KU-DMIS at BioASQ 9: Data-centric and model-centric approaches for biomedical question answering

Wonjin Yoon[1], Jaehyo Yoo[1], Sumin Seo[1], Mujeen Sung[1], Minbyul Jeong[1], Gangwoo Kim[1] and Jaewoo Kang*[1]

[1]*Korea University, Seoul, 02841, South Korea*
* *Corresponding Author*

Abstract

In this paper, we present approaches for our participation in the 9th BioASQ challenge (Task b - Phase B). Our systems are based on the transformer models with model-centric and data-centric approaches. For factoid-type questions we modified the dataset to increase label consistency, and for list-type questions we apply the sequence tagging model which is a more natural model design for the multi-label task. Our experimental results suggest two main points: better model design can be achieved by reflecting data characteristics such as the number of labels for a data point; and scarce resources such as BioQA datasets can greatly benefit from a data-centric approach with relatively little effort. Our submissions achieve competitive results with top or near top performance in the challenge.

Keywords

BioNLP, Biomedical Natural Language Processing, BioASQ, Biomedical Question Answering

1. Introduction

Question answering (QA) is the task of finding information about the given question from the given document. Biomedical question answering (BioQA) is a specific category of QA tasks where questions and/or the related documents, namely passages, are in the context of the biomedical domain.

BioASQ [1] challenge is a large-scale annual competition for biomedical literature, which encompasses document classification, document retrieval, and QA tasks, and is a one of the richest source for BioQA research. Questions of the challenge are categorized in 4 categories: Factoid-type, List-type, Yes/No-type, and Summary-type questions. Answers of QA tasks are in two formats: exact answer and ideal answer. For an exact answer, the output of the model is

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
wjyoon@korea.ac.kr (W. Yoon); jaehyoyoo@korea.ac.kr (J. Yoo); suminseo@korea.ac.kr (S. Seo); mujeensung@korea.ac.kr (M. Sung); minbyuljeong@korea.ac.kr (M. Jeong); kgw8965@gmail.com (G. Kim); kangj@korea.ac.kr (J. Kang*)
http://wonjin.info/ (W. Yoon); https://minstar.github.io/ (M. Jeong)
0000-0002-6435-548X (W. Yoon); 0000-0002-3600-6362 (J. Yoo); 0000-0001-8703-0322 (S. Seo); 0000-0002-7978-8114 (M. Sung); 0000-0002-1346-730X (M. Jeong); 0000-0003-4581-0384 (G. Kim); 0000-0001-6798-9106 (J. Kang*)

Table 1
Data cleaning operations for the answers.

Operation	Question	Original Answer	Normalized Answer
Sentence to phrase	What family do mDia proteins belong in?	['mDia proteins are members of the formin family']	['formin family']
	What does polyadenylate-binding protein 4 (PABP4) bind to?	['PABP4 binds mRNA poly(A) tails.']	['mRNA poly(A) tails']
	Where are the orexigenic peptides synthesized?	['The orexigenic peptides are sythesized in the hypothalamus.']	['the hypothalamus']
Punctuation marks	What is the effect of CRD-BP on the stability of c-myc mRNA?	['To protect c-myc CRD from endonucleolytic attack.']	['To protect c-myc CRD from endonucleolytic attack']
	What is gingipain?	['A keystone periodontal pathogen. ']	['A keystone periodontal pathogen']
	What is the role of the UBC9 enzyme in the protein sumoylation pathway?	['SUMO-conjugating enzyme']	['SUMO-conjugating enzyme', 'SUMO conjugating enzyme']

Table 2
Data cleaning operation for the questions. Misspelled words are corrected.

Original Question	Cleaned Question
Which **trancription** factor activates the betalain pathway?	Which **transcription** factor activates the betalain pathway?
What happens to H2AX upon DNA **bouble** strand breaks?	What happens to H2AX upon DNA **double** strand breaks?
What is a popular **mesaure** of gene expression in RNA-seq experiments?	What is a popular **measure** of gene expression in RNA-seq experiments?

a word or a short phrase, whereas, for an ideal answer, the output should be one or multiple sentences. Participants can choose to answer both formats or partially.

In this paper, we provide approaches of our participation in QA tasks of the challenge (Task 9b - Phase B) [2]. Our submissions are in both exact and ideal answer formats. Our approaches vary from the type of questions and answers. For exact answers, we utilize both model-centric approach (list-type question) and data-centric approach (factoid-type question) to improve our previous systems [3] for BioASQ 8b. For ideal answers, we apply abstractive summarization method using large-scale language model BART [4].

