

Query-focused Extractive Summarisation for Biomedical and COVID-19 Complex Question Answering

Macquarie University's Participation at BioASQ10 Synergy and BioASQ10b Phase B

Diego Mollá¹

¹Macquarie University, Australia

Abstract

This paper presents Macquarie University's participation to the two most recent BioASQ Synergy Tasks (as per June 2022), and to the BioASQ10 Task B (BioASQ10b), Phase B. In these tasks, participating systems are expected to generate complex answers to biomedical questions, where the answers may contain more than one sentence. We apply query-focused extractive summarisation techniques. In particular, we follow a sentence classification-based approach that scores each candidate sentence associated to a question, and the n highest-scoring sentences are returned as the answer. The Synergy Task corresponds to an end-to-end system that requires document selection, snippet selection, and finding the final answer, but it has very limited training data. For the Synergy task, we selected the candidate sentences following two phases: document retrieval and snippet retrieval, and the final answer was found by using a DistilBERT/ALBERT classifier that had been trained on the training data of BioASQ9b. Document retrieval was achieved as a standard search over the CORD-19 data using the search API provided by the BioASQ organisers, and snippet retrieval was achieved by re-ranking the sentences of the top retrieved documents, using the cosine similarity of the question and candidate sentence. We observed that vectors represented via sBERT have an edge over tf.idf. BioASQ10b Phase B focuses on finding the specific answers to biomedical questions. For this task, we followed a data-centric approach. We hypothesised that the training data of the first BioASQ years might be biased and we experimented with different subsets of the training data. We observed an improvement of results when the system was trained on the second half of the BioASQ10b training data.

Keywords

BioASQ, Synergy, query-focused summarisation, Biomedical, COVID-19, DistilBERT, sBERT, data-centric

1. Introduction

The BioASQ challenge¹ organises shared tasks on biomedical semantic indexing and question answering. In this paper, we present Macquarie University's participation in several of these tasks.²

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

✉ diego.molla-aliiod@mq.edu.au (D. Mollá)

🌐 <https://researchers.mq.edu.au/en/persons/diego-molla-aliiod> (D. Mollá)

🆔 0000-0003-4973-0963 (D. Mollá)



© 2022 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)

¹<http://www.bioasq.org/>

²Code related to this paper is available at <https://github.com/dmollaaliiod/bioasq10b-public> and <https://github.com/dmollaaliiod/bioasq10-synergy-public>.

Question Orteronel was developed for treatment of which cancer?

Type factoid

Snippet Pooled-analysis was also performed, to assess the effectiveness of agents targeting the androgen axis via identical mechanisms of action (abiraterone acetate, orteronel).

Exact answer castration-resistant prostate cancer

Ideal answer Orteronel was developed for treatment of castration-resistant prostate cancer.

Figure 1: An example question with its question type, a relevant snippet, an exact answer, and a correct ideal answer, extracted from the training data of BioASQ10b

The Synergy tasks aim to evaluate technologies useful for the development of an end-to-end question answering (QA) system for questions about COVID-19 asked by biomedical experts. In particular, the Synergy tasks evaluate the quality of document retrieval over a snapshot of COVID-19 [1], snippet retrieval, and the generation of “ideal answers” that may contain multiple sentences. We present our participation in the second BioASQ9 Synergy task that ran between May and June 2021, and the BioASQ10 Synergy task that ran between December 2021 and February 2022.

Task B of BioASQ focuses on biomedical semantic QA. Similar to the Synergy tasks, several technologies corresponding to components of an end-to-end QA system are evaluated. In contrast with the Synergy tasks, Task B of BioASQ has two distinct phases. Phase A evaluates the quality of document and snippet retrieval on a snapshot of PubMed³, whereas Phase B, given a question, its question type (“summary”, “factoid”, “yesno”, “list”), and a list of candidate snippets, evaluates the system’s ability to find short answers (“exact answers”) and long, possibly multi-sentence answers (“ideal answers”). Figure 1 shows an example of a question and its question type, a correct snippet for the question, a correct exact answer, and a correct ideal answer. We present our participation in Task B, Phase B of BioASQ10, that ran between March and May 2022 (henceforth BioASQ10b, Phase B).

All of our contributions to the above tasks are based on a common question-answering architecture that we will describe in Section 2. Section 3 presents our participation in the Synergy tasks. Section 4 presents our participation in BioASQ10b, Phase B. Finally, Section 5 concludes this paper.

