

End-to-end Biomedical Question Answering via Bio-AnswerFinder and Discriminative Language Representation Models

Ibrahim Burak Ozyurt¹

¹FDI Lab, Department of Neurosciences, University of California at San Diego, La Jolla, CA USA

Abstract

Generative Transformers based language representation models such as BERT and its biomedical domain adapted version BioBERT have been shown to be highly effective for biomedical question answering. Here, discriminative, sample-efficient biomedical language representation models based on ELECTRA language representation model architecture were introduced to enhance an end-to-end biomedical question answering system, Bio-AnswerFinder, for the BioASQ challenge. The introduced language representation models outperformed other language models including BioBERT in answer span classification, answer candidate re-ranking and yes/no answer classification tasks. The resulting end-to-end system participated in BioASQ Synergy and both phases of Task 9B with promising results.

Keywords

question answering, language representation models, biomedical information retrieval

1. Introduction

Transformers based language representation models such as BERT [1], XLNet [2] and ALBERT [3] are becoming increasingly popular for many downstream NLP tasks due to their ubiquitous performance advantages over previous methods. Domain adaptation of general language model BERT to the biomedical domain has shown significant performance improvements for downstream biomedical NLP tasks [4].

BERT, XLNet and ALBERT use a masked language modeling (MLM) approach by masking 15% of the training sentences and learning to guess the masked tokens in a generative manner resulting in only learning from 15% of the tokens per example. Recently a new pretraining approach named ELECTRA [5] is introduced for BERT transformer based encoder architecture, where a discriminative model is trained to detect whether each token in the corrupted input was replaced by a co-trained generator model sample or not. It is shown that ELECTRA is computationally more efficient than BERT and outperforms it given the same model size, data and computation resources [5]. The improvements over BERT by ELECTRA is most impressive at small model sizes and that effectiveness translates to biomedical domain for domain adapted small ELECTRA models [6].


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

✉ iozyurt@ucsd.edu (I. B. Ozyurt)

ORCID 0000-0003-3944-1893 (I. B. Ozyurt)



© 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)

The development and evaluation of a question answering system without an expert generated training/evaluation question answer data set is impossible. BioASQ, an EU-funded biomedical semantic indexing and question answering challenge [7, 8] yearly provides cumulative sets of biomedical question/gold standard answer data and evaluation platform for the advancement of biomedical question answering.

In this paper, enhancements to a sentence level end-to-end biomedical question answering system, Bio-AnswerFinder [9] to provide answers to all four types (factoid, list, yes/no and summary) of BioASQ challenge questions is introduced. To achieve this, three new biomedical domain adapted pretrained ELECTRA models are introduced. The introduced Bio-ELECTRA models are compared against many language representation models for question keyword selection, question answer span classification, answer candidate re-ranking, yes/no answer classification tasks showing superior performance. An abstractive summarization module based on the Transformers based text-to-text generation model T5 [10] is also introduced. The resulting system can answer any biomedical domain natural language questions and used in the the 9th BioASQ Challenge for the Synergy and 9B tasks.

The rest of the paper is organized as follows. After a brief overview of Bio-AnswerFinder and proposed enhancements, details of the pretraining of the ELECTRA based biomedical language models are provided. This is followed by the experiments on answer span classification for the BioASQ factoid/list questions, answer candidate re-ranking, search engine keyword selection and yes/no question answer determination. After the introduction of extractive and abstractive summarizer systems, details of the BioASQ Synergy and BioASQ 9B systems are provided. Following this, results of the challenge is discussed together with an error analysis on the BioASQ 8B ground truth data for factoid questions.

2. Overview of Bio-AnswerFinder

Bio-AnswerFinder [9] is a biomedical question answering system that takes a natural language question and returns a list of sentences from biomedical texts ranked in the order of confidence that they would answer the question. An overview of the system is shown in Figure 1. The original system is retrofitted with new modules to be able to provide answers for the four types of questions of the BioASQ Task B. The retrofitted modules together with the enhanced existing modules are shown in blue in Figure 1.

