Biomedical Question Answering with Transformer Ensembles

Raghav R[1,*,†], Jason Rauchwerk[1,†], Parth Rajwade[1,†], Tanay Gummadi[1,†], Eric Nyberg[1] and Teruko Mitamura[1]

[1]*Language Technologies Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA*

Abstract

Recent advancements in natural language processing, specifically transformers, have shown great promise in improving the performance of question-answering systems. However, we observe that a single transformer model may not achieve sufficient accuracy and reliability to meet the stringent requirements of biomedical question answering. Based on our participation in the BioASQ Challenge, we present a comprehensive approach for biomedical question answering using transformers, integrating an end-to-end data processing pipeline with the UMLS Metamap and different ensembling techniques. Our findings suggest that transformer ensembles achieve significant performance improvements when compared to individual models.

Keywords

biomedical question answering, transformer models, ensemble learning

1. Introduction

The rapid growth of on-line biomedical text has stimulated the research and development of robust, specialized language models that provide reliable, high-accuracy responses to queries posed against the medical literature. For example, the PubMed database contains more than 35 million citations and abstracts of biomedical articles[1]. To overcome the challenge of inadequate contextual representation when matching queries in the biomedical domain, researchers have turned to transformer-based language models such as BERT [1], which have demonstrated remarkable efficacy in capturing contextual information from large corpora. Various adaptations of BERT, namely Med-BERT [2], SciBERT [3], and Clinical-BERT [4] have been specifically designed to address the need for context-aware biomedical language models.

In this paper, we explore the hypothesis that an ensemble of transformer models can perform better than a single transformer alone for specific bioinformatic tasks. We tested our hypothesis by participating in the eleventh edition of the BioASQ Challenge [5], specifically focusing on Phase B of Task 11b [6]. Our primary objective is to deliver "exact" answers for yes/no, factoid, and list question types. The BioASQ Challenge consists of four rounds of test sets, providing

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

†These authors contributed equally.

rraghavr@cs.cmu.edu (R. R); jrauchwe@cs.cmu.edu (J. Rauchwerk); prajwade@cs.cmu.edu (P. Rajwade); tgummadi@cs.cmu.edu (T. Gummadi); ehn@cs.cmu.edu (E. Nyberg); teruko@andrew.cmu.edu (T. Mitamura)

© 2023 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]https://pubmed.ncbi.nlm.nih.gov/about/

CEUR-WS.org/Vol-3497/paper-014.pdf

participants with the opportunity to submit up to five systems for each test set. The organizers provide the dataset for Task 11b [7] in the form of a single training set and four test sets for each evaluation round. We submitted a total of four systems across three of the test sets. This allowed us to explore various approaches and methodologies, enhancing our understanding of the problem space.

After analyzing the performance of different systems in previous editions of the BioASQ Challenge, we decided to ensemble BioBERT [8] and BioM-Electra [9] for factoid and list questions. For yes/no questions, we employ BioM-Electra [10]. We use the Unified Medical Language System (UMLS) MetaMap tool [2] for preprocessing data, and for synonym removal during post-processing.

2. Related Work

There has been significant prior work done for question-answering in the biomedical domain. Following the advent of BERT [1], Lee et al. [8] introduced BioBERT, a language modeling approach that initializes a BERT model (pretrained on Wikipedia and BookCorpus) and continues pretraining using masked-language modeling (MLM) and next sentence prediction (NSP) on PubMed abstracts and PubMed Central (PMC) full-text articles. BlueBERT [11] follows a similar approach but finds performance improvements, in the clinical domain, by pre-training on PubMed abstracts and MIMIC-III clinical notes. However, Gu et al. [12] find the aforementioned mixed-domain pretraining objective to be inferior to domain-specific pretraining from scratch given the difference in vocabulary from the initial BERT model and the later biomedical context.

PubMedBERT [12] is a new BERT model, trained from scratch using PubMed abstracts, that outperforms BioBERT and BlueBERT; the authors attribute the performance improvements to having an in-domain vocabulary which the architecture can model completely in order to fully optimize for in-domain data. Jeong et al. [13] propose a sequential transfer learning method for fine-tuning biomedical models on intermediate datasets, before fine-tuning on the specific biomedical task; this helps to improve performance due to data scarcity for the final task. Specifically, the authors show a significant F1 gain by training on MNLI [14] and SQuAD [15] before BioASQ, and unifying context-length distributions between fine-tuning tasks. Ting et al. [16] present a method using BioBERT to generate snippets for ideal answers (BioASQ Task B), and then using these snippets to predict exact answers for factoid/list questions.