2. Question Answering Architecture

The question-answering system that is the focus of our participation in all of the tasks presented in this paper is based on query-focused extractive summarisation. The architecture of the system is illustrated in Figure 2, and follows the classification set up proposed by [2].

The query-focused summarisation system takes the question, a candidate sentence, and the sentence position⁴, and calculates a sentence score. The system computes the word embeddings

³<https://pubmed.ncbi.nlm.nih.gov/>

⁴The sentence position was incorporated as an absolute number: 1, 2, ... n , where n is the total number of input sentences. We chose to include the sentence position as earlier experiments in past BioASQ years showed an

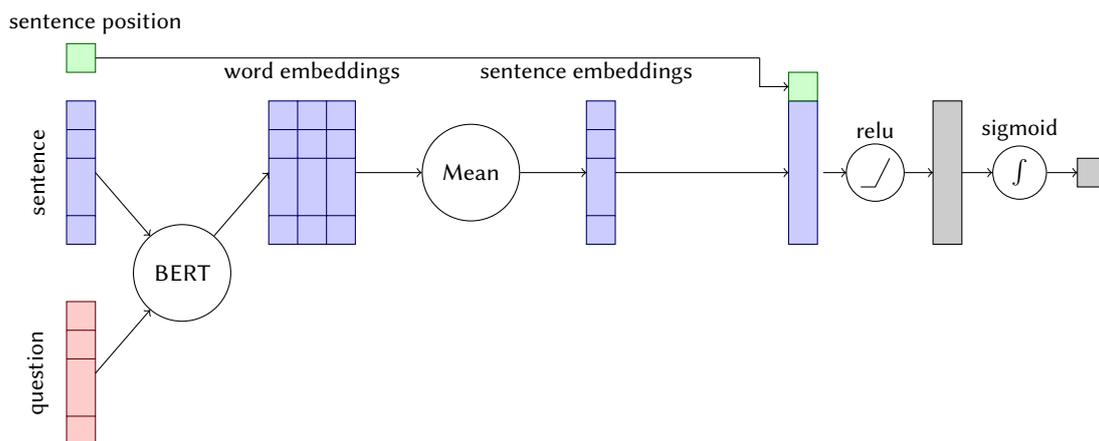


Figure 2: Architecture of the question answering system used for BioASQ9b, Phase B.

of the question and candidate sentence using a BERT architecture [3]. In particular, for the BioASQ9 Synergy task 2 we used ALBERT [4], which was the best-performing system in [5]’s participation in BioASB8b⁵. For the BioASQ10 Synergy task, we used DistilBERT [6], which performed very well in [2]’s participation in BioASQ9b, and even outperformed BioBERT [7]. For BioASQ10, Phase B, we also used DistilBERT. Average pooling is then used to merge the word embeddings of the candidate sentence into the sentence embeddings. The sentence position is then concatenated to the sentence embeddings, and an additional intermediate dense layer is added. A final classification layer predicts the sentence score.

The question and the sentence were fed to BERT in the same way as defined by the creators of BERT [3]. That is, the input consisted of an initial “[CLS]” token, followed by the question text, then a “[SEP]” token that indicates a new sentence, and finally the candidate sentence text. This information was passed to BERT, indicating the question and the candidate sentence as two separate text segments.

The classification labels used for training the system were automatically generated from the training data, based on the ROUGE score of the candidate sentence with respect to the annotated ideal answer. In particular, given a particular question, the top 5 sentences according to their ROUGE score were labelled as 1, and the rest were labelled as 0. For the Synergy tasks we used the BioASQ9b training data, whereas for BioASQ10b, Phase B, we used the BioASQ10b training data.

We used the pre-trained ALBERT and DistilBERT models available by Huggingface⁶. These models were frozen during training, so that only the weights of the additional layers shown in Figure 2 were updated.

improvement of the results.

⁵At the time of training the system for the BioASQ9 Synergy task 2, the final results of BioASQ9 had not been released yet.

⁶<https://huggingface.co/> – For ALBERT, we used ‘albert-xxlarge-v2’. For DistilBERT, we used ‘distilbert-base-uncased’.

3. The Synergy Tasks

This section describes the systems that participated in the Synergy task 2 of BioASQ9, and the Synergy task of BioASQ10 (in this paper, we will use the collective expression “the Synergy tasks” to refer to these). The Synergy task 2 of BioASQ9 ran in 2021 but the results were not made available at the time of the paper submission deadline for BioASQ9. For this reason, we are describing the system in this paper.