The modules of the Bio-AnswerFinder can be grouped logically into question processing, document processing and answer processing phases.

In the question processing phase, the natural language question is parsed, followed by the detection of the focus of the question. Afterwards, search keywords are selected from the words of the question using a supervised long short term memory (LSTM) [11] based keyword classifier. For BioASQ 9B, this module is replaced by a Bio-ELECTRA++ [6] based keyword tagger.

In the document processing phase, query relevant documents are retrieved from a traditional keyword based information retrieval system (Elasticsearch). Bio-AnswerFinder uses an iterative most specific to most generic keyword search guided by the keyword classifier selected keywords to retrieve a relevant set of documents from an Elasticsearch index. The order of keywords

dropped from iteration to iteration is learned from a set of annotated BioASQ 5B questions using a ranking classifier based on RankNet [12] with LSTM using attention [13].

In the answer processing phase, the question type (focus, definition question or other) is detected. Definition questions are handled by definition pattern based filtering of the sentences from the retrieved documents. For questions with a focus a detected entity type, the entity type is used for filtering out candidate sentences not having entities of the focus entity type. For both focus and other non-definition questions, the answer candidate sentences are ranked by a weighted version of the relaxed word mover's distance [14]. Afterwards, up to first 100 of these sentences are further re-ranked by a fine-tuned BERT [1] classifier.

For BioASQ 9, the BERT re-ranker is replaced by a better performing Bio-ELECTRA re-ranker. For factoid and list questions, a Bio-ELECTRA based question answer span classifier is used. For yes/no questions, two different Bio-ELECTRA based classification approaches are introduced. For summary questions, both extractive and abstractive summarization approaches are introduced. These approaches are explained in more detail in the following sections.

3. ELECTRA Based Biomedical Language Representation Models

For pretraining corpus both PubMed abstracts and PubMed Central (PMC) open access full-length papers were used. The main pretraining corpus was built using 21.2 million PubMed abstracts from the January 2021 baseline distribution. From the abstracts, title and abstract text sentences were extracted resulting in a corpus of 3.6 billion words. The second 12.3 billion words corpus was built using the sentences extracted from the sections of PMC open access papers excluding the references sections, which, unlike the other paper sections, do not have a regular sentence format. A domain specific word piece vocabulary was generated using SentencePiece byte-pair-encoding (BPE) model [15] from PubMed abstract texts. The Bio-ELECTRA Mid and Base models were pretrained for one million steps on the PubMed abstracts corpus followed by 200,000 steps training on the PMC open access papers corpus. The Bio-ELECTRA Mid Combined model was pretrained on the combination of abstract and full text paper corpus for 1.2 million steps.

Since the improvements over BERT and other transformers based language models by ELECTRA are pronounced at small model sizes [5], a mid sized model with a hidden layer size in the middle of the small and base ELECTRA architectures is introduced to investigate its competitiveness against the base model with a more than twice training parameters size. The Bio-ELECTRA model architectures are summarized in Table 1. All the mid and base sized Bio-ELECTRA models were pretrained on a single 8 core version 3 tensor processing unit (TPU) with 128 GB of RAM. The small Bio-ELECTRA++ model [6] was pretrained on a consumer grade 8 GB Nvidia RTX 2070 GPU. The hardware and pretraining times for all Bio-ELECTRA models were summarized in Table 2.

