Generating Biomedical Question Answering Corpora From Q&A Forums

ANDRE LAMURIAS[ID]**, DIANA SOUSA, AND FRANCISCO M. COUTO**[ID]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisbon, Portugal

Corresponding author: Andre Lamurias (alamurias@lasige.di.fc.ul.pt)

ABSTRACT Question Answering (QA) is a natural language processing task that aims at obtaining relevant answers to user questions. While some progress has been made in this area, biomedical questions are still a challenge to most QA approaches, due to the complexity of the domain and limited availability of training sets. We present a method to automatically extract question-article pairs from Q&A web forums, which can be used for document retrieval, a crucial step of most QA systems. The proposed framework extracts from selected forums the questions and the respective answers that contain citations. This way, QA systems based on document retrieval can be developed and evaluated using the question-article pairs annotated by users of these forums. We generated the BiQA corpus by applying our framework to three forums, obtaining 7,453 questions and 14,239 question-article pairs. We evaluated how the number of articles associated with each question and the number of votes on each answer affects the performance of baseline document retrieval approaches. Also, we demonstrated that the articles given as answers are significantly similar to the questions and trained a state-of-the-art deep learning model that obtained similar performance to using a dataset manually annotated by experts. The proposed framework can be used to update the BiQA corpus from the same forums as new posts are made, and from other forums that support their answers with documents. The BiQA corpus and the framework used to generate it are available at https://github.com/lasigeBioTM/BiQA.

INDEX TERMS Question answering, document retrieval, text mining, datasets, deep learning.

I. INTRODUCTION

Question Answering (QA) consists of the automatic retrieval of information that can directly answer user questions. This task is relevant for the biomedical domain, due to the large quantity and variety of information that is necessary to integrate to fully understand biomedical problems. QA systems can assist the study of biological systems by automatically retrieving information relevant to the queries made by a researcher. This is possible because of all the biomedical information that is available in text format on publicly available digital libraries, in the form of scientific literature. However, these systems require advanced techniques when dealing with large document repositories.

In recent years, many developments have been made in QA systems using deep learning techniques [1], [2]. QA systems are developed using gold standards, where automatic

The associate editor coordinating the review of this manuscript and approving it for publication was Abdel-Hamid Soliman[ID].

answers are compared to the manually annotated answers by domain experts. One of the most relevant gold standards for biomedical QA is BioASQ [3], which is also a community challenge where every year, a new dataset is released to train and evaluate QA systems. One limitation of this type of evaluation is that the questions are not obtained from real users. The BioASQ datasets are developed by experts, who create the questions and the answers. Therefore it lacks the variability that is inherent to real-world scenarios. furthermore, it requires the involvement of domain experts, which may not be possible in some cases.

Question-and-Answers (Q&A) forums, such as Stack Exchange and Quora, are websites where users can post questions and answer the questions of other users. In recent years, search engines have improved their ability to answer natural language queries, instead of relying just on keywords as input. It is more intuitive for users to retrieve information using natural language than through keyword-based search. QA systems could, therefore, benefit from Q&A

forums where users collaborate in posting and answering questions.

Although most of these forums have no selection process to restrict the participants, the users can vote on each answer, and, in some cases, the user who made the question can give their approval to one of the answers. One significant example of Q&A forums is Stack Exchange, which is a network of websites where users can ask questions about specific subjects, such as computer science, mathematics, art, and various languages. Reddit is another forum where users can post questions and get answers from communities focused on specific subjects, such as nutrition. Our work focused on Stack Exchange and Reddit since both these forums contain questions relevant to the biomedical domain, and provide an API to retrieve their contents.

For biomedical questions, it is essential to provide proper references that corroborate the given answers. Otherwise, the correctness of the answer may be questioned. Usually, Q&A forums that are related to biomedical subjects, including the ones used in this work, encourage users to support their posts with scientific articles. A significant advantage of open-access digital libraries such as PubMed is that anyone can analyze its documents to better understand biomedical problems. For this reason, improving biomedical document retrieval is of interest to the general community.

With this article, we present a framework to generate a gold standard for document retrieval QA, with questions and relevant articles retrieved automatically from Q&A forums. We demonstrated our approach on three forums that are relevant to biomedical sciences: Biology from Stack Exchange, Medical Sciences also from Stack Exchange, and Nutrition from Reddit.

We also present the BiQA corpus of question-article pairs, which can be used to develop and evaluate document retrieval and QA systems, as we demonstrate in this article. Unlike existing corpora, our corpus is based on real user questions and can evolve over time. The proposed framework could also be applied to other relevant forums, in any language, and update the dataset with new questions. To the best of our knowledge, this is the first work to automatically obtain document-based QA training data from web forums.

II. RELATED WORK
A. USER-BASED QA
Recently, there has been an increased focus on QA using real user questions. For example, the Natural Questions dataset released by Google uses real queries made by users of the Google search engine [4]. While the questions of other QA datasets are formulated by annotators whose function is to read a paragraph and write relevant questions [5], this dataset contains real queries from users seeking information. The authors then annotated each question with answers taken from the English Wikipedia. The objective of this dataset is to develop systems that can extract the correct answer to each question from a specific Wikipedia page.

Another example is the TREC LiveQA track [6], which promotes the development of new approaches to answer questions retrieved from various sources, such as Yahoo Answers and the U.S. National Library of Medicine. The results of the teams that have participated in this track highlight the low scores obtained when compared with human performance, particularly for consumer health questions [7].