Our participation in the Synergy tasks share the same question answering system architecture described in Section 2. The only difference between the two Synergy tasks is, as mentioned in Section 2, that the BioASQ9 Synergy 2 system used ALBERT, whereas the BioASQ10 Synergy system used DistilBERT. In both cases, the system was trained with the training data of BioASQ9b.

To generate the candidate sentences required by the question answering system, we followed this procedure:

1. Retrieve the most relevant documents as described in Section 3.1;
2. Split the retrieved documents into sentences and select the candidate sentences as described in Section 3.2.

3.1. Document Retrieval

The relevant documents were retrieved using the search API provided by the organisers of the BioASQ Synergy task. This API is based on a Web service that accepts a query and returns a JSON data structure. We simply used the unmodified question as the search query. In subsequent work we are exploring pre-processing and fine-tuning steps to improve the quality of the Document Retrieval stage.

The final runs submitted consist of the top 10 documents, after removing those that were in previous feedback, to conform with the submission requirements.

3.2. Snippet Retrieval

Every sentence from every retrieved document was a candidate snippet. This includes sentences from documents that were retrieved but were not submitted in the Document Retrieval runs. We then experimented with the combination of 2 dimensions to re-rank the candidate snippets, for a total of 4 different approaches.

The first dimension was based on the calculation of the similarity between the question and candidate snippet. We experimented with the following two options:

TfidfCosine. We represented the question and candidate sentences using tf.idf. Each candidate sentence was then scored based on the cosine similarity between the question vector and the sentence vector.

Table 1

Number of sentences selected, for each question type

	Summary	Factoid	Yesno	List
n	6	2	2	3

sBERTCosine. We used sBERT [8] to represent the question and the candidate sentences, and to determine the similarities between the question and the sentences. We used the default set up for sBERT, which computes the cosine similarity between the question vector and the sentence vector.

The second dimension was based on the criteria used for the final ranking of the candidate sentences. We experimented with local sorting and global sorting.

LocalSorting. For every relevant document, we extracted the top 3 sentences according to the cosine similarity approaches described above. The final list of sentences was composed of the top 3 sentences from the top document, followed by the top 3 sentences of the second document, and so on.

GlobalSorting. In contrast to the local sorting approach, all sentences of all documents were now sorted according to their cosine similarity with the question, regardless of what document the snippets were obtained from.

The final runs submitted consist of the first 10 snippets, after removing those that were in previous feedback, to conform with the submission requirements.

3.3. Answer Generation

As mentioned above, the question, candidate sentences, and sentence position were fed to the system illustrated in Figure 2. The sentence position was simply the unnormalised position of the sentence within the list of snippets, after the snippets have been ranked as described in Section 3.2. Given a question, the top-scoring n sentences according to the scores produced by the QA system were combined to form the final answer. These sentences were presented in the order of appearance in the list of snippets. The value of n was based on the question type and is shown in Table 1.

3.4. Results of the Synergy Tasks

This section describes the results of the runs submitted to the Synergy tasks.

Table 2 shows the F1 score of the documents returned by our systems. As mentioned in Section 3.1, these documents were found by submitting the unmodified question as the query to the search API provided by the developers of the Synergy task. As expected, the results were poor relative to other submissions.

Table 3 shows the F1 score of the snippets returned by our runs. For each run, we indicate the run name, the type of similarity used, and the type of sorting performed. We observe that,

Table 2

Document retrieval results of the submissions to the BioASQ9 Synergy 2 (top) and BioASQ10 Synergy (bottom) tasks. Metric: F1. The results of rows labelled “Best”, “Median”, and “Worst” refer to the results of other systems, other than our own, submitted to the challenge.

Run	Round 1	Round 2	Round 3	Round 4
Best	0.3693	0.2039	0.1327	0.1896
Median	0.2388	0.1423	0.0710	0.0800
Worst	0.0157	0.0067	0.0053	0.0175
MQ-BioASQ9	0.1978	0.1087	0.0483	0.0800
Best	0.3220	0.2221	0.1970	0.1564
Median	0.3100	0.1646	0.1327	0.1067
Worst	0.2729	0.1003	0.0655	0.0478
MQ-BioASQ10		0.1003	0.0754	0.0808

Table 3

Snippet retrieval results of the submissions to the BioASQ9 Synergy 2 (top) and BioASQ10 Synergy (bottom) tasks. Metric: F1. The best of our systems in each round is highlighted in **bold**. The results of rows labelled “Best”, “Median”, and “Worst” refer to the results of other systems, other than our own, submitted to the challenge.