BioM-ELECTRA and BioM-ALBERT [10] are variants of ELECTRA and ALBERT pretrained on PubMed abstracts. They subsequently fine-tune on a combination of SQuAD and MNLI, and finally the BioASQ dataset [17], achieving SOTA on BioASQ 10b for list questions [18]. BioLinkBERT [19] takes an alternative approach in adding an additional pretraining objective of document relation prediction, in order to learn contextual linked concepts between documents in the form of PubMed hyperlinks. Given limited resources, we selected BioBERT, the most commonly cited baseline model for bioinformatics [8], and BioM-ELECTRA, the best-performing model on the most recent BioASQ challenge [18] as the two baselines for our work.

[2] https://www.nlm.nih.gov/research/umls/implementation_resources/metamap.html

3. Model Overview

3.1. List and Factoid Questions

Following the methodology of Alrowili and Vijay-Shanker [9], we merge factoid and list questions to overcome the limited number of training examples. We split the lists into multiple factoid questions and search the golden snippets for the spans that contain an exact string match for the answer. Because there are multiple snippets for each question, this creates many snippet-answer pairs for each original question. Each of these pairs are rewritten into the SQuAD format to be fed into our models. We use BioBERT and BioM-ELECTRA for these questions.

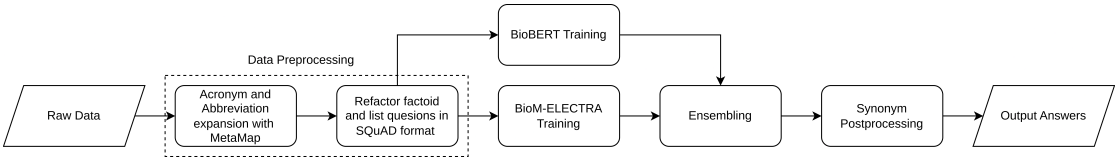

Figure 1: A flowchart of the entire pipeline

3.2. Yes/No Questions

We treated yes/no questions as a binary classification problem. We concatenate all of the golden snippets to create a paragraph and feed this context and the question to the model. We did not attempt to answer yes/no questions in our first batch, but submitted a model for the second and third batches to better compare our performance with submissions from previous years. We started with DistilBERT [20], BioBERT, and BioM-ELECTRA for these questions. However, we chose BioM-ELECTRA for our systems because of its superior performance.

4. Methodology

4.1. Dataset Preprocessing

Snippets that come from different articles may use different names or acronyms to refer to the same concept. For instance, the protein "transforming growth factor alpha" is variously referred to as "transforming growth factor alpha", "transforming growth factor", "TGFα", and "TGF-α". We use to the MetaMap tool to ensure that all answers in the snippets are properly identified. MetaMap queries the UMLS Metathesaurus (curated by the National Library of Medicine[3]) to determine the canonical form for each biomedical term. We run snippets through MetaMap to expand all acronyms and abbreviations before finding the answer spans.

[3] https://www.nlm.nih.gov/

4.2. Synonym Postprocessing

We also utilize UMLS MetaMap to remove synonyms from the model's predicted candidate answer list. In the Metathesaurus, each entity has a Concept Unique Identifier (ConceptUI), which is shared among all names that can refer to the same entity. Our system sorts candidate answers by their confidence scores and greedily constructs an answer set while making sure that all final answers have unique ConceptUIs.

4.3. Ensembling

We believe that ensembling models (which was also proposed by Alrowili and Vijay-Shanker [9] as a future prospect), specifically BioBERT and BioM-ELECTRA, can combine their strengths and form a better system. We performed a grid search to discover the weighting schemes that maximize F1 score (for list questions) and MRR (for factoid questions). Our ensembling weights were (0.004, 0.996) when maximizing F1 and (0.037, 0.963) when maximizing MRR for BioBERT and BioM-ELECTRA, respectively. We use these weights to compute a linear combination of weights and confidence scores for each predicted answer. We rerank and filter the candidate answers to return our ensembled predictions.

5. BioASQ Task 11b Systems

We performed an 80-20 split on the training set for our validation purposes (internal testing). The results are shown in Table 1.

Table 1
Results on validation dataset (internal testing)

System	Yes/No		Factoid	List
	F1	Acc.	MRR	F1
Distilbert	0.8657	0.7708	-	-
BioBERT	0.8790	0.8063	0.6488	0.4914
BioM-ELECTRA	0.9430	0.9130	0.6862	0.5504
Ensemble 1 (max. F1)	-	-	0.6910	0.5374
Ensemble 2 (max. MRR)	-	-	0.6969	0.5374

We participated in the BioASQ Challenge under the team name 'AsqAway', submitting 4 systems to the task. Our systems are described in Table 2. The BioASQ test performance of our systems is described in Tables 3, 4, and 5. The evaluation datasets are small and there is high variance in model performance across the batches, making it difficult to compare them directly.