B. DATASETS BASED ON WEB FORUMS
Other authors have also explored Q&A forums to generate datasets. Reference [8] presented an algorithm to extract Q&A pairs from forum threads, where it may not be clear which posts are questions and which are answers. Reference [9] generated a dataset based on Yahoo Answers and used crowdsourcing to rank the answers given to the same question. However, since Stack Exchange and Reddit users can vote on answers for any question (not just the ones they made), our approach does not require that extra annotation step. Reference [10] explored Stack Overflow to develop a learning-to-rank method to evaluate the quality of the answers, based on the scores given by the users. The authors extracted several textual and user-based features to classify each answer, including the number of links to external sources, which they found to be a good predictor of the quality of the answer. Nevertheless, to the best of our knowledge, we were the first to extract references to articles from the answers to automatically generate a QA dataset. Reference [11] demonstrated an approach to review-based QA using questions posted on Amazon product reviews. In this study, they highlight the difficulty of answering user questions in a real-world scenario due to many questions not being answerable. However, they modeled this task as a closed-world QA task, where it is known which snippets contain the answers. In our case, we consider an open-world scenario where any document from the biomedical literature can contain the answer.

C. BIOMEDICAL QA
The development of systems for biomedical QA requires gold standard datasets to both train and evaluate those systems. The biomedical domain poses a challenge to this type of system since the questions may be more challenging to answer, and the datasets available are more limited. BioASQ is one of the most significant community efforts in advancing biomedical QA systems. It is a series of evaluations where the participants compete on various QA tasks, using a gold standard provided by the organization. Each edition of BioASQ has had at least these two subtasks: one which consists of indexing PubMed abstracts with MeSH terms, while the second consists of answering biomedical questions. Each year the organizers provide new annotated questions that are curated by biomedical experts. In 2019 (BioASQ 7), task B phase A consisted of retrieving relevant documents, concepts, and snippets from designated article repositories to answer biomedical questions. The best Mean Average Precision (MAP) scores ranged between 0.1218 and 0.2898 on five

test batches. This dataset is similar to ours because PubMed abstracts are used to answer natural language questions. However, while the BioASQ questions are created and answered by experts, our dataset is made of questions and answers created by users of online forums. The consequence of this difference is that our questions are taken directly from real users, and the answers are given by non-experts, which are easier to obtain.

There have been other efforts in developing datasets for biomedical QA tasks. MediQA [12] is another relevant challenge, which focuses on clinical data. This dataset consists of question and candidate answers, which the participants had to classify and rank according to the relevance to the question. While this dataset was developed manually to include both correct and incorrect answers, the organizers also provided other datasets derived from online resources which contain only the correct answers [13]. The dataset released for this task contained 383 questions, while the answers were retrieved using a QA system and ranked by experts. emrQA is another dataset that also focused on the clinical domain [14]. It consists of 6,686 questions and explores the i2b2 dataset to generate QA pairs. Recently, [15] released the PubMedQA dataset, where they derive questions based on article titles and answer them with their respective abstracts.

III. CORPUS GENERATION

Fig. 1 summarizes our framework that we used to generate the BiQA corpus from Q&A forums. For each post of a forum, we retrieved its title, body text, and the number of votes. If the forum is not specifically for Q&A, such as Reddit, we selected only posts with question marks in the title or body. For each post, we retrieved all of its first-level replies, i.e., we ignored comments made on the question and answers because these usually are not direct answers to the question. We also did not take into account user data, although other authors have pointed out that that information could be relevant to rank answers [10]. However, we saved the original IDs of every post and reply so that additional information could be retrieved later.
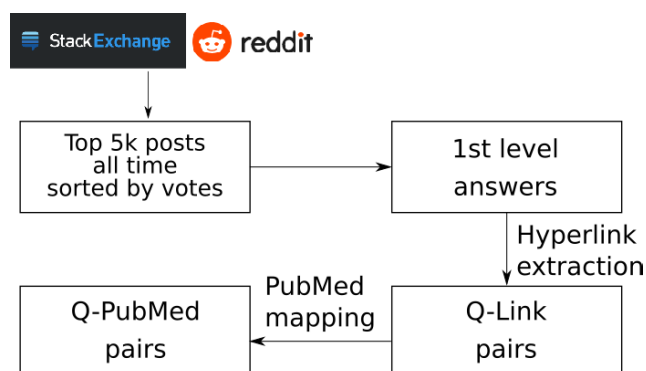

FIGURE 1. Framework used to generate the BiQA corpus.

Afterward, we parsed each answer to obtain hyperlinks mentioned throughout the answer text. Whenever possible,

we linked various types of URLs to PubMed: from PMC, doi.org, ScienceDirect, and ResearchGate. More types of URLs could be handled in the future using more complex mapping rules or using machine learning. Other users provided the full formatted citation without a hyperlink, which would require more complex methods to link to PubMed automatically.

Example 1 shows a question from our corpus, and a document used to answer it. The answer has some text written by the user as well as the passage from the article that contained the answer, which we omitted. This is a simple example where the user admits that they are not experts and simply used academic search engines. Indeed, if we search the question title text on PubMed, the article linked can be found in the results list. Example 2 provides another example where domain knowledge is necessary to retrieve the articles to answer the question. In this case, the question text should be processed to remove stop words, and it is necessary to have some understanding of the relationship between cholesterol levels and consumption of eggs.

Example 1: Question posted on Biology.SE where a user based their answer on an article retrieved from PubMed.

- **Question link:** https://biology.stackexchange.com/questions/84829
- **Question title:** Number of dopaminergic neurons in VTA
- **Answer text:** Approximately, 250000-440000 neurons in Humans (Rice *et al.*, 2016).
- **Answer PMID:** 25269834 (Mapping dopaminergic deficiencies in the substantia nigra/ventral tegmental area in schizophrenia.)

