LaRSA at BioASQ 10b: classical and novel approaches for biomedical document retrieval and question answering

Zakaria KADDARI[a], Toumi BOUCHENTOUF[a]

[a] *LaRSA laboratory, AIRES team, National School of Applied Sciences, Université Mohammed Premier, Oujda, Morocco*

Abstract

In this paper, we describe our participation in the 10th edition of the BioASQ challenge. We participated in both phases (A and B) of task B, with one system for each phase. In phase A, we used ElasticSearch with BM25 as a retriever, Roberta-base fine-tuned on SQuAD as a reader, along with a cross-encoder based reranker trained on MS Marco Passage Ranking task. In phase B, we submitted results for both exact, and ideal answers. For factoid and list answer types, we used BioBERT-large fine-tuned on SQuAD with our novel attention enriching mechanism. For yes/no answer type we used BioBERT-large fine-tuned on the BoolQA and PubMed QA datasets. For ideal answers, we used BART-large fine-tuned on the CNN, and the ebmsum datasets. Both of our systems yielded promising results, especially in phase B.

Keywords [1]

BioASQ, BQA, BioBERT, BM25, BART

1. Introduction

The task of biomedical question answering (BQA) [1] has seen considerable attention from the question answering (QA) research community in recent years. Especial after the COVID-19 pandemic, where biomedical researchers were in a race against time to analyze and make sense of the huge number of scientific studies on COVID-19 and the related coronaviruses, especially in the early days of the pandemic. Numerous, end-to-end QA systems were developed to address this challenge [2, 3, 4, 5, 6]. Some were built directly as a response to the pandemic, focusing mainly on COVID-19 scientific literature.

The BioASQ [7] challenge is an annual competition focusing on large-scale biomedical semantic indexing and QA. It is the only competition that addresses the overall components of end-to-end biomedical QA systems. From the initial document indexing and retrieval, to the extraction or the generation of the final answer. Also taking into account the four common question types, factoid, list, yes/no, and summary questions. This year marks the 10th edition of the competition, which indicates the success and impact it has on the BQA and other related tasks. This year also marks our team's (LaRSA) first participation in the competition.

This year, the competition has four distinct tasks, task Synergy on biomedical semantic QA for COVID-19, task A on large-scale online biomedical semantic indexing, task B on biomedical semantic QA which is further divided into phase A for document and snippets retrieval, and phase B for QA and summarization. The fourth task is called DisTEMIST which involves disease text mining and indexing.

In the rest of the paper, we describe our participation in this year's BioASQ challenge, where we participated in both phases A and B of task B. In phase B we submitted results for both exact (factoid, list, and yes/no) and ideal answers.

In the next section, we describe our systems for phase A and B of task B. For phase A, we present the detailed architecture of our system composed of three main components, a retriever, a ranker, and a

[1] CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

EMAIL: z.kaddari@ump.ac.ma (A. 1); t.bouchentouf@ump.ac.ma (A. 2)
ORCID: 0000-0003-4034-5612 (A. 1); 0000-0002-2689-8678 (A. 2)

reader. For phase B, we used one system with different models for factoid/list questions, yes/no, and summary/ideal answers. Therefore, we describe the model we used for each question type. In section three, we present and discuss our results for each phase and each question type. Finally we finish with a conclusion.

2. System description

2.1. Phase A

In this phase, a set of biomedical questions are provided for each batch of the competition. For each question, each participating system is required to return a list of at most 10 relevant biomedical articles from the designated article repositories, and a list of at most 10 relevant text snippets from the returned articles.

For this phase, we used ElasticSearch with BM25 as a retriever, Roberta-base [8] fine-tuned on SQuAD [9] as a reader, along with the ms-marco-MiniLM [2] cross-encoder based reranker trained on MS Marco Passage Ranking task [3]. We implemented our system for this phase using the haystack [4] framework. We set the number of documents to be returned by the retriever, the reader, and the ranker, to 200, 50, and 10 respectively.

Figure 1, shows the overall architecture of our system for phase A.

