

MDS_UNCC Question Answering System for Biomedical Data with Preliminary Error Analysis

Seethalakshmi Gopalakrishnan¹, Swathi Padithala², Hilmi Demirhan¹ and Wlodek Zadrozny^{1,2}

¹College of Computing and Informatics, University of North Carolina at Charlotte, United States

²School of Data Science, University of North Carolina at Charlotte, United States

Abstract

In this paper, we are describing the submission that we made for the 9th year BioASQ competition. The BioASQ challenge aims to promote methodologies and systems for large-scale biomedical semantic indexing and question answering. There are shared tasks that are the benchmark datasets through the BioASQ website yearly. The dataset represents the information needed from biomedical experts. This paper has worked on the 9th task with the BioASQ BioBERT model based on the Bidirectional Encoder Representations from the Transformers (BERT) model. We have fine-tuned the BioASQ BioBERT model and submitted our results both with training and without training. The results show that fine-tuning with training the model gives better results. For the Batch5 factoid submission, we have got an MRR of 0.52, which is higher than the original version of BioASQ BioBERT. For Yes/No, we got the F1 score of 0.81, and for the list, the F1 was 0.26.

We also present preliminary results of our error analysis, where we hypothesize about the causes of some errors, and run simple experiments to confirm or disprove them. For example, we see that the presence of the natural language modalities – which are quite common in questions, answers and snippets – influences the accuracy.

Keywords

BioASQ, Question Answering, Error Analysis, Factoid, Yes/No, List, Modal Verbs

1. Introduction

Question Answering (QA) focuses the Information Retrieval (IR) system on the actual information need of the user. Performance on a QA task reflects domain knowledge and natural language understanding ability of a system.

BioASQ, Biomedical Semantic Question Answering Challenge, is one of the famous competitions that ask practitioners to develop a system for unstructured biomedical text and decision support from PubMed articles. The system is supposed to automatically respond to biomedical questions with relevant concepts, articles, snippets, exact answers and summaries. BioASQ challenge is a competition of document classification, information retrieval and question answering. The participants are expected to achieve the goal of submitting ideal and user-understandable answers to the four types of natural language questions by combining the information from

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

✉ sgopala4@uncc.edu (S. Gopalakrishnan); spaditha@uncc.edu (S. Padithala); hdemirha@uncc.edu (H. Demirhan); wzadrozni@uncc.edu (W. Zadrozny)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

biomedical articles and ontologies. The four types of questions from benchmark datasets are Yes/No questions, Factoid questions, List questions, and Summary questions. One example of Yes/No questions is that "Do CpG islands colocalize with transcription start sites?". One example for a Factoid question is that "Which virus is best known as the cause of infectious mononucleosis?". One example for a List question is "Which are the Raf kinase inhibitors?". Also one example for a Summary question is that "What is the treatment of infectious mononucleosis?". There are two phases in the challenge: Phase A and Phase B. During phase A, participating systems have to reply with related concepts which are from designated terminologies and ontologies, related articles in English, related snippets and RDF triples. During phase B, participating systems need to have responses with exact and ideal answers in English.

This project builds on the 2019 BioASQ experiments reported in [1]. In the project, we are using Task 9b data (the 9th year BioASQ competition shared task). In Task 9a, participants are asked to classify new abstracts written in English, as they became available online. The classes came from the MESH hierarchy, i.e., the subject headings that are currently used to manually index the abstracts. As new manual annotations became available, they were used to evaluate the classification performance of participating systems (that classified articles before they were manually annotated), using standard information retrieval (IR) measures (e.g., precision, recall, accuracy), as well as hierarchical variants of these measures. The models we are using are BioBERT which is based on Bidirectional Encoder Representations from Transformers (BERT) model [2] which is pre-trained on Wikipedia articles. The BioBERT model was pre-trained with biomedical text using the PubMed and PMC articles [3]. As the participants, we had to annotate input natural language questions with biomedical concepts, and retrieve relevant documents, snippets and triples (Phase A). Eventually, we need to find and report the answers to the questions (Phase B), given as additional input the golden responses of the Phase A.