2. Methods

In this section, we describe details of our approaches. In the Section 2.1, 2.2, and 2.3, we describe our approaches for the exact answers. In the Section 2.4, we describe our unified model for ideal answers, which can answer all four types of questions: 3 aforementioned types and summary-type questions. For all types, we mainly fine-tune BioBERT [5], which has been proven to be effective on various NLP tasks in the biomedical or clinical domain [6, 7, 8, 9, 10, 11].

2.1. Data-centric approach; Factoid Questions

Recently, pre-trained models [12, 13] have achieved dramatic improvements of downstream tasks in both general and biomedical domain by harnessing large-scale models with transfer-learning methods [10, 13, 14, 15]. BioQA models have also benefited from transfer-learning

[3, 11, 16]. However, utilizing the maximum of scarce resources is susceptible to the rare error of the training samples, as opposed to relatively rich datasets where a few erroneous samples can be ignored by the model. Moreover, Jeong et al. [3] measured the proportion of questions, which is unanswerable if transformed to the extractive QA setting, in the test dataset of BioASQ 8b. Hence, the models trying to solve the task under the extractive QA setting are suffering from unexpected noise.

In this section, we introduce our data-centric approach for factoid-type questions of the BioASQ 9b challenge. Data-centric approach is a concept of improving the quality of training data to make it better fit into the model. The term *data-centric* forms a binary opposition term pair with *model-centric* approach which focuses more into improving model to achieve better performance. The concept is introduced by Ng [17] at a seminar, where he argued the benefit of data-centric approaches and showed that, for some datasets, larger improvements in performance can be made with data-centric approach than model-centric approach.

Our main aim of data-centric approach is to increase labeling consistency and exclude or clean noisy data points. Table 1 and 2 shows the data cleaning operations and the examples of them. The answers from the BioASQ 9b training samples are mostly in the format of noun phrases. However, some data points have answers in sentences format. We manually modify such sentence answers to a noun phrase format. Additional normalization processes are made to correct misspelled word and to remove punctuation marks. Word correction and minor normalization processes are made to the questions. We do not modify grammatical structure of the questions and we count both British and American spelling as correct.

152 changes are made to the 9b training dataset including 22 question corrections and 24 dropped data points. For the evaluation of our models, we do not apply normalization steps. Data cleaning resources are available at https://github.com/dmis-lab/bioasq9b-dmis .

2.2. Model-centric approach; List Questions

Extractive question answering is a task of finding answer spans of a question in the given passage. List-type questions are questions with multiple answers whereas factoid questions are questions that can be answered with one phrase. For list-type question, the number of answer for a given question is uninformed (i.e. not available as a metadata). Hence, deciding it remains a challenging and key operation to participating systems.

Previous works utilize factoid models with an additional steps to decide the number of answers for the questions. Factoid models are designed to predict a single answer span, and thus, they can not be trained on multi-label setup directly. In other words, for a training data point of list-type question, one answer span is trained for a training step and the other answers are acting as a noise since they are considered as non-answer tokens. We call this setting as "start-end span prediction", which is commonly used in biomedical extractive QA [3, 11, 16].

Following the approach of Yoon et al. [18], our systems for list-type questions are based on the sequence tagging approach. Specifically, a question and its corresponding passage are concatenated to construct a sequence, which is a training data point. Our systems adopt either BIO or IO scheme to annotate answer spans. For each tokens in the passage is tagged as B, I, or O tag which stands for Beginning, Inside, Outside, respectively.

Sequence tagging approach has two significant benefits over the previous models. First, the

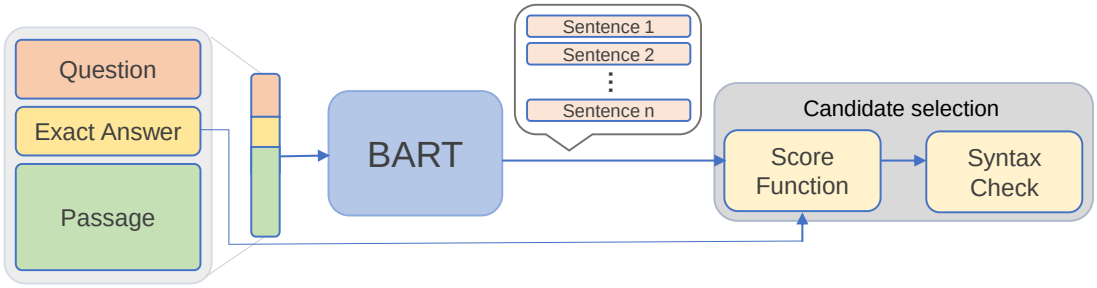

Figure 1: Overview of our systems for ideal answers. Question, passage and the exact answer form an input sequence. *Candidate selection* module scores the candidate sentences and select the best candidate as an ideal answer for a given question.