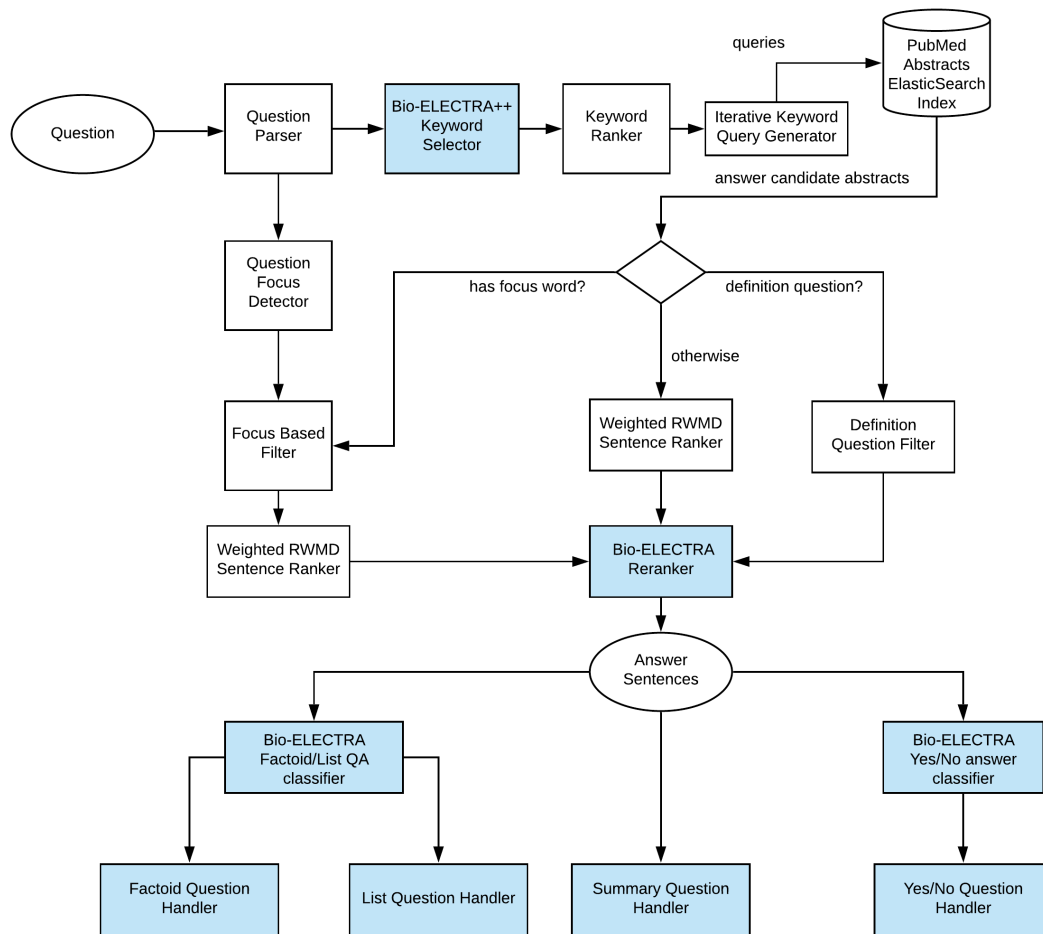


Figure 1: Overview of the Bio-AnswerFinder system

4. Experiments with Factoid/List Question Answer Span Detection

Since most factoid and list questions can be answered by a word or phrase (multiple word/phrases for list questions), the answers can be detected by learning to estimate scores for spans in the sequence of tokens of an answer candidate passage. To this end, the training/testing sets were generated from the factoid and list questions of the BioASQ 8b training data. From about 30% of the list and factoid questions which cannot be aligned to their exact answers, 152 more questions were recovered via manual inspection for synonyms and transliterations. The labeled data set is split into 85%/15% training/testing data sets of size 9557 and 1809, respectively. To increase performance over the smaller BioASQ data, the training set was combined with the

Table 1
ELECTRA Model Architectures for Biomedical Domain

Model	Params	Architecture
Bio-ELECTRA++ [6]	11M	hidden:256, layers:12, batch:64, attention heads:4
Bio-ELECTRA Mid	50M	hidden:512, layers:12, batch:256, attention heads:8
Bio-ELECTRA Base	110M	hidden:768, layers:12, batch:256, attention heads:12
Bio-ELECTRA Mid Combined	50M	hidden:512, layers:12, batch:256, attention heads:8

Table 2
Pretraining of ELECTRA Models for Biomedical Domain

Model	Params	Steps	Train Time/Hardware
Bio-ELECTRA++	11M	3.6M	48 days on RTX 2070 8GB GPU
Mid	50M	1.2M	6.5 days on 8 TPUv3s
Base	110M	1.2M	12.5 days on 8 TPUv3s
Mid Combined	50M	1.2M	6.5 days on 8 TPUv3s

out-of-domain SQuAD [16] v1.1 data set.