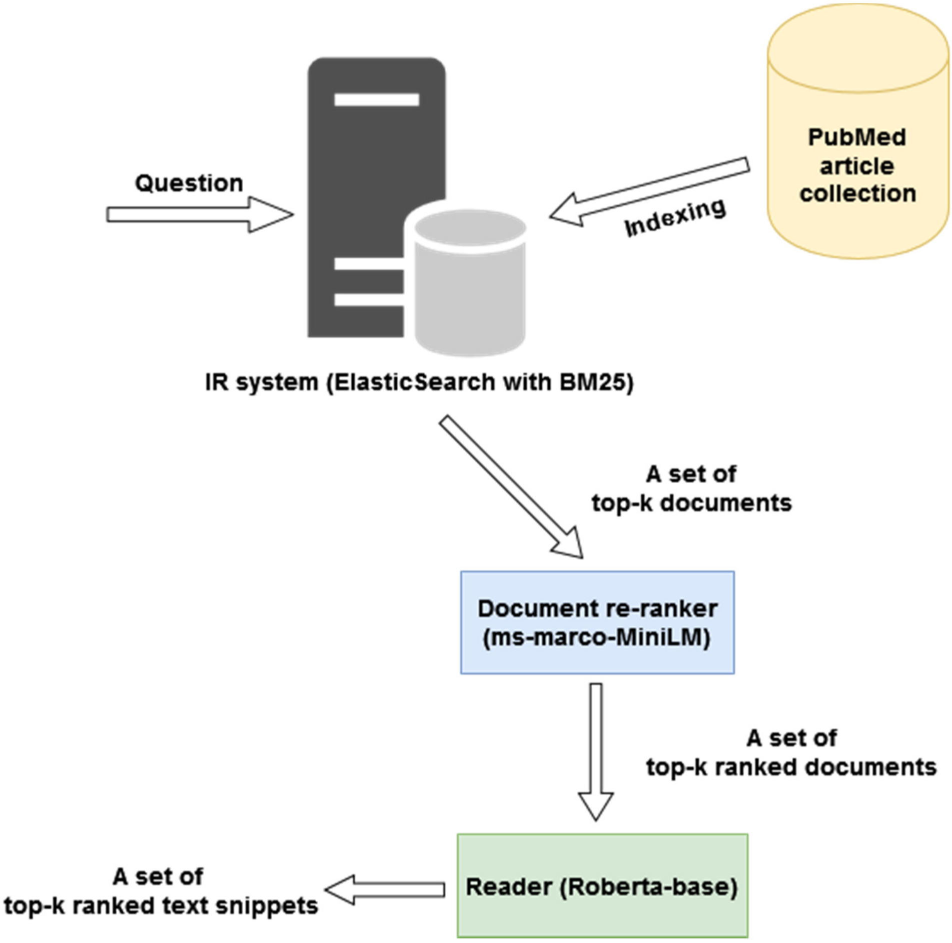
Question
PubMed article collection
Indexing
IR system (ElasticSearch with BM25)
A set of top-k documents
Document re-ranker (ms-marco-MiniLM)
A set of top-k ranked documents
Reader (Roberta-base)
A set of top-k ranked text snippets

Figure 1: The overall architecture of our system for phase A

[2] https://huggingface.co/cross-encoder/ms-marco-MiniLM-L-12-v2
[3] https://microsoft.github.io/MSMARCO-Passage-Ranking
[4] https://haystack.deepset.ai

2.2. Phase B

In this phase, the participants are provided with the same questions as in phase A along with the correct lists of articles and text snippets. For each question, a participating system must return exact answers for factoid, list, and yes/no questions along with an optional ideal answer which takes the form of an extracted or generated paragraph-sized full answer. For summary questions, only the ideal answer is required. Below, is an overview of the approaches we took for each question type.

2.2.1. Factoid and list questions

These are questions that require in general a particular biomedical entity name (eg: a drug, a disease,etc), a number, or a similar short expression as answer.