model can be trained on multiple answers simultaneously which is more natural model design for the multi-answer task. As all available training labels (i.e. answers) are used for the training, rather than acting as training noise, the model learn the maximum of the dataset. Second, the approach is an end-to-end model that finds all the answers in the passage, whereas the previous models require complex post-processing steps to decide the number of answers. Models for the previous BioASQ challenges [3, 11] decided the number of answers by threshold-based answer decision process, where the threshold value is a hyperparameter that needs to be tuned. Additionally, rule-based answer number detection are adopted under the assumption that if the numeric value exists in the question, the value is highly likely to be the required number of answers (ex. Question: List *3* apoE mimetics). In contrast, our approach does not require additional hyperparameter searching nor have to rely on the weak assumption.

2.3. Yes/No Questions

Following the systems of our participation for BioASQ 8b [3], we use BioBERT [5] with an additional pre-training step on MNLI dataset [19]. For BioASQ 9b task, our systems are based on $\text{BioBERT}_{\text{LARGE}}$, which has more model parameters than $\text{BioBERT}_{\text{BASE}}$ model from the previous participation. For yes/no-type questions, we adjust the ratio of *yes* questions and *no* questions in the training set to 1:1. The original training dataset is heavily skewed towards question with *yes* answers. As shown in Table 3, original 8b training set consists of 10,284 *yes* and 1,691 *no* question and passage pair samples. After our pre-processing steps, approximately 85.8% of answers for the yes/no-type questions are *yes*. Our systems are trained on the down-sampled training dataset to alleviate the class imbalance problem.

2.4. Ideal answers

Figure 1 shows an overview of our approach for the ideal answers. Our model utilize the predicted exact answer as a input for generating a ideal answer. In detail, we generate all the combinations of triples of an exact answer (A) and all the passages ($P_1, \ldots, P_n$) available for a question (Q). Then using the triples, we generate a candidate sentences ($C_1, \ldots, C_n$) for a question. We then select one ideal answer from the candidates using candidate selection process.

Table 3
Proportion of question samples having yes or no as its answers. Samples are pre-processed from original questions and consist of question and passage pairs.

	Yes	No
Original Training 8b	10284 (85.8%)	1691 (14.1%)
Down-sampled Training 8b	1691 (50.0%)	1691 (50.0%)

Table 4
Performance of our factoid model on original training dataset and cleaned dataset of BioASQ 8B. The performance is based on the macro average of 5 batches. We only describe results on full-abstract setting.

	SAcc	LAcc	MRR
Original Training 8b	37.75	57.62	45.66
Cleaned Training 8b	**43.71**	**58.94**	**49.14**

Candidate Selection Our model is designed to generate one answer candidate for a input sequences. In the original dataset, multiple articles/snippets are provided as corresponding passages for a question. Hence multiple answer candidates are generated for a question. We need a candidate selection process to submit one answer, and the quality of this process can largely impact the performance of the model. Our candidate selection process is composed of a scoring function and a syntax checker.

Score Function We have tried to score the generated ideal answer candidates ($C_1, \ldots, C_n$) by checking the presence of (candidate) exact answer(s). In order to check the presence, we employ BERN [9], a BioBERT based named entity recognition (NER) and linking system, to detect all the entities in the ideal answer candidates. For each candidate sentences for factoid and list questions, we calculate the F1-score using the tentative exact answer(s) and recognize entities from the candidate sentences. If one of the tentative answer(s) exists in the generated sentences, we add it to the set of recognized entities even if it is not detected by BERN.

Syntax Checker We use **language-check** python package [1], an automated syntax checker, to correct or filter out grammatically wrong candidate sentences. From the list of ideal answer candidates, we select one with the highest score and check it with the syntax checker. If grammar errors are detected, we try to correct it with the checker. However, if the checker finds impossible to fix it, we then skip the candidate and move to the second highest candidate.

3. Results and Discussion

Table 4 presents our experimental results on the factoid dataset. Owing to our data-centric approach (noisy data points exclusion and increasing label consistency), our model perfor-

[1] https://github.com/myint/language-check

Table 5
Experiments of list type on different settings. The scores in the table are reported based on the macro average of 5 batches in BioASQ 8B test sets.

