

# Large Biomedical Question Answering Models with ALBERT and ELECTRA

Sultan Alrowili<sup>1</sup>, K. Vijay-Shanker<sup>1</sup>

<sup>1</sup>Department of Computer and Information Science, University of Delaware, Newark, Delaware, USA

## Abstract

The majority of systems that participated in the BioASQ8 challenge are based on BioBERT model [1]. We adopt a different approach in our participation in the BioASQ9B challenge by taking advantage of large biomedical language models that are built on ELECTRA [2] and ALBERT [3] architectures, including both BioM-ELECTRA and BioM-ALBERT [4]. Moreover, we examine the advantage of transferability [5] between BioASQ and other text classification tasks such as The Multi-Genre Natural Language Inference (MultiNLI) [6]. Our results show that both BioM-ELECTRA and BioM-ALBERT significantly outperform the BioBERT model on the BioASQ9B task.

## Keywords

BERT, ELECTRA, ALBERT, BioASQ

## 1. Introduction

BioBERT model [7] represents the early success of domain adaptation of BERT [8] model in the biomedical domain. BioBERT model shows impressive results on the BioASQ7B challenge by taking the lead on most five batches of BioASQ7B challenge [9]. Furthermore, the BioBERT model is used in the majority of biomedical models that competed in the BioASQ8 challenge [1]. However, since the introduction of BERT model in 2018, new Transformer-based models have been introduced to NLP community including RoBERTa [10], ELECTRA [2], XLNET [11], MegaTron-LM [12], and ALBERT [3]. An adaptation of some of these models to the biomedical domain have been introduced later as BioRoBERTa [13], BioMegaTron [14] and PubMedBERT [15]. Additionally, we have introduced both BioM-ELECTRA and BioM-ALBERT models [4]. Both models are large-scale models that are adapted to the biomedical domain by pretraining both on Pubmed abstracts.

As noted earlier, a majority of participant systems in the BioASQ8B challenge are based on the BioBERT base-scale model. This motivates us to examine the effectiveness of large-scale biomedical models. The main findings of our investigations are that:

- (i) Both BioM-ALBERT and BioM-ELECTRA, models that we have recently developed are effective in addressing both BioASQ factoid and list questions.
- (ii) Treating BioASQ yes/no question as a classification problem is an effective approach that can lead to competitive performance.


---

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

✉ alrowili@udel.edu (S. Alrowili); vijay@udel.edu (K. Vijay-Shanker)



© 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)

## 2. System Description

We use large-scale biomedical language models, which is one of the primary differences between our system and other prior systems that participated in the BioASQ8B challenge. In our participation in the BioASQ9B challenge, we use both our models BioM-ELECTRA and BioM-ALBERT models [4].

### 2.1. BioM-ELECTRA

ELECTRA is a model that was built upon the idea of Transformer encoder, and attention mechanism [16] that BERT model uses. However, the ELECTRA model introduces novelty to the loss function by eliminating Next Sentence Prediction (NSP) objective, which is a similar decision taken by the RoBERTA model [10]. Moreover, ELECTRA improves the loss function by incorporating ideas from GAN model [17] where it generates corrupted (fake) tokens by employing a small Masked Language Model (MLM). Then, a discriminator model will judge those corrupted tokens and decide if they are "original" or "replaced" tokens.

To shift the contextual representation of ELECTRA, we pretrain ELECTRA on PubMed abstracts using specific domain vocabulary learned from PubMed abstracts. We pretrain our BioM-ELECTRA for 434K steps using TPUv3-512 units with a batch size of 4096.

### 2.2. BioM-ALBERT

ALBERT model [3] takes a similar decision to ELECTRA regarding the loss function by dropping Next Sentence Prediction (NSP) function. Furthermore, ALBERT introduces a self-supervised loss for sentence-order prediction (SOP) objective. Additionally, the ALBERT model improves the efficiency of the Transformer model by introducing both parameter-sharing and factorization of embedding layers techniques. The Parameter-sharing technique improves the architecture by reducing the parameters redundancy inside the model.

On the other hand, factorization of embedding layers allows the model to increase its hidden layer size up to 4096 while having only 235M parameters in the case of ALBERT-xxlarge. We build BioM-ALBERTxxlarge by pretraining ALBERTxxlarge on PubMed Abstracts using TPUv3-512 unit for 264K steps and a batch size of 8192. Similar to BioM-ELECTRA, we also pretrain BioM-ALBERT on PubMed abstracts only.