Our contributions in this paper are as follows:

- We describe our systems, which placed 4th in Task 9b Phase B. In data preparation, we created a new way of converting the BioASQ format to BioBERT format, and we make our code available <https://github.com/seetagopal/BioASQ-2021.git>
- We performed error analysis (on training data) of the Yes/No questions, and hypothesized about causes of some errors. These, e.g. include the presence of modal verbs/auxiliaries such as 'may'.
- We performed a statistical analysis of Wh-questions and of the modal verbs/auxiliaries.

2. BioASQ Related Work and the Competition Data

2.1. BERT and BioBERT

BERT stands for "Bidirectional Encoder Representations from Transformers" is a contextual word embedding model which was developed in 2018 by Google [2]. It is a contextualized word representation model that is pre-trained using bidirectional transformers [4]. The model takes a sentence as an input and outputs a contextual embedding of the word. Currently Google Search Engine uses BERT model for over 70 languages [5]. In addition to search, BERT can be used for additional tasks such as question answering and language inference. BERT's pre-trained

deep bidirectional representations from unlabeled text can be modified with an additional output layer for these different tasks. Bidirectional representations are crucial in biomedical text mining to represent relationships in a biomedical corpus [6]. In this work, we are using BERT for question answering tasks. Question and paragraph (context) are given as an input to the model.

In this work, we used BioBERT models. BioBERT (Bidirectional Encoder Representations from Transformers for Biomedical Text Mining) is a domain-specific language representation model pre-trained on large-scale biomedical corpora developed by Lee et al. [7]. BioBERT and BERT have very similar architecture. Since it is pre-trained on biomedical corpora, it achieves a better performance than BERT in biomedical text mining tasks.

Lee et al. [7] have fine-tuned BioBERT for question answering by using the same BERT architecture used for SQuAD [8]. They used the BioASQ factoid datasets for fine-tuning as factoid datasets have similarity to that of SQuAD. Some of the BioASQ factoid questions were unanswerable as exact answers were not present in the given texts. The unanswerable questions were removed from the training sets. They used the same pre-training process of Wiese et al. [9] as they have also used SQuAD. They have used strict accuracy, lenient accuracy and mean reciprocal rank as evaluation metrics. Yoon et al. [10] have also used BioBERT for answering biomedical questions including factoid, list, and yes/no type questions. Jin et al. [11] have reviewed the Biomedical Question Answering approaches by classifying them in 6 major methodologies namely open-domain, knowledge base, information retrieval, machine reading comprehension, question entailment and visual QA.

2.2. Prior BioASQ work

This year ninth edition of the BioASQ Challenge is being held. The previous year, 8th BioASQ challenge can be summarized as follows [12]. There were three tasks last year: Task 8a, task 8b, task MESINESP. Task 8a was a large-scale biomedical semantic indexing task. Task 8b was a biomedical question answering task. Task MESINESP was a new task on medical semantic indexing in Spanish (task MESINESP).

Some of the task 8b submissions were as follows. Kazaryan et al. [13] have participated as ITMO team. They used BERT fine-tuned on SQuAD [8]. They also used a model based on BioMed-RoBERTa [14] to improve the produced answers. Ozyurt et al. [15] used Electra [16] and BioBERT [7] on SQuAD and BioASQ datasets combined. Pappas et al. [17] experimented with a SciBERT-based model for exact answer extraction [18] modelled for cloze-style biomedical machine reading comprehension [19] (MRC).

2.3. BioASQ Data

For the 9th BioASQ tasks, training dataset for Task 9a is about semantic indexing. Task 9a data contains MeSH terms MEDLINE curators annotated in the biomedical articles from PubMed. Different years version means different MeSH terms used in the articles from PubMed besides different sizes. Task 9b is about Question Answering; and as part of the task in the data we get concepts, articles, snippets, RDF triples, “exact” answers and “ideal” answers in JSON format.

Task 9a, named “Large-scale online biomedical semantic indexing”, has 15,559,157 articles.