Example 2: Question posted on MS.SE where a user based their answer on several biomedical articles (we show only one of those articles).

- **Question link:** https://medicalsciences.stackexchange.com/questions/529
- **Question title:** How many eggs can one eat per day?
- **Answer text:** (…) The two studies cited indicate that consumption of eggs does not raise cholesterol levels significantly for about 3/4ths of the population. (…)
- **Answer PMID:** 16340654 (Dietary cholesterol provided by eggs and plasma lipoproteins in healthy populations.)

We did not filter the answers by their score; however, in our analysis, we tested various thresholds for the minimum number of votes and its effect on document retrieval engines (Section V-B). When an answer has more votes, it means that more users agreed that that answer is correct or at least relevant. While we could have applied stricter criteria for answer selection, we observed some answers supported by citations but without additional votes (every answer starts with one vote). This is possibly due to the limited size of the forums' user base. However, there were cases of negative scores, which can indicate that the answer is incorrect or non-related to the question, according to the users that voted.

IV. CORPUS EVALUATION

We evaluated the BiQA corpus using two strategies. First, we compared the results obtained with three document search methods on the questions of the dataset. This way, we could compare the references obtained through user answers with the ones obtained directly using search engines. The scores obtained indicate if search engine techniques can obtain the same answers that were provided by the users. Secondly, we evaluated the dataset by applying it to a biomedical question answering task. The objective was to determine if, with our dataset, we can obtain similar performance to a dataset developed by domain experts.

A. IR-BASED VALIDATION

In order to understand the potential of our dataset for document retrieval, we evaluated three robust baseline methods. We considered the articles obtained from the forum answers as the list of relevant documents of each question. The three methods were: i) the NCBI Eutils API (https://eutils.ncbi.nlm.nih.gov/entrez/eutils) which provides access to the PubMed search engine, ii) query likelihood scorer with Dirichlet term smoothing calculated by Galago [16] on our local version of PubMed, and iii) BM25 scorer [17], also calculated by Galago. The PubMed search engine also uses BM25 as a first stage, with additional improvements [18].

We tested querying each method with the title and body text of the questions, as well as with the answer text since the articles were usually more related to the answer text than to the question text (Fig. 2). Although in a QA task the answer text would not be available, we still wanted to study, if given the answer text, we could retrieve the relevant articles.
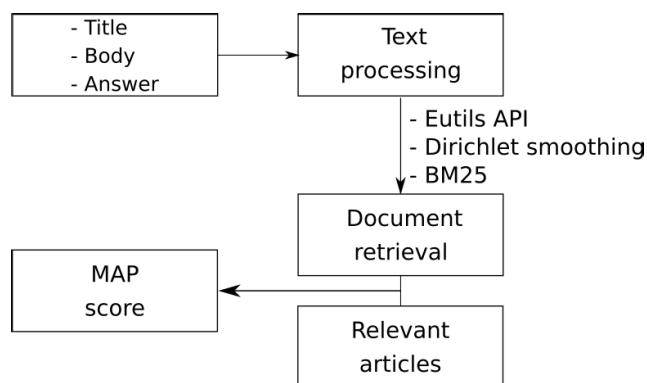

FIGURE 2. Baseline evaluation of the BiQA corpus.

Our framework first tokenized the text using Spacy (https://spacy.io/) and then removed HTML tags, stopwords, punctuation, and spaces. We used the ''OR'' operator with the PubMed search since otherwise, the articles would have to contain the same words as the query, and we would get very few results. We retrieved a maximum of 100 articles per question, ordered by the relevance score provided by the PubMed API, and the Dirichlet term smoothing score and BM25 score calculated by Galago. We retrieved 100 articles because this was the same number of articles retrieved by [19]

(see Section IV-B). Then we compared with the articles given by the users that answered each question, which we considered as correct if they had more than a given number of votes. If the correct articles appear at the top of the search results, it means that they could have been obtained through a document search engine. However, if the correct articles appear at the bottom or do not appear at all, it means that the document retrieval process fails for this type of question.

B. BioASQ6b VALIDATION

Document retrieval is a significant step of a QA pipeline since the answers are usually derived from the retrieved documents. For this reason, we used our dataset with a biomedical QA system that performed document retrieval tasks to evaluate the effect on its performance. We chose the AUEB system developed for BioASQ6 [19] since it obtained the highest MAP scores on that competition for the document retrieval tasks. The BioASQ6 dataset, where the AUEB system was evaluated, consists of a training set of 2,151 questions, a development set of 100 questions, and five testing batches of 100 questions each. On this dataset, the results of each testing batch are reported separately.

The AUEB team explored various deep learning models for BioASQ6b. We selected their extension of the Position-Aware Convolutional Recurrent Relevance (PACRR) model [20], dubbed TERM-PACRR, as it obtained better results in most test batches for document retrieval. This model computes a query-document similarity matrix and uses a convolution neural network and multi-layer perceptron to compute the relevance score of each query term.

We added the BiQA corpus to the BioASQ corpus to determine if the additional questions would bring improvements to the performance of the AUEB system. To accomplish this, we first executed the AUEB system with the original BioASQ6 data to reproduce the results. Then we reran it using the full BiQA dataset, in addition to the BioASQ data. The AUEB system first retrieves 100 documents for each question using Galago and uses the BM25 score to then rerank these documents. Therefore we did the same for the questions of the BiQA corpus, using our local version of PubMed indexed by Galago. At each epoch, the model was evaluated on the development set, and, at the end of each training run, the model that obtained the best performance on the development set was selected. We followed the procedure described by the original AUEB submission, which consisted in training ten versions of each model using different random seeds and combining the results of the ten models on each test batch. The purpose of this procedure is to reduce the variability of the results of different training runs.