Table 1 shows the architecture design and the reported results [4] of our models on SQuAD2.0 [18] and BioASQ7B-Factoid tasks against other SOTA models. We include this table to show a head-to-head comparison between different architectures that have been used by participants' systems in the BioASQ9B challenge [19]. We should also note that it is a common practice in the literature to fine-tune the biomedical language model on the SQuAD dataset first and then on the BioASQ dataset. The reason to follow this approach because SQuAD2.0 dataset has more than 130K examples, which is much larger than the BioASQ dataset.

**Table 1**

Results of BioM-ALBERT and BioM-ELECTRA on BioASQ7B-Factoid Task. Evaluation metrics are F1 score for SQuAD task and Mean Reciprocal Rank MRR for BioASQ task. We use reported results of BioBERT-Base, BioBERT-Large, and BioMegaTron [14]; PubMedBERT, BioM-ELECTRA and BioM-ALBERT [4].

Model	#Parameters	#Hidden Size	SQuAD2.0	BioASQ7B-Factoid
BioBERT-Base	110M	768	-	41.1
BioM-ELECTRA-Base	110M	768	84.4	52.3
PubMedBERT-PMC-base	110M	768	80.9	51.9
BioBERT-Large	335M	1024	-	50.1
BioMegaTron345m	345M	1024	84.2	52.5
BioM-ELECTRA-Large	335M	1024	88.3	54.1
BioM-ALBERT-xxLarge	235M	4096	87.0	56.9

### 3. Experimental Setup

#### 3.1. Pre-Processing phase

For BioASQ9B factoid and list questions, we converted all questions to SQuADv1.1 format. Therefore, we duplicate the snippet (context) for each question in the training and test dataset instead of having a group of snippets and one corresponding question. For yes/no questions, we adopted a binary classification approach to solve this task by having the context (snippet) as "sentence 1", questions as "sentence 2" and the answer (yes/no) as a "label." We use a pre-processing script developed by [15] to generate the BioASQ classification dataset.

#### 3.2. Environmental Design

We fine-tune our models on factoid and list questions using Google Cloud Compute Engine with TPUv3-8 units and TensorFlow 1.15. For the yes/no task, we use the Hugging-face Transformers library [20] and V100 GPU on the Google Colab Pro environment.

#### 3.3. Hyperparameters

For factoid and list questions, we use the same hyperparameters settings that we use in our previous work [4] as shown in Table 2. We made this decision to examine the consistency and reproducibility of both BioM-ELECTRA and BioM-ALBERT on the BioASQ9B challenge. For the yes/no question, we use the training and testing dataset of the BioASQ8B challenge to determine our choices of hyperparameters.

#### 3.4. Task-to-Task Transfer Learning

The early work done by [5] and [21] shows that the transferability (Task-to-Task Transfer Learning) between general domain tasks such as MultiNLI [6] and SQuAD helps to improve the results on SQuAD and BioASQ8B tasks. We did a similar approach by fine-tuning both BioM-ALBERT and BioM-ELECTRA on the MNLI task, then SQuAD, and later on the BioASQ

**Table 2**

Details of fine-tuning hyperparameters that we use for both BioM-ALBERT and BioM-ELECTRA. (MSL=Max Seq. Length)

Task	Model	Learning Rate	Batch	Epochs	MSL
Factoid/List	BioM-ELECTRA	2e-5	24	4	512
Factoid/List	BioM-ALBERT	1e-5	128	3	384
Yes/No	All our models	3e-5	8	5	256

training dataset. We investigate and report the impact of this transferability on BioASQ9B in the result section.

## 4. Results and Discussion

We participated in the BioASQ9B challenge under the name "UDEL-LAB". Our reported results in this section are obtained from the BioASQ9B official leader board. We participate in the BioASQ9B-Factoid challenge starting from batch 3, and we use batch 2 to test the format of our submission. Therefore, we only include results of BioASQ-Factoid challenge starting from batch 3. We participated in yes/no, and list questions on batch five only since both types of tasks require extra pre-processing that we could not develop at early stage.

### 4.1. Factoid Task

Table 3 shows the results of our system on the BiASQ9B-Factoid challenge. We show only the top five systems for each batch based on the mean reciprocal rank (MRR) score. The Fudan University team participated with four systems under the name of ir\_sys [19]. Their systems combined SpanBERT [22], PubMedBERT [15] and XLNet models [11]. On other hand, "bio-answerfinder" system uses the BioELECTRA model [23], which they have developed early based on ELECTRA architecture. The result of BioM-ALBERT and BioM-ELECTRA against other models on both batch three and batch five suggests that our models has more consistency on the BioASQ performance than other models. Results also highlight that language model scale is a dominant factor on the performance of BioASQ-Factoid questions. Only large-scale models that are based on ALBERT-xxlarge, ELECTRA-large, and XLNET are taking the lead in all three batches.