In each article, there are an average 12.68 MeSH terms annotated. 29,369 MeSH covered total. The dataset is 7.9 GB zipped and 25.6 GB unzipped. Task 9b, named “Introductory biomedical semantic QA,” uses benchmark datasets containing development and test questions, in English, along with gold standard (reference) answers. The benchmark datasets are being constructed by a team of biomedical experts from around Europe.

Below is the number of questions in each category in the 9b training data. Yes/No: 1033, Factoid: 1092, List: 719, Summary: 899. The number of questions in each category in Task 9b Batch5 test data: Yes/No: 19, Factoid: 36, List: 18, Summary: 27.

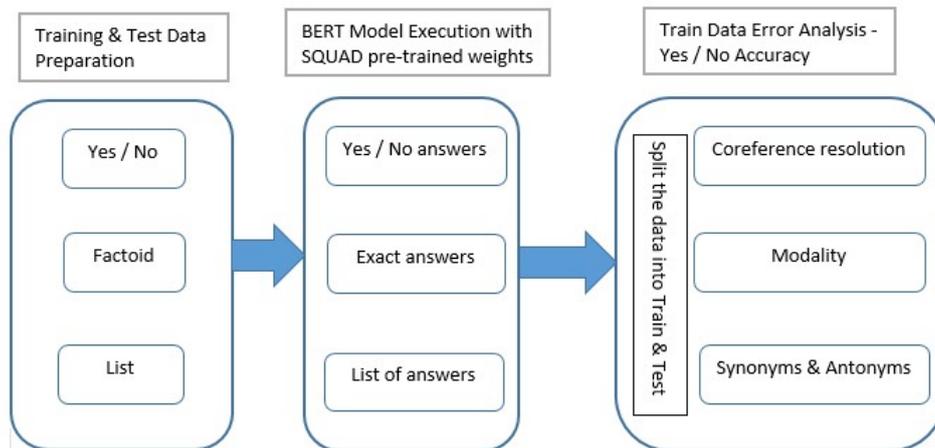


Figure 1: Overview of the model. First step is to prepare the data (Section 3) into BioBERT format which will then be used by the BioASQ-BioBERT model for training and prediction. The last step is the error analysis which is explained in Section 5. The major finding of our error analysis of Yes/No questions was that understanding coreference resolution, modality, synonyms and antonyms may help in improving the model accuracy

3. Data Preparation and Experiments

3.1. Data Preparation

The main purpose of our data preparation is to convert the BioASQ data into the format which is accepted by the BioBERT. For the factoid questions we are creating a dictionary in which the question id, question, answer, context will be added. BioBERT will expect the start index of the answers during the training. So, one of the main tasks of this data preparation is to find the start of the answers. In order to find the start index, we are performing the set operation on the ideal and exact answers to find the unique answers for that question. Once we have the answers the next step is to find out what is the index of those answer in the given snippet. If the answer is present in the snippet, then that will be set as the start index which we are doing by setting flags. Apart from creating a dictionary to append all the information given above we are also adding three more numbers to the given id. This is necessary because the given id will

be of length 24 but the evaluation script in BioBERT to convert the BioBERT predictions into BioASQ format will expect the id of length 28. The same procedure is followed for the list type questions. For Yes/No same procedure is followed except for finding the answer start index.

3.2. Experiments: Finetuning BioBERT

We are using the pretrained weights provided by the BioASQ BioBERT model [20, 3]. The first step for training this model is to convert the given data into the BioBERT format which includes the answer start index. The process of how we are doing this is explained in the data preparation (Section 3). In the BioASQ BioBERT model, they have released the pretrained weights for the factoid, list and Yes/No questions separately. These weights are pretrained on SQUAD dataset on top of the BioBERT model. We are using those weights to train the BioBERT model for the BioASQ training 9b data. Once the training is done, we are using the weights obtained as a result of training for the prediction. We are following this procedure for factoid, list and Yes/No question answer. Once we got the predictions we ran the script provided in the BioASQ BioBERT model to convert the predictions into the BioASQ format. The overview of the model is given in Figure 1.