Run	Similarity	Sorting	Round 1	Round 2	Round 3	Round 4
Best			0.3290	0.1726	0.1262	0.1355
Median			0.2288	0.1365	0.0732	0.0764
Worst			0.0311	0.0101	0.0231	0.0132
MQ-1-BioASQ9	tfidf	local	0.1031	0.1035	0.0707	0.0764
MQ-2-BioASQ9	tfidf	global	0.1100	0.0540	0.0324	0.0619
MQ-3-BioASQ9	sBERT	local	0.1071	0.0999	0.0692	0.0749
MQ-4-BioASQ9	sBERT	global	0.1923	0.1075	0.1044	0.0762
Best			0.2910	0.1525	0.1574	0.1217
Median			0.2757	0.1410	0.1087	0.0948
Worst			0.2296	0.0540	0.0273	0.0416
MQ-1-BioASQ10	tfidf	local		0.0660	0.0465	0.0771
MQ-2-BioASQ10	tfidf	global		0.0540	0.0273	0.0416
MQ-3-BioASQ10	sBERT	local		0.0683	0.0457	0.0770
MQ-4-BioASQ10	sBERT	global		0.0928	0.0725	0.0827

considering the poor quality of the documents retrieved, the snippets were of quality comparable to that of other runs of the BioASQ9 Synergy 2 task (but not the runs of the BioASQ10 Synergy task), but there is room for improvement. Among our runs, the most successful configuration was using sBERT cosine similarity and global sort.

Table 4 shows the human evaluation results of the ideal answers returned by our runs. Our runs are very competitive, especially given the relatively poor quality of the input snippets. Given the poor quality of the input snippets in all of our runs, it is dangerous to make generalisations about how the quality of the snippets affect the quality of the answers. Having said that, we can observe that, in the BioASQ9 Synergy 2 task, the runs that generated the best snippets

Table 4

Ideal answer results of the submissions to the BioASQ9 Synergy 2 (top) and BioASQ10 Synergy (bottom) tasks. Metric: Average of human evaluation scores. The best of our systems in each round is highlighted in **bold**. The results of rows labelled “Best”, “Median”, and “Worst” refer to the results of other systems, other than our own, submitted to the challenge.

Run	Similarity	Sorting	Round 1	Round 2	Round 3	Round 4
Best			4.375	3.850	3.630	3.295
Median			3.625	3.100	3.450	3.045
Worst			1.042	0.450	3.290	2.060
MQ-1-BioASQ9	tfidf	local	3.250	3.100	3.450	3.045
MQ-2-BioASQ9	tfidf	global	3.210	3.075	3.290	3.295
MQ-3-BioASQ9	sBERT	local	3.372	3.250	3.520	3.067
MQ-4-BioASQ9	sBERT	global	2.250	3.025	3.490	3.292
Best			3.790	3.810	3.562	3.180
Median			3.367	3.160	3.250	2.617
Worst			3.287	1.550	0.827	0.372
MQ-1-BioASQ10	tfidf	local		3.270	3.415	2.617
MQ-2-BioASQ10	tfidf	global		3.160	3.305	2.990
MQ-3-BioASQ10	sBERT	local		3.360	3.517	2.925
MQ-4-BioASQ10	sBERT	global		3.490	3.547	2.690

(MQ-4) did not lead to generating the best ideal answers. The impact of and interplay between the document and snippet retrieval stages, and the question-answering stage, deserves further exploring.

4. BioASQ10b, Phase B

For BioASQ10b, Phase B, we used the question answering system described in Section 2, using DistilBERT as the BERT variant chosen to compute the word embeddings. Following a data-centric approach, the main difference between the Synergy tasks and BioASQ10, Phase B, is the choice of training data. We hypothesised that the training data that corresponds to the early years of BioASQ, that is, the first samples of the BioASQ10b training data, might be biased. We therefore tested the use of different portions of the training data as shown in Table 5, by incrementally removing the first samples of the training data. We can observe that best evaluation results are obtained with only 50% of the training data.

To double-check that indeed the first samples of the training data are biased, we conducted another round of experiments, but this time removing the last samples of the training data. Table 6 shows that results worsen as the amount of training data diminishes, as one might expect in systems that are based on supervised approaches to machine learning.

Hyperparameter search showed that the same hyperparameters give optimal results when training using the entire training data, or using only 50% of the training data: dropout=0.6, number of epochs=1.

Table 5

Results of 10-fold cross-validation after removing the *first* samples of the BioASQ10b training data. Metric: Average ROUGE-SU4 F1. Best result shown in **bold**.