For these two types of questions, we used BioBERT-large [10] fine-tuned on SQuAD with our novel attention enriching mechanism. BioBERT is based on the transformer [11] architecture, which use a stacked set of self-attentions to encode contextual information for input tokens. The calculation of self-attention weights inside BioBERT is as follow, giving a sequence of input vectors $H = [h_1, h_2, \dots, h_{|H|}]$ its self-attention representation is computed from the scaled dot-product of the query vector $Q = W^Q H$ and the key vector $K = W^K H$ followed by a softmax normalization. Where d_k is the dimension size of the key vector K and V is the value vactor that is also projected from the hidden vectors from the previous layer along with the query vector $Q$and the key vector K. The result is the matrix of attention weights $\alpha \in \mathbb{R}^{|H|\times|H|}$

$$\alpha = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \qquad (1)$$

We based our attention enriching mechanism approach on the hypothesis that tokens with bigger attention scores may have bigger probabilities of being selected as the response start and end tokens by BioBERT's QA layer [12]. First we tagged all biomedical entities in the question and context of the training and test sets. We also used spaCy[5] named entity recognition (NER) module to tag dates, organizations, percent, quantity, and cardinal values. Using the tagged biomedical and NER entities, we constructed a two-dimensional matrix M from the question and context where only biomedical and NER tokens are set to one, and zero for all other tokens. We fed the resulting matrix to BioBERT along with the question and context during training and testing phases following the standard usage of BERT [13] for QA. The resulting matrix is then added to BioBERT's attention scores during self-attention.

$$\alpha = softmax\left(\frac{QK^T + M}{\sqrt{d_k}}\right)V \qquad (2)$$

We adopted this approach for both list and factoid questions. For list questions, we also used the sequence tagging technique [14]

2.2.2. Yes/No questions

For this type of questions, the systems may return only "yes" or "no" as answers. Here, we used BioBERT-large fine-tuned on the BoolQA [15] and PubMedQA [16] datasets.

2.2.3. Summary and ideal answers

These are questions that can only be answered by extracting or generating a short paragraph-sized answer. This is the case of summary questions. Nevertheless, extracted or generated ideal answers can

[5] https://spacy.io

also be provided for factoid, list, and yes/no questions. Here, we used BART-large [17] fine-tuned on the CNN [18], and the ebmsum datasets [19]

3. Results and Discussion

We participated with one system in each phase under the name LaRSA. The reported results below are obtained from the BioASQ10B official leader board. For each phase and question type, we report the batch number, our score, the top score obtained by the best performing system, along with our ranking compared to the total number of systems participating in each batch and each question type. We counted our ranking based on the scores, we considered multiple systems having the same score as one system.

3.1. Phase A

We only began our participation in this phase from the fourth batch. In Table 1 we present our results for document retrieval and snippet extraction. The relative poor performance of our system for this phase demonstrate that the generic IR models and algorithms that we used are not enough for the case of biomedical documents. Models and techniques specific to the biomedical domain must be used in order to yield competitive results.

Table 1
Results of our participation in phase A

Batch	Phase	Mean precision			F-Measure			MAP		
		Score	Top	Rank	Score	Top	Rank	Score	Top	Rank
4	Documents	11.52	25.00	7/27	17.03	27.18	10/27	33.42	40.58	16/27
	Snippets	05.53	12.70	14/16	07.20	16.19	14/16	13.93	66.06	14/16
5	Documents	08.78	29.12	21/25	13.71	31.60	21/25	30.78	41.54	19/25
	Snippets	05.81	10.09	12/12	08.27	14.61	12/12	21.47	56.13	12/12
6	Documents	06.22	07.66	11/26	08.72	10.79	11/26	09.62	17.04	13/26
	Snippets	03.57	06.01	6/14	04.24	06.99	6/14	04.47	14.20	12/14

3.2. Phase B

3.2.1. Factoid questions

We submitted results for exact answers for all the six batches of phase B. In the first batch, we scored first for the lenient accuracy metric. Except for the second batch, we were able to be in the top three for at least one metric for all the other batches. We consider these results promising, giving that this is our first participation in BioASQ.