4. Results

We have submitted our results to the 9th Batch of BioASQ competition by finetuning the BioASQ BioBERT model on the BioASQ Task9B – Phase B dataset. We have submitted our predictions in the Batch 5 of the BioASQ competition and we are placed 4th in the leader board. Our system name is MDS_UNCC. The results that we got for our Test Batch5 submission is given in Table 1.

System	Yes/No			Accuracy	Factoid		MRR
	Macro F1	F1 Yes	F1 No		Strict Acc.	Lenient Acc.	
KU-DMIS-2	0.8246	0.88	0.7692	0.8421	0.3889	0.6111	0.4722
KU-DMIS-3	0.8246	0.88	0.7692	0.8421	0.4167	0.5556	0.4745
KU-DMIS-5	0.8246	0.83	0.7692	0.8421	0.3889	0.5833	0.4583
MDS_UNCC	0.7841	0.8182	0.75	0.7895	0.4167	0.6944	0.5204

Table 1

BioASQ competition results for the Batch5 submission. KU-DMIS is placed first in the leaderboard. MDS_UNCC is our submission which is placed 4th in the competition. For the List questions we got an Mean Precision of 0.3259, recall of 0.2963 and F-measure of 0.2678 (which is about 12% below the best system).

5. Error Analysis

In this section we discuss some aspects of our error analysis. We start with Batch 3, but we mostly focus on Batch 5. We perform both error analysis and data analysis in this section. In particular we discuss the presence of the modalities in the data, types of questions, and the accuracy on these types.

The Yes/No test data prediction file from the Batch3 has identified 87% Yes Accuracy and 54% No Accuracy; that is, most of the Yes were identified correctly but almost half of No answers were identified wrong. Upon doing the further analysis, we see that the probability of No questions was underestimated. We hypothesize that corpus expansion [21, 22] would help in improving the accuracy, but we have not yet tested this hypothesis.

For Factoid questions, we observed a pattern of lower probabilities for the best and longest match, and the higher probability for initial sentences was higher. For List questions, the probability was again smaller for the long text and the probability for initial sentences higher; moreover, the best matching words were often left out. Thus, we need to find a solution to include the best match words and for matching the longest sentence.

	Questions	Answers	Snippets
Can	58	319	140
Could	0	55	45
May	6	201	107
Might	0	22	21
Should	5	30	16
Will	0	24	33
Would	4	12	9
Must	0	11	4
Possible	8	19	23
Percent of modals	7%	59%	33%

Table 2

Number of modals present in the BioASQ training9b dataset. In this table, column indicates the modal verb and the row indicates the count of the modal verb for questions, answers and snippets in the dataset. The last row indicates the percentage of modals present in each of the questions, answers and snippets. We hypothesize that a better understanding of the role of modals can improve the accuracy of question answering.

In our BioASQ Batch5 submission we got an accuracy of 78% for Yes/No questions. In order to find out where the machine fails with the Yes/No questions, we are presenting a more detailed error analysis for the Yes/No questions in this section. Since the golden data for the task9b is not yet published, we have divided the training data in train and test data using the scikit learn library with train data 80% and test data 20%. We had 835 Yes/No questions in the train data and 207 questions as test data. We ran BioASQ-BioBERT on this train data and made some predictions. Out of 207 Yes/No questions, we got 184 questions answered correctly and 23 questions answered incorrectly. For further analysis, we have checked these 23 questions manually along with the snippets. Our analysis show that in many of the questions, the machine cannot understand the synonyms and the antonyms. Also, for few complex questions coreference resolution is needed. For example:

Example 1:

Question: Does dronedarone affect T3 and T4 levels?

Actual answer: No

Predicted answer: Yes

Snippet: Amiodarone resulted in increased T4, T4/T3 and rT3, whereas dronedarone did not alter the thyroid hormone profile in normal animals.

In the above example the expected answer is No but the predicted answer was Yes. In the snippet, the first part talks about Amiodarone, whereas the second part gives the answer for the above question and says that the dronedarone did not alter the thyroid hormone profile. This question can be answered correctly only if the machine can understand that T3, T4 refers to the thyroid hormone.

In order to check whether our hypothesis is correct, we have changed the above question into "Does dronedarone alter the thyroid hormone level?" and ran BioASQ-BioBERT. Changing the question predicted this answer correctly. We can infer that the proper dealing of coreference resolution is needed.