6. Discussion

One of our initial observations was that the models returned a large number of probable answers, despite having set probability thresholds. While this meant that all possible answers were being covered in most of the cases, there were a large number of false positives which led to a drop in

Table 2
Our BioASQ Task 11b Systems

System	Yes/No Model	List and Factoid Model
AsqAway_1	BioM-ELECTRA	BioBERT
AsqAway_2	BioM-ELECTRA	BioM-ELECTRA
AsqAway_3	BioM-ELECTRA	Ensemble 1 (max. F1)
AsqAway_4	BioM-ELECTRA	Ensemble 2 (max. MRR)

Table 3
Results from BioASQ Task11B Batch 1

System	Yes/No		Factoid	List
	F1	Acc.	MRR	F1
BioBERT	-	-	0.3158	0.4595
BioM-ELECTRA	-	-	0.3947	0.4804
Ensemble 1 (max. F1)	-	-	0.3947	0.5106
Ensemble 2 (max. MRR)	-	-	0.4211	0.5106

Table 4
Results from BioASQ Task11B Batch 2

System	Yes/No		Factoid	List
	F1	Acc.	MRR	F1
BioBERT	-	-	0.4545	0.1756
BioM-ELECTRA	0.8693	0.8750	0.4545	0.2327
Ensemble 1 (max. F1)	-	-	0.4773	0.2329
Ensemble 2 (max. MRR)	-	-	0.4773	0.2329

Table 5
Results from BioASQ Task11B Batch 3

System	Yes/No		Factoid	List
	F1	Acc.	MRR	F1
BioBERT	-	-	0.3154	0.4290
BioM-ELECTRA	0.9091	0.8750	0.4423	0.4813
Ensemble 1 (max. F1)	-	-	0.4615	0.4431
Ensemble 2 (max. MRR)	-	-	0.4615	0.4431

precision. Upon further observations, we noticed that both acronyms and their expanded forms were included in the training data. This formed the motivation behind the UMLS Preprocessing step as described in Section 4.1.

Another observation was that a lot of the answers returned by our models were synonyms of each other. Since the challenge requires the systems to remove synonyms in the candidate answers, we performed the Synonym Postprocessing step as described in Section 4.2. We present a quantitative analysis of our results, based on UMLS Preprocessing and Synonym Postprocessing in Tables 6 and 7 respectively. Figures 2 and 3 show a qualitative example of the same. The number of answers returned is reduced significantly while maintaining accuracy.

Table 6
Model Performance with and without UMLS Preprocessing

System	With UMLS Preprocessing		Without UMLS Preprocessing	
	Factoid	List	Factoid	List
BioBERT	0.6488	0.4914	0.5968	0.4435
BioM-ELECTRA	0.6862	0.5504	0.6026	0.4670

Table 7
Model Performance with and without Synonym Postprocessing

System	With Synonym Postprocessing		Without Synonym Postprocessing	
	Factoid	List	Factoid	List
BioBERT	0.6488	0.4914	0.5866	0.4430
BioM-ELECTRA	0.6862	0.5504	0.6254	0.4808


```
"5518e7da622b194345000004": {
    "predicted_ans": [
        "c-Jun NH2-terminal kinase (Jnk)",
        "JNK)",
        "NH2-terminal kinase",
        "c-Jun",
        "Mitogen-Activated Protein kinases",
        "c",
        "c",
        "c",
        "NK)"
    ],
    "base_answers": [
        "c-Jun NH2-terminal kinase",
        "JNK"
    ],
    "type": "factoid"
},
```


Figure 2: With UMLS


```
"5518e7da622b194345000004": {
    "predicted_ans": [
        "c-Jun NH2-terminal kinase",
        "NH2-terminal kinase",
        "c-Jun NH2-terminal kinase (JNK",
        "NH2-terminal kinase (JNK",
        "c-Jun NH2-terminal kinase (JNK)",
        "Jun NH2-terminal kinase",
        "JNK",
        "H2-terminal kinase",
        "terminal kinase",
        "NH2-terminal kinase (JNK)",
        .
        .
        .
    ],
    "base_answers": [
        "c-Jun NH2-terminal kinase",
        "JNK"
    ],
    "type": "factoid"
}
```


Figure 3: Without UMLS

7. Future Work

We anticipate significant opportunities for improvement both on the data and modeling side. As the model is currently only pre-trained on full-text articles and abstracts, there are instances where our architecture chooses an adjacent but incorrect medical term (low precision) or returns a nonsensical answer given the sparseness of the correct term in the full-text (low recall). The former case is more prevalent among our experiments than the latter and we hypothesize adding titles to the pre-training procedure can improve both precision and recall; precision is improved as the title acts effectively as a distillation for the context and recall is improved in the form of providing additional context for the model to train upon. Similarly, in the training procedure,

we can leverage a combination of the abstract from the provided documents rather than solely relying on snippets as context.