Finally, we also experimented only using the BiQA corpus to train a model with the AUEB system. We followed the same procedure to evaluate this version of the model.

C. EVALUATION MEASURES

We calculated the Mean Average Precision (MAP) on each subset of the BiQA corpus and each validation method, which

is also the metric used on BioASQ Task B Phase A. This way, we can indirectly compare the scores obtained on our dataset with those obtained on the BioASQ datasets. To calculate the MAP score, first, we calculate the average precision obtained with each question, and then the macro-average over all questions. The average precision is given by:

$$AP = \frac{\sum_{r=1}^{|L|} P(r) \cdot rel(r)}{|L_R|} \qquad (1)$$

where $|L|$ is the number of retrieved articles and $|L_R|$ is the number of relevant articles obtained from the answers, $P(r)$ is the precision obtained with the first r retrieved articles, and $rel(r)$ is 1 if the rth retrieved article is relevant, and 0 otherwise. Considering Example 1, since it only has one relevant article, its AP would be equal to $\frac{1}{n}$ where n is the position of the relevant article in the retrieved list. For the BioASQ6b validation, we used a variation of this measure where L_R is always equal to 10, which is how it is implemented on the evaluation tool. Then we calculate MAP as:

$$MAP = \frac{\sum_{i=1}^{|Q|} AP(q_i)}{|Q|} \qquad (2)$$

where $q_i \in Q$ and Q is the list of all questions.

V. BiQA CORPUS ANALYSIS
A. DATA SOURCES
We focused on three online forums to extract QA data: Biology and Medical Sciences from Stack Exchange (we refer to them as Biology.SE and MS.SE) and Nutrition from Reddit (r/nutrition). We assumed that the users of these communities would give more value to answers supported by scientific articles, instead of accepting answers that could be based on personal experiences, for example. On Reddit, only an account is necessary to create and answer posts, as well as to vote on posts and replies. However, r/nutrition does not allow posts from new accounts or accounts with negative karma (balance of positive and negative votes). On Stack Exchange, it is possible to answer as a guest without creating an account, but it has a reputation system so that only established users can vote on other answers. Both forums have volunteer moderators who define community guidelines and can delete posts. Table 1 provides some statistics on these forums, obtained at the time of writing.

TABLE 1. Statistics of the Q&A forums used to generate the BiQA corpus. Reddit does not provide the total number of posts and the number of daily visitors.

Forum stats	#Posts	Users	Visits/day	Posts/day	Age
Biology.SE	24k	44k	10k	11	8y6m
MS.SE	6.5k	15k	2.8k	3.9	5y3m
r/nutrition	-	887k	-	20	11y

B. CORPUS STATISTICS
The BiQA corpus consists of a collection of three datasets, obtained from each of the forums previously described. We show the number of questions and question-article pairs of each dataset in Table 2. We considered only questions with answers that contained citations that we could automatically extract and map to PMIDs. As we did not normalize the number of votes, we can see a relation between the average number of votes and size of the community; r/nutrition has a larger community than Biology.SE and MS.SE; hence the average number of votes is higher.

Each answer has a number of votes associated with it, as well as a list of PMIDs, retrieved at the timestamp shown in Table 2. Fig. 3 shows the distribution of each dataset in terms of the number of votes on the answers, while Fig. 4 shows the distribution of the number of PMIDs per answer. We can see a similar distribution of votes as the one reported by [10]. The MS.SE subset is the smallest of the three forums; hence it also does not have answers with as many votes as the other two. However, it has a considerable number of answers with many PMIDs, as can be seen in Fig. 4. The r/nutrition subset has a higher average number of PMIDs per answer than Biology.SE. However, many answers had links to other sources of information, such as blog posts and Wikipedia pages. We also observed that these communities did not always value answers with a large number of citations because more succinct answers are easier to understand.

C. Q-ARTICLE SIMILARITY
To confirm if the articles provided as answers are actually relevant to the questions, we calculated the cosine similarity between each question title and respective article title. We used the vectors from the `en_vectors_web_lg` package of Spacy, which are based on a vocabulary of 1M words and 300d vector representations. These vectors were calculated using the GloVe algorithm [21] on the Common Crawl dataset.[1] Then we calculated the similarity of each question text with an answer article from a random question of the same dataset. We calculated the mean and standard deviation of the Q-article pairs and Q-random article pairs, as well as the t-test statistic, for each dataset (Table 3). The p-value for every dataset was <0.001, showing that the articles provided by the crowd are significantly more similar to the question than random articles.

VI. SEARCH ENGINE EVALUATION
As a baseline evaluation of the dataset, we attempted to retrieve the same PubMed documents that were referenced by

[1]https://commoncrawl.org/the-data/

TABLE 2. Statistics of the three datasets that compose the BiQA corpus, along with the total number of questions and Q-PMID pairs and macro-averages.