	Setting	Prec	Recall	F1
Start-End	Snippet	40.72	**60.14**	43.20
	Full Abstract	38.63	54.03	40.54
BIO Tagging	Full Abstract	40.10	55.25	42.62
IO Tagging	Full Abstract	**40.79**	55.46	**43.82**

Table 6
Experiments of Yes/No type on different sampling strategy. The performance is based on the macro average of 5 batches in BioASQ 8B test sets. We only describe results on snippet setting.

	Acc	Macro F1	Yes F1	No F1
Original Training 8b	0.82	0.80	0.87	0.73
Down-sampled Training 8b	**0.94**	**0.93**	**0.95**	**0.92**

mance shows 3.48 score improvement on the mean reciprocal rank (MRR) score and 5.96 point improvement on the strict accuracy (SAcc) metric.

Table 5 shows the experimental results on the list dataset for sequence tagging model and the baseline model, namely Start-End model[3] (for both full abstract and snippet datasets). Our model outperforms the baseline model, which we used for the last year, with large gap. Our model with IO tagging even outperforms the start-end model with snippet dataset as input without using complex post-processing and rule-based processing steps. Please note that snippets are more concentrated sources of information than the full abstract documents since the answer should exist in a snippet or a full abstract but a snippet is a sentence-length document whereas a full abstract is a paragraph-length document. In the challenge, we also submitted ensemble system, which consists of sequence tagging model and start-end model. However, ensemble system did not show significant improvement over single models.

For yes/no-type questions, we introduce the down-sampling method which we balance the number of yes and no questions from the training set by sampling 1,691 yes questions out of 10284 yes questions (i.e. sampled 16% of yes questions). We have empirically shown that a down-sampled dataset alleviates the underlying class imbalance issue of yes/no-type questions. The results of yes/no model trained on down-sampled and original training data are described in Table 6. For models trained on down-sampled data, macro F1 score is improved from 0.80 to 0.93 (0.13 improvement). We have found that the down sampling method achieves better model performance for imbalanced yes/no-type questions. We conduct the experiments following the same hyperparameter settings, including learning rate and the number of learning steps.

The official evaluation metric for the ideal models is *human evaluation score*, which is scored by the human experts on the following four criteria: recall, precision, repetition, and readability. Since the human evaluation score is not yet released at the moment[2], we provide the qualitative

[2]The official result (human evaluation) will be on : http://participants-area.bioasq.org/results/9b/phaseB/

Table 7
Qualitative analysis of ideal answers on BioASQ 9B test sets. Tentative exact answers, which are the outputs of our exact models, are denoted as *Exact Answer* in the table.

Type	Question - Exact Answer - Ideal Answer
Factoid	Question: Which disease is monitored in the BIOCURA cohort? Exact Answer: **rheumatoid arthritis** Ideal Answer: Baseline disease activity in patients with **rheumatoid arthritis** is monitored in the BIOCURA cohort.
List	Question: What are the years of the initiation and completion of the Human Genome project? Exact Answer: **1990**, **2003** Ideal Answer: The Human Genome Project (HGP) was initiated in **1990**, and the completion of the genome project was in **2003**.
Yes/No	Question: Is Tocilizumab (Actemra) used to block/antagonize the IL-6 receptor? Exact Answer: **yes** Ideal Answer: **Yes**, Tocilizumab (Actemra) is a humanized monoclonal antibody that binds to and inhibits the IL-6 receptor. It is used for treatment of type 1 diabetes.

Table 8
BioASQ 9B results for our systems. The scores are obtained from the leaderboard. Score in **bold font** denotes that our system achieved the highest score from the leaderboard.

# of Batch	Factoid (MRR)		List (F1)		Yes/No (Macro F1)	
	Ours	Top-1	Ours	Top-1	Ours	Top-1
9B Batch 1	38.79	46.32	**53.39**		**92.58**	
9B Batch 2	52.94	55.39	46.44	48.92	88.54	94.54
9B Batch 3	42.34	61.49	54.21	58.87	90.23	95.32
9B Batch 4	57.26	69.29	**70.61**		**94.80**	
9B Batch 5	47.45	58.80	45.60	51.75	**84.21**	

analysis of the model on Table 7. As we denoted in Section 2.4, tentative exact answers are used as one of the inputs to generate ideal answer candidates and to select best candidates for the questions. By utilizing the exact answer as one of the input to the ideal answer generation model, the model is imposed to consider the given exact answer in the sentence generation step and potentially include the exact answer in the generated sentence. The examples in the Table 7 show that the generated ideal answer sentences successfully include the given exact answers (potential exact answers). We did not exclude the snippets that do not contain the exact answer since our model is expected to include the given exact answers in the generated sentences even if they dose not exist in the passage.