	Training	Batch5 Test
Starts with "Wh"	54%	64%
Starts with "What is"	23%	29%
in the form "What is X?"	5%	6%
in the form "What is .. of/using .. ?"	12%	16%
Starts with "What .." but not "What is/are"	4%	4%
Which	22%	25%
What are	3%	3%
Where	1%	1%
When	0.3%	1%
Complement of "Wh"	46%	36%

Table 3

Percentage allocation of most frequent question types in training for Task 9b and Batch5 test dataset. Here the Complement of "Wh" indicates the questions which do not start with "Wh." (The percentages are rounded-off; and some types of "what is" questions are not counted here).

One other example which explains the understanding of modality is important is given below:

Example 2:

Question: Is cardiac magnetic resonance imaging indicated in the pre-participation screening of athletes?

Actual answer: No

Predicted answer: Yes

Snippet: As modern imaging further enhances our understanding of the spectrum of athlete's heart, its role may expand from the assessment of athletes with suspected disease to being part of comprehensive pre-participation screening in apparently healthy athletes.

In the above example the expected answer was No and it was predicted as Yes. Looking into the snippet, we can understand that modern imaging may be used for pre-participation but this is not necessary. If the machine can understand the modal verb "may" then this question can be answered correctly.

To further analyze the role of modals in BioASQ data we have collected the count of modals in

	Yes/No	List	Factoid
Starts with "Wh"	N/A	37%	30%
starts with "Which"	N/A	44%	43%
starts with "List"	N/A	53%	N/A
starts with "What is"	N/A	0%	32%
starts with "What are"	N/A	19%	0%
starts with "Where"	N/A	0%	33%
other type of "What questions"	N/A	12.5%	16%
complement of wh and list	N/A	42%	N/A
starts with "What is X?"	N/A	N/A	0%
starts with what is of/using ..?	N/A	N/A	13%
Complement of wh questions	89%	N/A	24%
Overall accuracy	89%	43%	31%

Table 4

Accuracy of different type of questions and the over all accuracy. There are no "wh" questions in the Yes/No type. There are no "What is X" and "What is of/using.." questions in the list, Yes/No type. In most of the categories with 0% accuracy the number of questions is very less with a minimum count of 1 and a maximum count of 10 questions.

the training data which is given in Table 2. This count shows that there are a number of modal verbs present in the training data. If the machine can understand the modals during the train, this may help in improving the accuracy. Also, from Table 2 we can see that the percentage of modals in the answers is higher than that of the questions. A better understanding of the relationship between the modals in the questions and answers may help improve the accuracy further.

In order to investigate our claim about the role of the modalities, we have changed the above question in Example 2 into "May cardiac magnetic resonance imaging be indicated in the pre-participation screening of athletes?". After changing the question the model has predicted the answer correctly which suggests the importance of understanding the modality. Obviously, we need to do more work on a larger set of examples to prove or disprove this hypothesis.

Example 3: Question: Does the BRAFV600E mutation have an effect on clinical response to radioiodine therapy?

Actual answer: Yes

Predicted answer: No

Snippet: Preclinical studies showed that BRAF mutation significantly reduced radioiodine uptake and decreased the sensitivity to radioactive iodine (RAI) therapy.

In the above example, the question asks about the effect of BRAFV600E. In the snippet, it shows that BRAF significantly reduced radioiodine which is an effect of BRAF. This question can be answered correctly if the machine can understand that 'effect' and 'reduced' are synonyms here. To investigate this claim we have changed the question to "Does the BRAFV600E mutation have reduced on clinical response to radioiodine therapy?". However, the system still predicts the answer wrongly. We can infer from this result that a better understanding about how the machine interprets these synonyms and antonyms is necessary.

Another aspect of the analysis is shown in Table 3. Namely we see that the distribution of types of questions is different in the training and test data.