Percentage removed	ROUGE-SU4 F1
10%	0.281
20%	0.288
30%	0.298
40%	0.309
50%	0.311
60%	0.308

Table 6

Results of 10-fold cross-validation after removing the *last* samples of the BioASQ10b training data. Metric: Average ROUGE-SU4 F1.

Percentage removed	ROUGE-SU4 F1
10%	0.275
20%	0.268
30%	0.270
40%	0.255
50%	0.241
60%	0.229

Table 7

Preliminary results of the submissions to BioASQ10b, Phase B. The best of our systems in each batch is highlighted in **bold**. The results of rows labelled “Best”, “Median”, and “Worst” refer to the results of all systems, including our own, submitted to the challenge.

Run	Training Data	ROUGE-SU4 F1					
		Batch 1	Batch 2	Batch 3	Batch 4	Batch 5	Batch 6
Best		0.3715	0.4168	0.3689	0.4165	0.3916	0.1705
Median		0.3339	0.3521	0.3387	0.3556	0.3389	0.1581
Worst							
MQ-1	All BioASQ10b	0.3490	0.3484	0.3344	0.3525	0.3415	0.1581
MQ-2	Last 50% of BioASQ10b	0.3339	0.3480	0.3316	0.3556	0.3431	0.1640

4.1. Submission Results to BioASQ10b, Phase B

Table 7 shows the results of our submissions to BioASQ10b, Phase B⁷. Note that the results reported in the BioASQ website⁸ may change in the future after the test data is potentially enriched with further annotations.

Our runs are comparable to the median of those of other participating systems. Surprisingly, there is little difference between using all training data or only the latter 50%. When we visually inspected the outputs of the runs, we noticed that the output of all runs in each batch were

⁷At the time of writing, only the automated evaluation results were available.

⁸<http://bioasq.org>

virtually identical, with only a few differences.

5. Summary and Conclusions

We have presented Macquarie University’s contribution to the BioASQ9 Synergy task 2, the BioASQ10 Synergy task, and BioASQ10b, Phase B (Ideal Answers). In all of our runs, the base question answering architecture was virtually the same, the only differences being the choice of DistilBERT vs. ALBERT, and the training data used.

For the synergy tasks, we used a system that has been trained using BioASQ9b training data. We experimented with approaches for snippet retrieval based on two dimensions: vectors used for similarity comparison, and final ranking approach. Cosine similarity using sBERT gave the best results, and we observed that not always the best snippets for the snippet retrieval task led to best answers in the question answering task.

Overall, the results of the question answering parts were competitive, especially given the relatively poor quality of the documents and snippets retrieved. We will investigate approaches to increase the quality of the retrieval stages, and explore the relation between quality of retrieval vs. quality of final answers.

For the BioASQ10b, Phase B task, we followed a data-centric approach and experimented with training regimes that incrementally removed samples from the training data. During our preliminary cross-validation experiments we observed an improvement of results using only the latter 50% of the training data, but this difference of results vanished in the submitted runs.

With a data-centric approach in mind, we plan to conduct further experiments that test the impact of changes and transformations of the training data. For example, besides further examining the impact of using portions of the training data, we will investigate the use of data augmentation techniques.

Acknowledgments

This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

References

- [1] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, *CORD-19: The COVID-19 open research dataset*, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL: <https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1>.
- [2] D. Mollá, U. Khanna, D. Galat, V. Nguyen, M. Rybinski, *Query-focused extractive summarisation for finding ideal answers to biomedical and COVID-19 questions*, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), Working Notes of CLEF 2021 — Conference and

Labs of the Evaluation Forum, Bucharest, 2021. URL: <http://ceur-ws.org/Vol-2936//paper-20.pdf>.

- [3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
- [4] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, ALBERT: A lite BERT for self-supervised learning of language representations, in: Proceedings of the 8th International Conference on Learning Representations, Virtual, 2020. URL: https://iclr.cc/virtual_2020/poster_H1eA7AEtvS.html.
- [5] D. Mollá, C. Jones, V. Nguyen, Query-focused multi-document summarisation of biomedical texts, in: L. Cappellato, C. Eickhoff, N. Ferraro (Eds.), Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum, Thessaloniki, 2020. URL: http://ceur-ws.org/Vol-2696/paper_119.pdf.
- [6] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), 2019. arXiv:1910.01108.
- [7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: pre-trained biomedical language representation model for biomedical text mining, *Bioinformatics* 36 (2019) 1234–1240. doi:10.1093/bioinformatics/btz682.
- [8] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Hong Kong, 2019, pp. 3982–3992. URL: <https://www.aclweb.org/anthology/D19-1410/>.