Forum	Query date	Qs w/PMID	Q-article pairs	Avg # votes	Avg # PMID
Biology.SE	04/05/2020	3961	6925	4.96	1.62
MS.SE	04/05/2020	1383	3053	3.91	2.06
r/nutrition	04/05/2020	2109	4261	6.13	1.79
Total	-	7396	14133	5.01	1.82

FIGURE 3. Votes distribution of the three datasets.
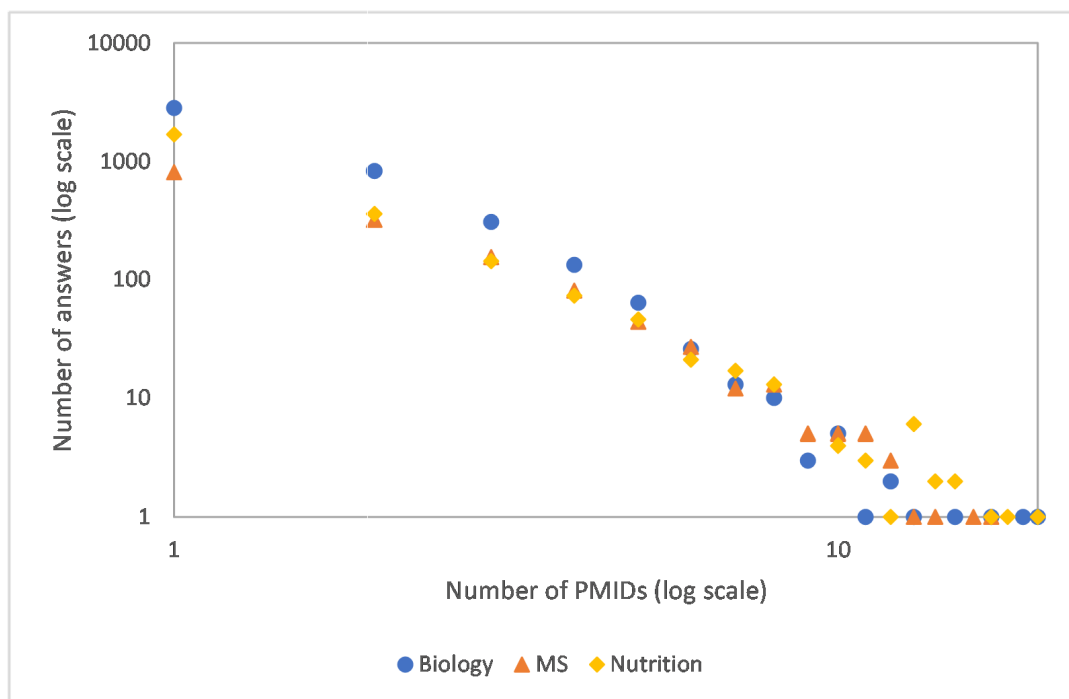

FIGURE 4. PMID count distribution of the three datasets.

the answers, using various search engines. This way, we can observe if a search engine could obtain the same documents as those provided by the users.

Since we do not have a score for each PMID but only for the answer as a whole (which may contain several PMIDs), we have no way to rank the PMIDs from the same answer.

TABLE 3. Mean and standard deviation of the similarity between question and respective answer articles, compared with a random answer.

	Q-A		Q-random A		
	Mean	σ	Mean	σ	p-value
Biology.SE	0.7191	0.1045	0.6287	0.1156	<0.001
MS.SE	0.7334	0.0993	0.6551	0.1100	<0.001
r/nutrition	0.7172	0.1029	0.6820	0.1100	<0.001

Nevertheless, the number of votes can provide a measure of the relevance of all PMIDs. On the other hand, it is also difficult to evaluate questions for which we only have a small number of PMIDs, regardless of the number of votes. For this reason, we experimented with limiting both the minimum number of votes (Table 4) and the number of PMIDs associated with the answers (Table 5) while performing a baseline evaluation.

TABLE 4. Effect of the minimum number of votes per answer on the MAP score and corpus size. The highest score of each subset is bolded.

	Biology		MS		Nutrition	
Min # Votes	# Qs	MAP	# Qs	MAP	# Qs	MAP
1	3714	0.0240	1332	0.0347	1963	0.0132
2	2896	**0.0260**	1114	0.0343	1550	0.0126
3	2202	0.0251	813	0.0377	1173	0.0136
4	1614	0.0240	577	0.0311	882	0.0145
5	1211	0.0259	408	0.0324	748	0.0147
6	943	0.0239	270	0.0269	616	0.0141
7	736	0.0222	191	0.0268	517	0.0143
8	593	0.0213	130	0.0331	459	0.0147
9	496	0.0214	98	0.0401	402	**0.0158**
10	415	0.0226	73	**0.0403**	354	0.0151

TABLE 5. Effect of the minimum number of PMIDs per question on the MAP score and corpus size. The highest score of each subset is bolded.

	Biology		MS		Nutrition	
Min # PMIDs	# Qs	MAP	# Qs	MAP	# Qs	MAP
1	3961	0.0237	1383	0.0359	2109	0.0133
2	1508	0.0313	688	0.0441	764	0.0144
3	661	0.0355	383	0.0453	395	0.0187
4	335	0.0448	216	0.0509	228	**0.0190**
5	163	0.0603	127	**0.0550**	153	0.0083
6	88	0.0829	78	0.0459	108	0.0075
7	51	0.1200	48	0.0504	77	0.0098
8	36	0.1462	34	0.0359	55	0.0135
9	21	0.1731	23	0.0299	43	0.0153
10	18	**0.2019**	19	0.0287	36	0.0171

As explained in Section IV-A, we retrieved the top 100 articles given by the query likelihood with Dirichlet term smoothing, which is the default scorer of Galago. Overall, we can see that increasing the minimum number of PMIDs can lead to higher MAP scores, although this results in a smaller corpus. The maximum MAP score is achieved when selecting only questions with at least 10 PMIDs for Biology. SE, 5 PMIDs for MS.SE, and 4 PMIDs for r/nutrition. We were also able to obtain higher MAP scores by filtering by the number of votes, on the MS.SE and r/nutrition datasets. Ideally, a balance should be found between the minimum number of votes and PMIDs and the size of the corpus.

TABLE 6. Effect of search method and query text on the MAP scores, using the BiQA corpus.