Table 2
Results of our model for factoid questions in phase B

Batch	Strict Acc.			Lenient Acc.			MRR		
	Score	Top	Rank	Score	Top	Rank	Score	Top	Rank
1	0.2941	0.4118	12/21	**0.5588**		**1/21**	0.4118	0.4608	9/21
2	0.4412	0.5588	15/23	0.5882	0.6765	10/23	0.5098	0.6000	15/23
3	0.5000	0.5313	3/29	0.6563	0.6875	2/29	0.5677	0.5792	4/29
4	0.5806	0.4516	5/29	0.6129	0.6774	3/29	0.5129	0.5995	11/29
5	0.4138	0.4828	3/27	0.5517	0.6207	3/27	0.4540	0.5098	5/27
6	0.1667	0.3333	2/20	0.1667	0.5000	3/20	0.1667	0.3333	9/20

3.2.2. List questions

Our model for list questions yielded poor results compared to our model for factoid questions. Even though, we used the same attention enriching mechanism technique for factoid and list questions. We plan to investigate this fact now that the competition has ended.

Table 3

Results of our model for list questions in phase B

Batch	Mean Prec.			Recall			F-Measure		
	Score	Top	Rank	Score	Top	Rank	Score	Top	Rank
1	0.5714	0.7201	8/20	0.4821	0.8464	11/20	0.4959	0.7469	11/20
2	0.6143	0.7042	7/23	0.4281	0.7530	9/23	0.4490	0.7051	9/23
3	0.4923	0.6273	14/28	0.4128	0.6742	11/28	0.4052	0.5655	13/28
4	0.4736	0.6162	9/28	0.3104	0.5844	14/28	0.3048	0.5386	15/28
5	0.4486	0.6799	17/25	0.4188	0.6407	15/25	0.4191	0.6016	17/25
6	0.5044	0.5730	5/19	0.3669	0.4690	8/19	0.3854	0.4534	5/19

3.2.3. Yes/No questions

Our model for yes/no questions ranked in the top three positions in at least one metric for batches one, four, and six. In the second batch, a technical issue in our result generation script prevented us from submitting the model results, therefore, we set the response to No for all questions, due to time constraints. This explains the big difference in scores in the second batch.

Table 4

Results of our model for Yes/No questions in phase B

Batch	Accuracy			Macro F1		
	Score	Top	Rank	Score	Top	Rank
1	0.9565	1	2/25	0.9464	1	2/25
2	0.3889	1	10/35	0.3378	1	11/35
3	0.8800	1	4/38	0.8252	1	4/38
4	0.9583	1	2/39	0.9473	1	3/39
5	0.7500	0.9286	6/41	0.7333	0.9282	10/41
6	0.6667	1	3/32	0.6250	1	4/32

3.2.4. Summary and ideal answers

We only began submitting results for summary questions and ideal answers starting from the fourth batch. Our model for summary questions, and ideal answers, scored first for the R-2 (F1) metric in batch 6. We were also able to score second for two metrics in the fourth batch.

Table 5

Results of our model for summary questions and ideal answers in phase B

Batch	R-2 (Rec)		R-2 (F1)		R-SU4 (Rec)		R-SU4 (F1)	
	Score	Top	Score	Top	Score	Top	Score	Top
3	0.4616	0.5851	0.2877	0.3761	0.4663	0.5948	0.2735	0.3689
4	0.5458	0.5752	0.4132	0.4229	0.5482	0.5884	0.4022	0.4165
5	0.4926	0.6071	0.3500	0.4020	0.4975	0.5984	0.3423	0.3916
6	0.1782	0.1927	**0.1528**		0.2043	0.2347	0.1691	0.1705

4. Conclusion and future work

In this paper, we presented our participation in the 10th edition of the BioASQ annual challenge. We participated in both phases of task B. We submitted results for both exact and ideal answers. For phase A, we used a classical IR pipeline composed of a retriever based on ElasticSearch with BM25, a RoBERTA based reader, and a ranker. For phase B, we extended BioBERT with our novel attention enriching mechanism for factoid and list questions. We made use of transferability for yes/no questions. And, for summary and ideal answers, we used BART. While our participating systems yielded medium to relatively poor results in general. We were able to rank in the top three positions for some metrics in numerous batches. We plan to further analyze the results of our first participation this year, and exploit the resulting insights in next year's BioASQ edition.