Finally, Table 8 shows the result of our participation and the best performing system in the challenge. If our system scored the highest performance for a given batch, we marked it using bold font.

4. Conclusion

In this paper, we report our participation with data-centric and model-centric approaches, and the results of our systems for the BioASQ 9b task. We introduce a data-centric approach for factoid-type questions, in which we train the model using the dataset with increased label consistency, and improves the performance of our model. For list-type questions, we apply the sequence tagging model and achieve better performance while at the same time lower the cost of pre- and post-processing. Furthermore, for yes/no-type questions in the BioASQ 9b, our models have shown the best performance by utilizing down-sampling. We can conclude that it is beneficial for BioQA models to use data-centric and/or model-centric approaches in consideration of the features of questions and answers.

Acknowledgments

We express gratitude towards Dr. Jihye Kim and Dr. Sungjoon Park from Korea University for their invaluable insight into our systems' output. This research is supported by National Research Foundation of Korea (NRF-2020R1A2C3010638) and a grant of the the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: HR20C0021)

References

[1] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, et al., An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, BMC bioinformatics 16 (2015) 1–28.

[2] A. Krithara, A. Nentidis, G. Paliouras, M. Krallinger, A. Miranda, Bioasq at clef2021: Large-scale biomedical semantic indexing and question answering., in: ECIR (2), 2021, pp. 624–630.

[3] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, J. Kang, Transferability of natural language inference to biomedical question answering, arXiv preprint arXiv:2007.00217 (2020).

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, 2019. arXiv:1910.13461.

[5] 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.

[6] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://www.aclweb.org/anthology/W19-1909. doi:10.18653/v1/W19-1909.

[7] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, 2019. arXiv:1903.10676.

[8] Q. Jin, B. Dhingra, W. W. Cohen, X. Lu, Probing biomedical embeddings from language models, 2019. arXiv:1904.02181.

[9] D. Kim, J. Lee, C. So, H. Jeon, M. Jeong, Y. Choi, W. Yoon, M. Sung, J. Kang, A neural named entity recognition and multi-type normalization tool for biomedical text mining, IEEE Access 7 (2019) 73729–73740. doi:10.1109/ACCESS.2019.2920708, funding Information: This work was supported in part by the National Research Foundation of Korea under Grant NRF-2017R1A2A1A17069645 and Grant NRF-2016M3A9A7916996, and in part by the National IT Industry Promotion Agency, Development Project of the Precision Medicine Hospital Information System (P-HIS), under Grant C1202-18-1001. Publisher Copyright: © 2013 IEEE.

[10] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, 2019. arXiv:1906.05474.

[11] W. Yoon, J. Lee, D. Kim, M. Jeong, J. Kang, Pre-trained language model for biomedical question answering, 2019. arXiv:1909.08229.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contesxtualized word representations, 2018. arXiv:1802.05365.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[14] N. Kim, R. Patel, A. Poliak, P. Xia, A. Wang, T. McCoy, I. Tenney, A. Ross, T. Linzen, B. Van Durme, S. R. Bowman, E. Pavlick, Probing what different NLP tasks teach machines about function word comprehension, in: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 235–249. URL: https://www.aclweb.org/anthology/S19-1026. doi:10.18653/v1/S19-1026.

[15] J. Phang, T. Févry, S. R. Bowman, Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks, 2019. arXiv:1811.01088.

[16] G. Wiese, D. Weissenborn, M. Neves, Neural domain adaptation for biomedical question answering, in: Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 281–289. URL: https://www.aclweb.org/anthology/K17-1029. doi:10.18653/v1/K17-1029.

[17] A. Y. Ng, A chat with andrew on mlops: From model-centric to data-centric ai, 2021. URL: https://www.youtube.com/06-AZXmwHjo.

[18] W. Yoon, R. Jackson, J. Kang, A. Lagerberg, Sequence tagging for biomedical extractive question answering, arXiv preprint arXiv:2104.07535 (2021).

[19] A. Williams, N. Nangia, S. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1112–1122. URL: https://www.aclweb.org/anthology/N18-1101. doi:10.18653/v1/N18-1101.