	Section	Biology.SE	MS.SE	r/nutrition
Pubmed	Title	0.0066	0.0103	0.0040
	Body	0.0052	0.0103	0.0031
	Answer	0.0787	0.0859	0.0641
BM25	Title	0.0267	0.0358	0.0139
	Body	0.0207	0.0298	0.0098
	Answer	0.1410	0.1571	0.0476
Dirichlet	Title	0.0237	0.0359	0.0133
	Body	0.0179	0.0285	0.0087
	Answer	0.1370	0.1744	0.0597

Finally, we studied two other variables on the baseline evaluation: the document retrieval method and the text used as a search query (title, body, or answer). Table 6 shows the results of this study, using each subset of the BiQA corpus. PubMed search obtained lower MAP scores, even though their method is an improvement over the BM25 scoring method. Their search engine is tuned for user queries, so one possible explanation is that questions easily answered by a PubMed search are less likely to be posted on Q&A forums. Furthermore, in every case, using the body text results in lower scores, and the best scores are achieved using the answer text. This is expected as the answer text introduces concepts that were missing from the question text and body.

VII. EVALUATION ON THE BioASQ6b DATASET

We applied the whole BiQA corpus to a deep learning QA system to study its impact on this type of system, merging the questions of the three subsets into a single corpus, without any restrictions of number of votes or PMIDs. We explored two scenarios: using BiQA in conjunction with the BioASQ6b training set and using just the BiQA corpus to train the model. The results of both experiments, as well as the results we obtained using only the BioASQ6b training data, are shown in Table 7, for each test batch.

TABLE 7. MAP scores obtained with the AUEB system using only the BioASQ6b training set, using both BioASQ6b and BiQA, and using only the BiQA corpus to train the model.

Test batch	BioASQ only	+BiQA	Δ	BiQA only	Δ
1	0.2221	0.2235	-0.0014	0.2125	0.0096
2	0.2267	0.2231	0.0035	0.2025	0.0241
3	0.2415	0.2436	-0.0021	0.2279	0.0136
4	0.1686	0.1712	-0.0026	0.1680	0.0006
5	0.1340	0.1355	-0.0015	0.1254	0.0086

We obtained higher MAP scores by adding the BiQA corpus to the BioASQ6b training set on 4 out of 5 batches, although these were only marginal improvements. We can see that our approach obtains MAP scores close to the ones obtained by the AUEB system using only the BioASQ training data. In fact, in 3 out of 5 batches, the difference is less than 0.01. These results show that the BiQA corpus could effectively be used as training data for biomedical QA, and expert curation could lead to better results.

VIII. DISCUSSION

Comparing Tables 4 and 5, we can see that, for the Biology.SE dataset, the number of PMIDs per answer is a more efficient way to obtain higher MAP scores, since for similar corpus sizes, the MAP score is higher using a PMID threshold. Even for higher threshold values, which results in small corpus sizes, the Biology.SE dataset obtains higher MAP scores. We did not show results for higher values given the small number of documents that remained. We also did not include answers with less than one vote since there were only 444 answers in this situation and did not change the results of this analysis. We expected MS.SE and r/nutrition to have the same behavior as Biology.SE on Table 5, in terms of how the MAP score increases with the minimum number of PMIDs. Since this was not the case, those two datasets proved to be more challenging and may require better query processing and natural language understanding in order to retrieve the correct documents. For example, multi-label question classification could also improve the scores of this type of task [22].

We compare the MAP scores obtained on the BiQA corpus with the ones obtained on the BioASQ6 challenge because it has a subtask that consists in document retrieval for QA (Task B Phase A). BioASQ is organized yearly, and every year there are new batches of questions. The participating teams are evaluated on each batch separately, obtaining different scores on each one. While the best score achieved on the 2019 edition was 0.2898 on test batch 3, on test batch 5, the best score was 0.1218. We also obtained a range of MAP scores on our corpus, in most cases lower than the BioASQ test batches. Furthermore, our corpus is closer to a real-world scenario since we are using user-submitted questions, which requires natural language understanding to correctly answer them. One possibility is that users had access to a search engine before posting a question and did not find an answer, therefore increasing the complexity of the questions. Novel systems tuned for this type of data can obtain better scores, especially considering that more data is made available each day as users post questions and answers on those forums.

We were able to train a deep learning model with our corpus that obtained MAP scores similar to a model trained on a corpus annotated by experts. One possible explanation for these results is that our corpus has more questions, and these questions were retrieved from various sources. Furthermore, adding the questions from our corpus to the model trained on the BioASQ6b training data led to slightly better MAP scores

on four test batches. These results highlight the importance of having more datasets for biomedical QA in order to improve the existing systems.

Our approach has some limitations when compared to expert annotations. Since Q&A forums are anonymous, there is no accountability of the answers, and personal biases may be more apparent than within a group of expert annotators. This approach is also dependent on how engaged the users are within the community: for example, r/nutrition has more users and a higher maximum number of votes than the other two, but it has fewer answers that we could link to PubMed articles, resulting in a lower average number of PMIDs. Another limitation is that there is no context for each citation used to answer. In some cases, an article is used as an example and not necessarily to answer the question. Our framework treats every article associated with an answer the same way because we do not know the rank of each article within an answer, so it cannot distinguish which ones are more relevant.

We select the biomedical domain given its complexity and because it is not usually the focus of QA systems. Although some biomedical QA systems have been proposed [23], these are limited in comparison to general ones. Biomedical-specific approaches are essential due to the relevance of this topic to the society. On the web, there are frequent health-related questions that are not easily answered but could be clarified through an efficient retrieval of the relevant literature. Questions from the BiQA corpus that were answered with several citations were related to vaccines, genetically modified food products, and meat consumption, for example, topics for which much information can be found in the scientific literature.