5. Acknowledgements

This work was supported by the Moroccan national center for scientific and technical research (CNRST) and the Google Cloud Research Credits program.

6. References

[1] Z. Kaddari, Y. Mellah, J. Berrich, T. Bouchentouf, and M. G. Belkasmi, "Biomedical question answering: A survey of methods and datasets," pp. 1–8, 2020.

[2] J. Lee, S. S. Yi, M. Jeong, M. Sung, W. Yoon, Y. Choi, M. Ko, and J. Kang, "Answering questions on COVID-19 in real-time," in Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Online: Association for Computational Linguistics, Dec. 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.nlpcovid19-2.1

[3] K. Suderman, N. Ide, V. Marc, B. Cochran, and J. Pustejovsky, "AskMe: A LAPPS Grid-based NLP query and retrieval system for covid-19 literature," in Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Online: Association for Computational Linguistics, Dec. 2020. [Online]. Available: https://www.aclweb.org/anthology/2020.nlpcovid19-2.28

[4] D. Su, Y. Xu, T. Yu, F. B. Siddique, E. J. Barezi, and P. Fung, "Caire-covid: A question answering and query-focused multi-document summarization system for covid-19 scholarly information management," 2020.

[5] A. Esteva, A. Kale, R. Paulus, K. Hashimoto, W. Yin, D. Radev, and R. Socher, "Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization," 2020

[6] Z. Kaddari, J. Berrich, N. Rahmoun, S. Belouali and T. Bouchentouf, "INKAD COVID-19 IntelliSearch: a multilingual search engine for answering questions about COVID-19 in real-time from the scientific literature," 2021 Fifth International Conference On Intelligent Computing in Data Sciences (ICDS), 2021, pp. 1-6, doi: 10.1109/ICDS53782.2021.9626759.

[7] M. e. a. Tsatsaronis, Balikas, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics, 2015

[8] Liu, Yinhan and Ott, Myle and Goyal, Naman and Du, Jingfei and Joshi, Mandar and Chen, Danqi and Levy, Omer and Lewis, Mike and Zettlemoyer, Luke and Stoyanov, Veselin, "RoBERTa: A Robustly Optimized BERT Pretraining Approach", DOI: https://doi.org/10.48550/arXiv.1907.11692

[9] P. Rajpurkar, R. Jia, and P. Liang, "Know what you don't know: Unanswerable questions for squad," 2018

[10] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," Bioinformatics, vol. 36, no. 4, pp. 1234–1240, 09 2019. [Online]. Available: https://doi.org/10.1093/bioinformatics/btz682

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, "Attention Is All You Need", DOI: https://doi.org/10.48550/arXiv.1706.03762

[12] Cui, Yiming & Zhang, Weinan & Che, Wanxiang & Liu, Ting & Chen, Zhigang. (2021). Understanding Attention in Machine Reading Comprehension.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", DOI: https://doi.org/10.48550/arXiv.1810.04805

[14] Wonjin Yoon, Richard Jackson, Jaewoo Kang, Aron Lagerberg, "Sequence Tagging for Biomedical Extractive Question Answering", DOI: https://doi.org/10.48550/arXiv.2104.07535

[15] 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 and Short Papers), pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

[16] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

[17] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, Online. Association for Computational Linguistics.

[18] Karl Moritz Hermann and Tomáš Kočiský and Edward Grefenstette and Lasse Espeholt and Will Kay and Mustafa Suleyman and Phil Blunsom, Teaching Machines to Read and Comprehend, (2015), arXiv:1506.03340

[19] Mollá, D., Santiago-Martínez, M.E., Sarker, A. et al. A corpus for research in text processing for evidence based medicine. Lang Resources & Evaluation 50, 705–727 (2016). https://doi.org/10.1007/s10579-015-9327-2