Moradi and Samwald [21] find vulnerabilities in BioBERT when exposed to word-level and character-level noise; we corroborate this observation in instances where training data has key medical terms misspelled or misused. Adversarial training offers robustness to such errors: Jia and Liang present the "AddSent" model-independent procedure from [22] and Du et al. [23] finds performance gains in the context of BioASQ. However, the alternative "AddAny" procedure by Jia and Liang [22] can also be implemented for more rigorous examples.

Using larger models (e.g. Large, X-Large, XX-Large variants) for BioM-ELECTRA and BioM-ALBERT empirically leads to incremental performance gains [10], however don't address a strata of error. Alternative models such as LinkBERT [19] show similar performance as the Bio-M variants from preliminary experimentation. Changing the order of finetuning procedures mentioned by Jeong et al. [13] in swapping the ordering of MNLI and SQuAD prior to tuning for the BioASQ data has potential for marginal gains.

We hypothesize the use of adversarial methods to have the most promise in delivering performance improvements, followed by the use of alternative architectures.

8. Conclusion

In this paper, we address the challenge of generating accurate information retrieval systems for biomedical information, specifically focusing on Phase B of Task 11b in the eleventh BioASQ Challenge. Given the complex and sensitive nature of biomedical data, we adopt a novel approach that involves ensembling state-of-the-art transformer models that have previously performed well in BioASQ challenges, along with implementing data processing techniques based on UMLS MetaMap. Our efforts aim to contribute towards the development of highly precise answers for the list and factoid question types. Our approach yields promise for data-oriented techniques towards improving performance on the task. Our code files are publicly available in a GitHub repository[4].

References

[1] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[2] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction, NPJ digital medicine 4 (2021) 86.

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

[4] B. Yan, M. Pei, Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 2982–2990.

[4] https://github.com/parthsr5/asqaway

[5] A. Nentidis, G. Katsimpras, A. Krithara, S. Lima-López, E. Farré-Maduell, L. Gasco, M. Krallinger, G. Paliouras, Overview of BioASQ 2023: The eleventh BioASQ challenge on Large-Scale Biomedical Semantic Indexing and Question Answering, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), 2023.
[6] A. Nentidis, G. Katsimpras, A. Krithara, G. Paliouras, Overview of BioASQ Tasks 11b and Synergy11 in CLEF2023, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, 2023.
[7] A. Krithara, A. Nentidis, K. Bougiatiotis, G. Paliouras, BioASQ-QA: A manually curated corpus for Biomedical Question Answering, Scientific Data 10 (2023) 170.
[8] 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.
[9] S. Alrowili, K. Vijay-Shanker, Exploring biomedical question answering with biom-transformers at bioasq10b challenge: Findings and techniques, in: Conference and Labs of the Evaluation Forum, 2022.
[10] S. Alrowili, K. Vijay-Shanker, Biom-transformers: building large biomedical language models with bert, albert and electra, in: Proceedings of the 20th Workshop on Biomedical Language Processing, 2021, pp. 221–227.
[11] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets, arXiv preprint arXiv:1906.05474 (2019).
[12] 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, ACM Transactions on Computing for Healthcare (HEALTH) 3 (2021) 1–23.
[13] 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).
[14] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, arXiv preprint arXiv:1704.05426 (2017).
[15] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, Squad: 100,000+ questions for machine comprehension of text, arXiv preprint arXiv:1606.05250 (2016).
[16] H.-H. Ting, Y. Zhang, J.-C. Han, R. T.-H. Tsai, Ncu-iisr/as-gis: Using bertscore and snippet score to improve the performance of pretrained language model in bioasq 10b phase b (2022).
[17] S. Alrowili, V. Shanker, Large biomedical question answering models with albert and electra., in: CLEF (Working Notes), 2021, pp. 213–220.
[18] S. Alrowili, K. Vijay-Shanker, Exploring biomedical question answering with biom-transformers at bioasq10b challenge: Findings and techniques (2022).
[19] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, arXiv preprint arXiv:2203.15827 (2022).
[20] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. arXiv:1910.01108.
[21] M. Moradi, M. Samwald, Improving the robustness and accuracy of biomedical language

models through adversarial training, Journal of Biomedical Informatics 132 (2022) 104114.
[22] R. Jia, P. Liang, Adversarial examples for evaluating reading comprehension systems, arXiv preprint arXiv:1707.07328 (2017).
[23] Y. Du, J. Yan, Y. Lu, Y. Zhao, X. Jin, Improving biomedical question answering by data augmentation and model weighting, IEEE/ACM Transactions on Computational Biology and Bioinformatics (2022).