All together ten language representation models including BERT based biomedical domain specific BioBERT model were evaluated for factoid and list question answer span detection task. The performance of the models are evaluated by the standard SQUAD evaluation metrics, exact match and F_1 score. Ten randomly initialized answer span classifiers are fine-tuned for each language representation model. The experiment results are summarized in Table 3. All non-small Bio-ELECTRA had significantly outperformed BioBERT on this task. Given that mid sized Bio-ELECTRA models have less than half of the parameters of BioBERT, the results are very encouraging. The best performing Bio-ELECTRA Mid pretrained for 1.2 million steps was chosen to be used in the final system.

Snippets provided by BioASQ challenge were first passed through Bio-AnswerFinder bypassing the candidate document retrieval section. The re-ranked candidate sentences were then used as input for the factoid/list question classifier. The answer candidate word sequences were scored by a combination of their span classification probabilities, number of occurrence and rank of the sentence in which they have occurred first. The answer candidates were normalized and filtered to remove sub-phrases, singular/plural differences and acronyms. For list questions, answers candidates were enriched by coordinated phrase detection and processing. A classifier score threshold of 0.65 was selected to maximize F_1 performance on a holdout set of questions for selecting a subset of list question answer candidate span of words for the BioASQ challenge.

5. Experiments with Answer Candidate Re-ranking

In Bio-AnswerFinder, answer candidate sentences are first ranked by the inverse document frequency weighted relaxed word mover’s distance on PubMed abstract trained GloVe word and phrase embeddings. While this ranking usually results in decent results, supervised re-ranking improves performance as measured on blind, multiple curator tests [9]. By casting the ranking

Table 3
Biomedical Question Answering Test Results

Model	Exact Match	F_1
Bio-ELECTRA++	57.93 (0.66)	67.48 (0.44)
ELECTRA Small++	57.78 (0.64)	67.10 (0.55)
BERT	59.98 (0.66)	70.25 (0.48)
BioBERT	63.58 (0.66)	72.72 (0.48)
ELECTRA Base	65.01 (0.84)	72.82 (0.70)
Bio-ELECTRA Mid (1M)	68.71 (0.76)	75.52 (0.49)
Bio-ELECTRA Mid (1.2M)	69.50 (0.54)	75.82 (0.40)
Bio-ELECTRA Base (1M)	68.44 (0.56)	75.02 (0.60)
Bio-ELECTRA Base (1.2M)	68.44 (0.38)	75.50 (0.35)
Bio-ELECTRA Mid Combined (1.2M)	66.46 (0.65)	74.05 (0.44)

problem as a 0/1 loss classification problem, the learned probability estimates can be used to rank the candidate sentences by relevance.

For Bio-AnswerFinder, up to 100 answer candidates per question as returned by the weighted rWMD ranker were annotated as relevant or not (up to the first occurrence of a correct answer). The questions were selected from the BioASQ 5b training set. In total, 44933 sentences for 492 training questions and 9064 sentences for 100 testing questions were annotated.

Nine language representation models were tested. Due to highly unbalanced nature of the data set (on average one positive example per 99 negative examples), a weighted loss function where the errors for the positive examples weighted 99 times more than a negative example error was also used. The mean reciprocal rank (MRR) results averaged on ten randomly initialized runs using 14 language representation models (including weighted models) are summarized in Table 4. Based on the results, Bio-ELECTRA Mid (1M) was chosen for BioASQ 9 challenge since it had more stable score distribution than the Bio-ELECTRA++ model and twice as fast as the larger Bio-ELECTRA Base re-ranking models. While all Bio-ELECTRA models were significantly better than both BioBERT and BERT Base, the performance differences among the best performing Bio-ELECTRA models were not statistically significant.