On the other hand, using the transferability between MNLI and SQuAD tasks improves the score of our systems in the third batch by almost 2% in MRR score. However, this improvement is not consistent in both batch 4 and 5. We attribute this inconsistency to the fact that the fine-tuning layer of BERT-like models is randomly initialized. This randomness causes a fluctuation in the results, especially if we have small evaluation data set [15]. On the other hand, the score of BioM-ALBERT and BioM-ELECTRA in both batches 3 and 5 suggest that having an ensemble model could help further improve the results.

**Table 3**

Results of BioM-ALBERT and BioM-ELECTRA on BioASQ9B-Factoid Task. Strict Acc. is based on the evaluation of the first predicted answer by the system. Lenient Acc. is based on whether the system returns the exact answer in the top five predicted answers.

Batch	Model	Strict Acc.	Lenient Acc.	MRR
9B Batch 3	BioM-ALBERTxxlarge+MNLI+SQuAD+BioASQ	0.5405	<b>0.7027</b>	<b>0.6149</b>
	lr_sys2	<b>0.5946</b>	0.6486	0.6135
	BioM-ALBERTxxlarge+SQuAD+BioASQ	0.5405	0.6757	0.5946
	BioM-ELECTRA-large+SQuAD+BioASQ	0.5135	<b>0.7027</b>	0.5923
	bio-answerfinder	0.5676	0.5946	0.5811
9B Batch 4	lr_sys1	<b>0.6429</b>	<b>0.7857</b>	<b>0.6929</b>
	lr_sys2	0.6071	0.7500	0.6464
	BioM-ELECTRA-large+SQuAD+BioASQ	0.5357	<b>0.7857</b>	0.6351
	BioM-ELECTRA-large+MNLI+SQuAD+BioASQ	0.5000	<b>0.7857</b>	0.6321
	BioM-ALBERTxxlarge+SQuAD+BioASQ	0.5357	0.7143	0.5982
9B Batch 5	BioM-ELECTRA-large+SQuAD+BioASQ	<b>0.5000</b>	<b>0.7222</b>	<b>0.5880</b>
	BioM-ELECTRA-large+MNLI+SQuAD+BioASQ	0.4722	0.6944	0.5694
	finetuning1	<b>0.5000</b>	0.6667	0.5671
	BioM-ALBERTxxlarge+SQuAD+BioASQ	0.4444	<b>0.7222</b>	0.5588
	BioM-ALBERTxxlarge+MNLI+SQuAD+BioASQ	0.4722	0.6667	0.5556

**Table 4**

Results of BioM-ALBERT and BioM-ELECTRA on BioASQ9B challenge for a list and yes/no questions. Official evaluation metrics for Yes/No task is Macro-F1 score, and for List questions is F-Measure. We only participated in batch five for these type of questions.

Task	Model	#Rank	Score
List	BioM-ALBERTxxlarge+MNLI+SQuAD+BioASQ	#1	<b>0.5175</b>
	BioM-ALBERTxxlarge+SQuAD+BioASQ	#2	0.4927
	lr_sys2	#3	0.4804
	BioM-ELECTRA-large+SQuAD+BioASQ	#7	0.4031
	BioM-ELECTRA-large+MNLI+BioASQ	#8	0.3936
Yes/No	KU-DMIS-2	#1	<b>0.8246</b>
	BioM-ALBERTxxlarge+SQuAD+BioASQ	#4	0.7564
	BioM-ELECTRA-large+SQuAD+BioASQ	#5	0.6801

## 4.2. List and Yes/No Tasks

Table 4 shows the results of our system on the BiASQ9B List and Yes/No challenge. In the list task, our systems ranked in first and second place. We achieved this score for list questions despite using the same hyperparameters that we use for the factoid task. On yes/no task, BioM-ALBERT performs significantly better than BioM-ELECTRA but falls behind the performance of "KU-DMIS-2" system, which uses BioBERT-Large [19]. We should also note that the number of both list questions (18) and yes/no (19) questions are relatively smaller than factoid questions (36). Tasks with small data sets usually are sensitive to hyperparameter choice and fluctuate

between each fine-tuning run, especially in the case of binary classification (yes/no) task.

## 5. Conclusion and Future Work

We demonstrate that BioM-ELECTRA and BioM-ALBERT models are effective in addressing the BioASQ challenge. Our systems take the lead in two batches of factoid tasks and by a significant margin (2%) in batch 5. Additionally, we show that applying transferability between MNLI and SQuAD led our systems to score at first place on factoid (batch 3) and list (batch 5) questions. For future work, we plan to build a large ensemble QA system based on both BioM-ELECTRA and BioM-ALBERT to address the BioASQ and pandemic challenges.

## Acknowledgement

We would like to acknowledge the support we have from Tensorflow Research Cloud (TFRC) team to grant us access to TPUv3 units.