Our framework could potentially be applied to other communities where biomedical questions are made, for example, Quora, or social networks such as Twitter and Facebook. We selected PubMed as the reference library since it is widely used by the biomedical community and can be easily accessed. Additionally, information extraction and machine reading comprehension techniques could be applied to obtain the answer span of the selected documents [24], [25].

A. COMPARISON TO OTHER CORPORA

Table 8 provides a summary of existing resources related to User QA and Biomedical QA that can be compared with

TABLE 8. Comparison of BiQA to other Biomedical and User QA corpora. DR - Document Retrieval; QA - Question Answering.

Corpus	Domain	Task	Size	Method	Ref
BiQA	Biomedical	DR/QA	7k	Information retrieval	-
BioASQ	Biomedical	DR/QA	3k	Expert annotation	[3]
MEDIQA	Clinical	QA	208	Expert annotation	[12]
emrQA	Clinical	QA	455k	Template	[14]
PubMedQA	Biomedical	QA	211k	Template	[15]
Cong2008	General	Text classication	3M	Information retrieval	[8]
Shah2010	General	Ranking	120	Crowdsourcing	[9]
AmazonQA	Product reviews	QA	923k	Information retrieval	[11]

our approach. In comparison to other existing corpora for Biomedical QA, ours is the first to automatically obtain user questions and documents associated with them. Corpora used in community challenges such as BioASQ and MEDIQA requires questions and answers obtained from user experts, while PubMedQA and emrQA use synthetically generated questions. These methods, although useful for some purposes such as training models, lack the diversity that user-generated questions will provide, while expert datasets are limited in size. Our approach overcomes these two limitations. In comparison to other User QA corpora, our approach focused on the biomedical domain, and as such, fewer questions can be extracted from the web [8], [11]. However, our method takes advantage of the user-interaction aspect of web forum, therefore not requiring crowdsourcing techniques such as the one used by [9].
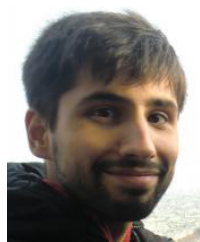
IX. CONCLUSION

This manuscript presents an effective open-source framework to generate corpora for QA document retrieval using Q&A forums, demonstrated in the form of the BiQA corpus. These forums are a useful resource for QA systems since the questions are real-world examples of user questions, and the answers are crowdsourced from multiple users. To the best of our knowledge, this is the first attempt at generating this type of dataset for QA. The task of document retrieval is a crucial step of a QA system, so these systems must optimize this step according to the user needs. We demonstrated the feasibility and performance of the framework and the BiQA corpus, namely their relevance for training and evaluating QA systems, particularly in the biomedical domain, for which resources are more scarce. The BiQA corpus is composed of 7,453 questions and 14,239 question-article pairs obtained from three different Q&A forums. The results obtained with our baseline methods indicate that higher MAP scores can be obtained on questions with a higher number of PMIDs, and more advanced methods as well as manual curation could help retrieve articles for all questions correctly. We release the dataset and the pipeline developed to extract it from the web to promote other works that tackle the challenges of QA systems that address real-world biomedical questions. As future work, we will use semantic representations to find similar questions, using word embeddings specific for community QA [26], [27]. These embeddings could be calculated using the corpora retrieved using our method.

REFERENCES

[1] L. Pang, Y. Lan, J. Guo, J. Xu, J. Xu, and X. Cheng, "DeepRank: A new deep architecture for relevance ranking in information retrieval," in *Proc. ACM Conf. Inf. Knowl. Manage.*, New York, NY, USA, Nov. 2017, pp. 257–266.

[2] J. Guo, Y. Fan, Q. Ai, and W. B. Croft, "A deep relevance matching model for ad-hoc retrieval," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2016, pp. 55–64.

[3] G. Tsatsaronis *et al.*, "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," *BMC Bioinf.*, vol. 16, no. 1, p. 138, Dec. 2015.

[4] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, and K. Toutanova, "Natural questions: A benchmark for question answering research," *Trans. Assoc. Comput. Linguistics*, vol. 7, no. 29, pp. 453–466, 2019.

[5] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for SQuAD," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 2. Melbourne, VIC, Australia: Association Computational Linguistics, Jul. 2018, pp. 784–789.

[6] E. Agichtein, D. Carmel, D. Pelleg, Y. Pinter, and D. Harman, "Overview of the TREC 2015 LiveQA track," in *Proc. 24th Text Retr. Conf. (TREC)*, Gaithersburg, MD, USA, Nov. 2015, pp. 1–9.

[7] A. B. Abacha, E. Agichtein, Y. Pinter, and D. Demner-Fushman, "Overview of the medical question answering task at TREC 2017 LiveQA," in *Proc. 26th Text Retr. Conf. (TREC)*, Gaithersburg, MD, USA, Nov. 2017, pp. 1–12.

[8] G. Cong, L. Wang, C.-Y. Lin, Y.-I. Song, and Y. Sun, "Finding question-answer pairs from online forums," in *Proc. 31st Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 2008, pp. 467–474.

[9] C. Shah and J. Pomerantz, "Evaluating and predicting answer quality in community QA," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 2010, pp. 411–418.

[10] D. H. Dalip, M. A. Gonçalves, M. Cristo, and P. Calado, "Exploiting user feedback to learn to rank answers in q&a forums: A case study with stack overflow," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, New York, NY, USA, 2013, pp. 543–552.

[11] M. Gupta, N. Kulkarni, R. Chanda, A. Rayasam, and Z. C. Lipton, "AmazonQA: A review-based question answering task," in *Proc. 28th Int. Joint Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, Aug. 2019, pp. 4996–5002.