6. Search Engine Keyword Selection via Bio-ELECTRA++

Selection of keywords is a vital step in the question answering step, since missing of even a single important keyword could prevent retrieval of relevant candidate documents. The original Bio-AnswerFinder had used a LSTM based multi class classifier using GloVe word embeddings trained on PubMed abstracts. To minimize out-of-vocabulary (OOV) effects on GloVe embeddings LSTM based model uses also inputs from part of speech tags of the question words encoded by a separate LSTM layer.

Encouraged by the performance of the discriminative language representation models, a Bio-ELECTRA++ model based approach was introduced. The keyword selection from a question

Table 4
Biomedical Question Answer Candidate Re-ranking Test Results

Model	MRR
ELECTRA Small++	0.281 (0.014)
ELECTRA Small++ (weighted)	0.281 (0.008)
Bio-ELECTRA++	0.335 (0.017)
Bio-ELECTRA++ (weighted)	0.332 (0.013)
BERT Base	0.246 (0.007)
BioBERT	0.283 (0.020)
ELECTRA Base	0.294 (0.017)
Bio-ELECTRA Mid (1M)	0.333 (0.017)
Bio-ELECTRA Mid (1M) (weighted)	0.336 (0.017)
Bio-ELECTRA Mid (1.2M)	0.316 (0.015)
Bio-ELECTRA Mid (1.2M) (weighted)	0.322 (0.015)
Bio-ELECTRA Base (1M)	0.333 (0.024)
Bio-ELECTRA Base (1.2M)	0.328 (0.013)
Bio-ELECTRA Base (1.2M) (weighted)	0.336 (0.023)

Table 5
Test Performance for Keyword Selection Classifiers

Model	Precision	Recall	F_1
LSTM Multi-input Model	91.72 (0.99)	89.39 (1.70)	90.53 (0.67)
Bio-ELECTRA++	97.58 (0.47)	96.93 (0.54)	97.25 (0.24)

task is cast as a sequence tagging problem. Bio-ELECTRA++ was selected over other larger Bio-ELECTRA models, because of its inference time performance is up to eight times better than the larger models. From BioASQ 5b, 752 training questions and 100 test questions were annotated for each word in the question being a keyword or not. The performance of both models averaged over ten randomly initialized training/testing phases is shown in Table 5, which shows that Bio-ELECTRA++ based keyword selection significantly outperforms LSTM based multi-input model.

7. Yes/No Question Answer Determination

Yes/no question answer determination from provided passages can be cast as a binary classification similar to sentiment classification to determine the implicit sentiment positive (yes) or negative (no) from the given context. However, some of the candidate passages might not provide enough evidence for either of the sentiments. In these cases due to the binary nature of the decision making, spurious decisions can be introduced. This is especially a problem with sentence level operation nature of the Bio-AnswerFinder. To remedy this, a third label (neutral) is introduced.

Negative sampling for neutral label was done using weighted rWMD based sentence similarity where sentences from the snippets are selected based on their weighted rWMD score being less than or equal to 0.6 compared to the sentences of the ideal answer. Snippet sentences having weighted rWMD score greater than or equal to 0.8 were chosen as additional label support sentences besides ideal answers. The thresholds were selected by minimizing the number of questions without any neutral sentence given the threshold values. While random sampling from other questions could be easily used for negative (neutral) sampling, the goal is to differentiate between candidate sentences related to the question but not provide an answer. The neutral sentences selected this way were afterwards checked and labeled manually.

From BioASQ 8b training data, 727 yes/no questions were selected for training and 128 for testing. Training/testing instances were prepared from sentences of the ideal answers and snippets. For yes/no classification there were 727 training instances and 128 test instances. For yes/no/neutral classification there were 2938 training instances and 539 testing instances. Nine language representation models were evaluated for yes/no classification to decide on the model for further yes/no/neutral answer classification. Test results for the average of ten randomly initialized classifiers per language representation model together with their standard deviations are shown in Table 6. The best performing Bio-ELECTRA Base model pretrained for 1M steps was selected for comparison experiments between yes/no and yes/no/neutral classifiers together with three different voting strategies for final decision.

During inference time, either the first ten highest ranked answer candidate sentences selected by Bio-AnswerFinder or the snippets as they are provided by BioASQ challenge is passed to the classifiers to make yes/no decision on each one of the candidate sentences/snippets. The final decision is made by a voting strategy. To this end, three voting strategies were used. The majority voting strategy uses the most common yes/no decision as the final decision. The best score strategy uses the decision of the answer candidate with the highest score as the final decision. The score voting strategy uses the highest sum of scores for the yes and no predicted answer candidates as the final decision. For evaluation, snippets provided for 128 test questions were scored by both types of classifiers. For yes/no/neutral classifier, any snippets with a neutral score greater than 0.5 were excluded from the voting. The test results are shown in Table 7. The yes/no/neutral classifier with score voting was the best performing classifier, which was chosen to be used for BioASQ 9 challenge.

8. Summary Question Handling

8.1. Extractive Summarization for BioASQ Summary Questions

In extractive summarization, a summary is generated by selecting sentences for the documents/snippets to be summarized. The introduced salient sentence selection strategy leverages the ranked sentences outputted by the Bio-AnswerFinder where the top 10 ranked answer candidate sentences are used. To minimize repetition, a hierarchical agglomerative clustering using weighted relaxed word mover's distance (wRWMD) similarity is introduced to group sentences where the cluster merge stop similarity threshold to maximize ROGUE-2 score was determined on training set answer summaries. From each cluster, the highest Bio-AnswerFinder ranked sentence is selected. The selected sentences are then ordered by their abstract occurrence

Table 6
Biomedical Yes/No Question Answer Classification Test Results

Model	P (Yes)	R (Yes)	F_1 (Yes)	P (No)	R (No)	F_1 (No)
Bio-ELECTRA++	91.24 (1.57)	95.29 (2.31)	93.19 (0.75)	78.91 (7.41)	63.85 (7.92)	69.84 (3.87)
ELECTRA Small++	88.18 (0.71)	94.31 (1.74)	91.14 (1.00)	69.92 (7.34)	50.38 (3.19)	58.40 (3.61)
BERT Base	87.02 (2.57)	95.49 (2.64)	90.99 (1.00)	65.15 (22.99)	43.46 (15.20)	51.71 (17.49)
BioBERT	92.94 (1.19)	93.04 (1.55)	92.63 (0.91)	71.94 (4.10)	69.23 (5.16)	70.42 (3.46)
ELECTRA Base	94.73 (1.67)	96.19 (0.82)	95.44 (0.88)	82.02 (3.08)	76.32 (8.01)	78.86 (5.06)
Mid (1M)	98.07 (0.71)	94.71 (1.40)	96.36 (0.96)	81.83 (4.15)	92.69 (2.69)	86.89 (3.23)
Base (1M)*	97.43 (1.14)	95.98 (1.20)	96.69 (0.58)	85.31 (3.64)	90.00 (4.62)	87.46 (2.31)
Mid (1.2M)	95.71 (1.97)	94.80 (1.76)	95.23 (0.89)	80.69 (4.44)	83.08 (8.28)	81.52 (4.16)
Base (1.2M)	97.22 (1.21)	95.49 (1.26)	96.34 (0.83)	83.61 (3.78)	89.23 (4.80)	86.23 (3.17)

Table 7
Yes/No versus Yes/No/Neutral Classification Performance

Model	P (Yes)	R (Yes)	F_1 (Yes)	P (No)	R (No)	F_1 (No)
Yes/No classifier Bio-ELECTRA Base (1M)						
majority voting	77.78	95.79	85.85	88.57	54.39	67.39
best score	80.91	93.68	86.83	85.71	63.16	72.73
score voting	80.91	93.68	86.83	85.71	63.16	72.73
Yes/No/Neutral classifier Bio-ELECTRA Base (1M)						
majority voting	84.62	93.62	88.89	86.67	70.91	78.00
best score	84.85	89.36	87.05	80.00	72.73	76.19
score voting	85.44	94.68	89.90	88.89	72.73	79.21
Yes/No/Neutral classifier Bio-ELECTRA Base (1M) seq length: 256						
majority voting	84.76	94.68	89.45	86.64	70.91	78.79
best score	85.29	92.55	88.78	85.11	72.73	78.43
score voting	85.58	94.68	89.90	88.89	72.73	80.0

order and concatenated.

8.2. Abstractive Summarization for BioASQ Summary Questions

Unlike extractive summarization where the summary is generated from the sentences of candidate documents/snippets, in abstractive summarization new content summarization the candidate documents/snippets is generated. To this end, a unified text-to-text transformer model called T5 [10] is trained with combined snippets as the document and the ideal answer as the summary for all summary questions from the BioASQ 8B training data. As a preprocessing step, any overlapping snippets are detected and only the longest of the overlapping snippets are included in generating document to be summarized. A T5 Base model is fine-tuned with a

maximum input sequence length of 512, batch size of 2 for 2 epochs to generate summaries of 150 tokens maximum.

9. BioASQ 2021 Synergy Task Systems

In the BioASQ 9 Synergy task, all questions were on the developing problem of COVID-19 without any guarantee that all them could be answered at the moment. There are no separate information retrieval and question answering from provided snippets, making the task only suitable to end-to-end systems. Also, feedback from the domain experts is provided after each round of the task allowing the participating systems to take advantage of the provided feedback in the next round. For the BioASQ Synergy document retrieval and snippets selection, Bio-AnswerFinder [9] was used. Instead of the LSTM based keyword selection classifier, a Bio-ELECTRA++ [6] model based keyword selection classifier described in Section 6 was used for better performance.

Starting from Synergy round 2, provided expert feedback data was used to augment the training data used for the BERT [1] based reranker classifier the Bio-AnswerFinder uses after weighted relaxed word mover's distance (wRWMD) similarity based ranking and focus word based filtering. At each round the BERT Base based reranker was retrained with the cumulative Synergy expert feedback.

Also for the rounds 2 and 3, an alternative keyword search engine (instead of Elasticsearch) was used after keyword query generation which was based on Pyserini search engine with MonoT5 based document re-ranking [17].

For round 4, a GloVe [18] embedding vector similarity based boolean search engine was developed where an approximate KNN GloVe vector similarity index was used for efficient similarity based retrieval of expansions for query keywords. The candidate set of abstracts retrieved by this search engine was combined with the Elasticsearch retrieved results for downstream processing by the Bio-AnswerFinder. The results of this system was entered as 'bio-answerfinder-2' to the Synergy task web site.

GloVe vectors used for round 1 and 2 were generated from the 2017 PubMed abstracts thus having no COVID-19 related terms resulting Bio-AnswerFinder excluding COVID-19 related terms from selected keywords for abstract retrieval, weighted relaxed word mover's distance (wRWMD) similarity based ranking resulting in system degradation. After this was noticed, new GloVe vectors were trained on the 2021 base PubMed abstracts and used to retrain affected classifiers in the Bio-AnswerFinder which were used in rounds 3 and 4.

9.1. Exact Answers/Ideal Answers

The re-ranked candidate sentences from the Bio-AnswerFinder are the input to the Synergy challenge subsystems.

9.1.1. Factoid and List Questions

For factoid and list questions, answer span classifier and post-processing described in Section 4 was used. