-of-the-art recurrent model of ESIM w/ BioELMo, probably because BioELMo weights are fixed while all BioBERT parameters can be fine-tuned, which better benefit from the pre-training settings.

Comparison of Training Schedules: Multi-phase fine-tuning setting gets 5 out of 9 model-wise best accuracy/macro-F1. Due to lack of annotated data, training only on the PQA-L (final phase only) generates similar results as the majority baseline. In phase I + Final setting where models are pre-trained on PQA-A, we observe significant improvements on accuracy and macro-F1 and some models even achieve their best accuracy under this setting. This indicates that a hard task with limited training instances can be at least partially solved by pre-training on a large automatically collected dataset when the tasks are similarly formatted.

Improvements are also observed in phase II + Final setting, though less significant than those of phase I + Final. As expected, multi-phase fine-tuning schedule is better than single-phase, due to different properties of the subsets.

Additional Supervision: Despite its simplicity, the auxiliary task of long answer BoW prediction clearly improves the performance: most results (28/40) are better with such additional supervision than without.

Model	Final Phase Only		Single-phase		Phase I + Final		Phase II + Final		Multi-phase	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Majority	55.20	23.71	–	–	–	–	–	–	–	–
Human (single)	78.00	72.19	–	–	–	–	–	–	–	–
w/o A.S.										
Shallow Features	53.88	36.12	57.58	31.47	57.48	37.24	56.28	40.88	53.50	39.33
BiLSTM	55.16	23.97	55.46	39.70	58.44	40.67	52.98	33.84	59.82	41.86
ESIM w/ BioELMo	53.90	32.40	61.28	42.99	61.96	43.32	60.34	44.38	62.08	45.75
BioBERT	56.98	28.50	66.44	47.25	66.90	46.16	66.08	50.84	67.66	52.41
w/ A.S.										
Shallow Features	53.60	35.92	57.30	30.45	55.82	35.09	56.46†	40.76	55.06†	40.67†
BiLSTM	55.22†	23.86	55.96†	40.26†	61.06†	41.18†	54.12†	34.11†	58.86	41.06
ESIM w/ BioELMo	53.96†	31.07	62.68†	43.59†	63.72†	47.04†	60.16	45.81†	63.72†	47.90†
BioBERT	57.28†	28.70†	66.66†	46.70†	67.24†	46.21†	66.44†	51.41†	**68.08**†	**52.72**†

Table 5: Main results on PQA-L test set under reasoning-required setting. A.S.: additional supervision. †with A.S. is better than without A.S. Underlined numbers are model-wise best performance, and bolded numbers are global best performance. All numbers are percentages.

Model	w/o A.S.		w/ A.S.	
	Acc	F1	Acc	F1
Majority	92.76	48.12	–	–
Shallow Features	93.01	54.59	93.05	55.12
BiLSTM	94.59	73.40	94.45	71.81
ESIM w/ BioELMo	94.82	74.01	95.04	75.22
BioBERT	96.50	84.65	96.40	83.76

Table 6: Results of Phase I (eq. 1). Experiments are on PQA-A under reasoning-required setting. A.S.: additional supervision.

Model	Eq. 2		Eq. 3	
	Acc	F1	Acc	F1
Majority	92.76	48.12	55.20	23.71
Human (single)	–	–	90.40†	84.18†
Shallow Features	93.11	56.11	54.44	38.63
BiLSTM	95.97	83.70	71.46	50.93
ESIM w/ BioELMo	97.01	88.47	74.06	58.53
BioBERT	98.28	93.17	80.80	63.50

Table 7: Bootstrapping results. Experiments are on PQA-A (eq. 2) and PQA-L (eq. 3) under reasoning-free setting. †Reasoning-free human performance.

5.3 Intermediate Results

In this section we show the intermediate results of multi-phase fine-tuning schedule.

Phase I: Results are shown in Table 6. Phase I is fine-tuning on PQA-A using question and context. Since PQA-A is imbalanced due to its collection process, a trivial majority baseline gets 92.76% accuracy. Other models have better accuracy and especially macro-F1 than majority baseline. Fine-tuned BioBERT performs best.

Model	w/o A.S.		w/ A.S.	
	Acc	F1	Acc	F1
Majority	55.10	23.68	–	–
Shallow Features	76.66	66.12	77.71	67.97
Majority	56.53	24.07	–	–
BiLSTM	85.33	81.32	85.68	81.87
Majority	55.10	23.68	–	–
ESIM w/ BioELMo	78.47	63.32	79.62	64.91
Majority	54.82	24.87	–	–
BioBERT	80.93	68.84	81.02	70.04

Table 8: Phase II results (eq. 5). Experiments are on pseudo-labeled PQA-U under reasoning-required setting. A.S.: additional supervision.

Bootstrapping: Results are shown in Table 7. Bootstrapping is a three-step process: fine-tuning on PQA-A, then on PQA-L and pseudo-labeling PQA-U. All three steps are using question and long answer as input. Expectedly, models perform better in this reasoning-free setting than they do in reasoning-required setting (for PQA-A, Eq. 2 results in Table 7 are better than the performance in Table 6; for PQA-L, Eq. 3 results in Table 7 are better than the performance in Table 5).

Phase II: Results are shown in Table 8. In Phase II, since each model is fine-tuned on its own pseudo-labeled PQA-U instances, results are not comparable between models. While the ablation study in Table 5 clearly shows that Phase II is helpful, performance in Phase II doesn't necessarily correlate with final performance on PQA-L.

6 Conclusion

We present PubMedQA, a novel dataset aimed at biomedical research question answering using yes/no/maybe, where complex quantitative reasoning is required to solve the task. PubMedQA has substantial automatically collected instances as well as the largest size of expert annotated yes/no/maybe questions in biomedical domain. We provide a strong baseline using multi-phase fine-tuning of BioBERT with long answer as additional supervision, but it's still much worse than just single human performance.

There are several interesting future directions to explore on PubMedQA, e.g.: (1) about 21% of PubMedQA contexts contain no natural language descriptions of numbers, so how to properly handle these numbers is worth studying; (2) we use binary BoW statistics prediction as a simple demonstration for additional supervision of long answers. Learning a harder but more informative auxiliary task of long answer generation might lead to further improvements.

Articles of PubMedQA are biased towards clinical study-related topics (described in Appendix B), so PubMedQA has the potential to assist evidence-based medicine, which seeks to make clinical decisions based on evidence of high quality clinical studies. Generally, PubMedQA can serve as a benchmark for testing scientific reasoning abilities of machine reading comprehension models.

7 Acknowledgement

We are grateful for the anonymous reviewers of EMNLP who gave us very valuable comments and suggestions.

References

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2016. Enhanced lstm for natural language inference. *arXiv preprint arXiv:1609.06038*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

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

William Hersh, Aaron Cohen, Lynn Ruslen, and Phoebe Roberts. 2007. Trec 2007 genomics track overview. In *TREC 2007*.

William Hersh, Aaron M. Cohen, Phoebe Roberts, and Hari Krishna Rekapalli. 2006. Trec 2006 genomics track overview. In *TREC 2006*.

Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*.

Seongsoon Kim, Donghyeon Park, Yonghwa Choi, Kyubum Lee, Byounggun Kim, Minji Jeon, Jihye Kim, Aik Choon Tan, and Jaewoo Kang. 2018. A pilot study of biomedical text comprehension using an attention-based deep neural reader: Design and experimental analysis. *JMIR medical informatics*, 6(1):e2.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gáabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association of Computational Linguistics*, 6:317–328.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Rhinehart, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, et al. 2019. Natural questions: a benchmark for question answering research.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Shuming Ma, Xu Sun, Yizhong Wang, and Junyang Lin. 2018. Bag-of-words as target for neural machine translation. *arXiv preprint arXiv:1805.04871*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

Roser Morante, Martin Krallinger, Alfonso Valencia, and Walter Daelemans. 2012. Machine reading of biomedical texts about alzheimers disease. In *CLEF*

2012 Conference and Labs of the Evaluation Forum-Question Answering For Machine Reading Evaluation (QA4MRE), Rome/Forner, J.[edit.]; ea, pages 1–14.

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

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. Bioread: A new dataset for biomedical reading comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Anselmo Peñas, Eduard Hovy, Pamela Forner, Álvaro Rodrigo, Richard Sutcliffe, and Roser Morante. 2013. Qa4mre 2011-2013: Overview of question answering for machine reading evaluation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 303–320. Springer.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Sampo Pyysalo, Filip Ginter, Hans Moen, Tapio Salakoski, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. *arXiv preprint arXiv:1809.01494*.

Hiroaki Sakamoto, Yasunori Watanabe, and Masataka Satou. 2011. Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting? *Annals of thoracic and cardiovascular surgery*, 17(4):376–382.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

A Yes/no/maybe Answerability

Not all naturally occuring question titles from PubMed are answerable by yes/no/maybe. The first step of annotating PQA-L (as shown in algorithm 1) from pre-PQA-U is to manually identify questions that can be answered using yes/no/maybe. We labeled 1091 (about 50.2%) of 2173 question titles as unanswerable. For example, those questions cannot be answered by yes/no/maybe:

- "Critical Overview of HER2 Assessement in Bladder Cancer: What Is Missing for a Better Therapeutic Approach?" (wh- question)
- "Otolaryngology externships and the match: Productive or futile?" (multiple choices)

B Over-represented Topics

Clinical study-related topics are over-represented in PubMedQA: we found proportions of MeSH terms like:

- "Pregnancy Outcome"
- "Socioeconomic Factors"
- "Risk Assessment"
- "Survival Analysis"
- "Prospective Studies"
- "Case-Control Studies"
- "Reference Values"

are significantly higher in the PubMedQA articles than those in 200k most recent general PubMed articles (significance is defined by $p < 0.05$ in two-proportion z-test).

C Annotation Criteria

Strictly speaking, most yes/no/maybe research questions can be answered by "maybe" since there will always be some conditions where one statement is true and vice versa. However, the task will be trivial in this case. Instead, we annotate a question using "yes" if the experiments and results in the paper indicate it, so the answer is not universal but context-dependent.

Given a question like "Do patients benefit from drug X?": certainly not all patients will benefit from it, but if there is a significant difference in an outcome between the experimental and control group, the answer will be "yes". If there is not, the answer will be "no".

"Maybe" is annotated when (1) the paper discusses conditions where the answer is True and conditions where the answer is False or (2) more than one intervention/observation/etc. is asked, and the answer is True for some but False for the others (e.g.: "Do Disease A, Disease B and/or Disease C benefit from drug X?"). To model uncertainty of the answer, we don't strictly follow the logic calculations where such questions can always be answered by either "yes" or "no".

Model		**Accuracy (%)**	**Macro-F1 (%)**
Majority		55.20	23.71
Supervised Learning only		56.98	28.50
	w/ A.S.	57.28	28.70
Pre-trained on PQA-A		66.90	46.16
	w/ A.S.	67.24	46.21
Multi-Phase Fine-tuning		67.66	52.41
	w/ A.S.	**68.08**	**52.72**
Human Performance		78.00	72.19