[12] A. B. Abacha, C. Shivade, and D. Demner-Fushman, "Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering," in *Proc. 18th BioNLP Workshop Shared Task*. Florence, Italy: Association Computational Linguistics, Aug. 2019, pp. 370–379.

[13] A. B. Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," 2019, *arXiv:1901.08079*. [Online]. Available: http://arxiv.org/abs/1901.08079

[14] A. Pampari, P. Raghavan, J. Liang, and J. Peng, "EmrQA: A large corpus for question answering on electronic medical records," in *Proc. Conf. Empirical Methods Natural Lang. Process.* Brussels, Belgium: Association Computational Linguistics, 2018, pp. 2357–2368.

[15] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "PubMedQA: A dataset for biomedical research question answering," 2019, *arXiv:1909.06146*. [Online]. Available: http://arxiv.org/abs/1909.06146

[16] W. B. Croft, D. Metzler, and T. Strohman, *Search Engines: Information Retrieval in Practice*, vol. 520. Reading, MA, USA: Addison-Wesley, 2010.

[17] S. E. Robertson, S. Walker, M. Beaulieu, and P. Willett, "Okapi at TREC-7: Automatic ad hoc, filtering, VLC and interactive track," *Nist Special Publication SP*, vol. 500, pp. 253–264, Nov. 1999.

[18] N. Fiorini, K. Canese, G. Starchenko, E. Kireev, W. Kim, V. Miller, M. Osipov, M. Kholodov, R. Ismagilov, S. Mohan, J. Ostell, and Z. Lu, "Best match: New relevance search for PubMed," *PLOS Biol.*, vol. 16, no. 8, Aug. 2018, Art. no. e2005343.

[19] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, and I. Kakadiaris, "Results of the fifth edition of the BioASQ challenge," in *Proc. BioNLP*, 2017, pp. 48–57.

[20] K. Hui, A. Yates, K. Berberich, and G. de Melo, "PACRR: A position-aware neural IR model for relevance matching," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 1–10.

[21] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[22] M. Wasim, W. Mahmood, M. N. Asim, and M. U. Khan, "Multi-label question classification for factoid and list type questions in biomedical question answering," *IEEE Access*, vol. 7, pp. 3882–3896, 2019.

[23] F. Galkó and C. Eickhoff, "Biomedical question answering via weighted neural network passage retrieval," in *Proc. Eur. Conf. Inf. Retr.* Cham, Switzerland: Springer, 2018, pp. 523–528.

[24] L. Qiu, D. Ru, Q. Long, W. Zhang, and Y. Yu, "Qa4ie:A question answering based system for document-level general information extraction," *IEEE Access*, vol. 8, pp. 29677–29689, 2020.

[25] D. Park, Y. Choi, D. Kim, M. Yu, S. Kim, and J. Kang, "Can machines learn to comprehend scientific literature?" *IEEE Access*, vol. 7, pp. 16246–16256, 2019.

[26] G. Zhou, T. He, J. Zhao, and P. Hu, "Learning continuous word embedding with metadata for question retrieval in community question answering," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1. Beijing, China: Association Computational Linguistics, Jul. 2015, pp. 250–259.

[27] G. Zhou, Y. Zhou, T. He, and W. Wu, "Learning semantic representation with neural networks for community question answering retrieval," *Knowl.-Based Syst.*, vol. 93, pp. 75–83, Feb. 2016.

DIANA SOUSA received the B.S. degree in biochemistry and the M.Sc. degree in bioinformatics and computational biology from the Faculty of Sciences, University of Lisbon, Lisbon, Portugal, in 2017 and 2019, respectively, where she is currently pursuing the Ph.D. degree in informatics.

Since 2016, she has been a Researcher with the LASIGE, Department of Informatics, Faculty of Sciences, University of Lisbon. Her research interest includes the development of text mining systems that aim at extracting biomedical information from biomedical literature, mainly focusing on relations between biomedical entities and on using deep learning approaches that resort to external sources of knowledge.

ANDRE LAMURIAS received the B.S. degree in biochemistry and the M.Sc. degree in bioinformatics and computational biology from the Faculty of Sciences, University of Lisbon, Lisbon, Portugal, in 2013 and 2015, respectively, and the Ph.D. degree in systems biology from the Faculty of Sciences, University of Lisbon.

He is currently a Senior Researcher with Priberam Labs. From 2013 to 2020, he was a Researcher with the LASIGE, Department of Informatics, Faculty of Sciences, University of Lisbon. He has participated on several text mining competitions, such as BioCreative, SemEval, and BioASQ. His research interests include biomedical text mining, including tasks such as named entity recognition, relation extraction, and question answering, using machine learning approaches and ontologies.

FRANCISCO M. COUTO received the degree and the master's degree in informatics and computer engineering from IST, in 2000 and 2001, respectively, and the Ph.D. degree in informatics, specialization in bioinformatics, from the Universidade de Lisboa, in 2006.

He was with the faculty at IST, from 1998 to 2001. Since 2001, he has been with the Faculdade de Ciências, Universidade de Lisboa (FCUL). He was an Invited Researcher with EBI, AFMB-CNRS, and BioAlma during his Ph.D. studies. He is currently an Associate Professor with habilitation, a member of coordination board of the master in bioinformatics and computational biology, and a member of LASIGE, where he founded the HBI-Health and Biomedical Informatics research line. He received the Young Engineer Innovation Prize 2004 from the Portuguese Engineers Guild and the Honorable Mention in 2017 and the prize in 2018 of the ULisboa/Caixa Geral de Depósitos (CGD) Scientific Prizes.

• • •