The accuracy results for the three types of questions (Yes/No, List, and Factoid) are computed on the test data we obtained by splitting the original training data, as described earlier. We calculate the accuracy by finding the exact match between the correct answers in the training data and the predictions we got for the List and Factoid questions. Out of the obtained list of predictions, even if one of the predictions is correct, we consider that question to be answered correctly. The results of the accuracy we got for the different types of questions are given in the Table 4.

6. Summary and Conclusions

In this article, we described our contribution to the 9th BioASQ competition. We showed the process of retraining the BioASQ-BioBERT model used in our experiments leads to better results. However, to improve our results further, we need better understanding of the models, as well as of the data; that is, questions, answers, and snippets. Regarding the models, in accordance with common knowledge, we believe that both larger data sets for training and corpus expansion for background knowledge [21, 22] should lead to improved results. Based on a few examples, we also predict that deeper language understanding, e.g. coreference and synonym resolution could have an impact on accuracy. As for the data, we showed that modalities are potentially important, and are present surprisingly frequently in both questions and answers. Again, this suggests that we should experiment with deeper NLP methods.

Acknowledgments

The authors acknowledge the help of David Ruddell in data preparation.

References

- [1] S. K. Telukuntla, A. Kapri, W. Zadrozny, UNCC biomedical semantic question answering systems. BioASQ: Task-7B, Phase-B, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 695–710.
- [2] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: NAACL-HLT, 2018.
- [3] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz682.
- [4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, arXiv preprint arXiv:1706.03762 (2017).
- [5] Wikipedia contributors, Bert (language model), 2021. URL: [https://en.wikipedia.org/wiki/BERT_\(language_model\)](https://en.wikipedia.org/wiki/BERT_(language_model)), [Online; accessed 11-May-2021].

- [6] M. Krallinger, O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez, et al., Overview of the biocreative vi chemical-protein interaction track, in: Proceedings of the sixth BioCreative challenge evaluation workshop, volume 1, 2017, pp. 141–146.
- [7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2020) 1234–1240.
- [8] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, *arXiv preprint arXiv:1606.05250* (2016).
- [9] G. Wiese, D. Weissenborn, M. Neves, Neural domain adaptation for biomedical question answering, *arXiv preprint arXiv:1706.03610* (2017).
- [10] W. Yoon, J. Lee, D. Kim, M. Jeong, J. Kang, Pre-trained language model for biomedical question answering, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 727–740.
- [11] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, C. Tan, M. Chen, S. Huang, X. Liu, S. Yu, Biomedical question answering: A comprehensive review, *arXiv preprint arXiv:2102.05281* (2021).
- [12] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, Overview of bioasq 8a and 8b: Results of the eighth edition of the bioasq tasks a and b (2020).
- [13] A. Kazaryan, U. Sazanovich, V. Belyaev, Transformer-based open domain biomedical question answering at bioasq8 challenge (????).
- [14] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, *arXiv preprint arXiv:2004.10964* (2020).
- [15] I. B. Ozyurt, A. Bandrowski, J. S. Grethe, Bio-answerfinder: a system to find answers to questions from biomedical texts, *Database* 2020 (2020).
- [16] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, *arXiv preprint arXiv:2003.10555* (2020).
- [17] D. Pappas, P. Stavropoulos, I. Androutsopoulos, Aueb-nlp at bioasq 8: Biomedical document and snippet retrieval, CLEF (cit. on p. 41) (2020).
- [18] D. Chen, A. Fisch, J. Weston, A. Bordes, Reading wikipedia to answer open-domain questions, *arXiv preprint arXiv:1704.00051* (2017).
- [19] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, *arXiv preprint arXiv:1903.10676* (2019).
- [20] W. Yoon, J. Lee, D. Kim, M. Jeong, J. Kang, Pre-trained language model for biomedical question answering, in: P. Cellier, K. Driessens (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2020, pp. 727–740.
- [21] N. Schlaefel, J. Chu-Carroll, E. Nyberg, J. Fan, W. Zadrozny, D. Ferrucci, Statistical source expansion for question answering, in: Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 345–354.
- [22] J. Chu-Carroll, J. J. Fan, N. M. Schlaefel, W. W. Zadrozny, Source expansion for information retrieval and information extraction, 2014. US Patent 8,892,550.