## References

- [1] A. Nentidis, A. Krithara, K. Bougiatiotis, M. Krallinger, C. Rodriguez-Penagos, M. Villegas, G. Paliouras, Overview of bioasq 2020: The eighth bioasq challenge on large-scale biomedical semantic indexing and question answering, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, Springer, 2020. URL: [https://link.springer.com/chapter/10.1007/978-3-030-58219-7\\_16](https://link.springer.com/chapter/10.1007/978-3-030-58219-7_16).
- [2] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, 2020. [arXiv:2003.10555](https://arxiv.org/abs/2003.10555).
- [3] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, R. Soricut, Albert: A lite bert for self-supervised learning of language representations, 2020. [arXiv:1909.11942](https://arxiv.org/abs/1909.11942).
- [4] S. Alrowili, V. Shanker, BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA, in: Proceedings of the 20th Workshop on Biomedical Language Processing, Association for Computational Linguistics, Online, 2021, pp. 221–227. URL: <https://www.aclweb.org/anthology/2021.bionlp-1.24>. doi:10.18653/v1/2021.bionlp-1.24.
- [5] M. Jeong, M. Sung, G. Kim, D. Kim, W. Yoon, J. Yoo, J. Kang, Transferability of natural language inference to biomedical question answering, 2021. [arXiv:2007.00217](https://arxiv.org/abs/2007.00217).
- [6] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 353–355. URL: <https://www.aclweb.org/anthology/W18-5446>. doi:10.18653/v1/W18-5446.
- [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* (2019). URL: <https://doi.org/10.1093/bioinformatics/btz682>. doi:10.1093/bioinformatics/btz682.



- [8] 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. URL: <https://www.aclweb.org/anthology/N19-1423>. doi:10.18653/v1/N19-1423.
- [9] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, Results of the seventh edition of the bioasq challenge, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Springer, 2019. URL: <https://arxiv.org/pdf/2006.09174.pdf>.
- [10] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.
- [11] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, 2020. arXiv:1906.08237.
- [12] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, B. Catanzaro, Megatron-lm: Training multi-billion parameter language models using model parallelism, 2020. arXiv:1909.08053.
- [13] P. Lewis, M. Ott, J. Du, V. Stoyanov, Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art, in: Proceedings of the 3rd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Online, 2020, pp. 146–157. URL: <https://www.aclweb.org/anthology/2020.clinicalnlp-1.17>. doi:10.18653/v1/2020.clinicalnlp-1.17.
- [14] H.-C. Shin, Y. Zhang, E. Bakhturina, R. Puri, M. Patwary, M. Shoeybi, R. Mani, BioMegatron: Larger biomedical domain language model, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4700–4706. URL: <https://www.aclweb.org/anthology/2020.emnlp-main.379>. doi:10.18653/v1/2020.emnlp-main.379.
- [15] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, 2021. arXiv:2007.15779.
- [16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, u. Kaiser, I. Polosukhin, Attention is all you need, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17, Curran Associates Inc., Red Hook, NY, USA, 2017, p. 6000–6010.
- [17] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, volume 27, Curran Associates, Inc., 2014. URL: <https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf>.
- [18] P. Rajpurkar, R. Jia, P. Liang, Know what you don’t know: Unanswerable questions for SQuAD, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 784–789. URL: <https://www.aclweb.org/anthology/P18-2124>. doi:10.18653/v1/P18-2124.

- [19] A. Nentidis, G. Katsimpras, E. Vantorou, A. Krithara, L. Gasco, M. Krallinger, G. Paliouras, Overview of bioasq 2021: The ninth bioasq challenge on large-scale biomedical semantic indexing and question answering, 2021. [arXiv:2106.14885](https://arxiv.org/abs/2106.14885).
- [20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>. doi:10.18653/v1/2020.emnlp-demos.6.
- [21] H. Zhang, H. Zhao, C. Liu, D. Yu, Task-to-task transfer learning with parameter-efficient adapter, in: X. Zhu, M. Zhang, Y. Hong, R. He (Eds.), Natural Language Processing and Chinese Computing, Springer International Publishing, Cham, 2020, pp. 391–402.
- [22] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, O. Levy, SpanBERT: Improving pre-training by representing and predicting spans, arXiv preprint arXiv:1907.10529 (2019).
- [23] I. B. Ozyurt, On the effectiveness of small, discriminatively pre-trained language representation models for biomedical text mining, in: Proceedings of the First Workshop on Scholarly Document Processing, Association for Computational Linguistics, Online, 2020, pp. 104–112. URL: <https://aclanthology.org/2020.sdp-1.12>. doi:10.18653/v1/2020.sdp-1.12.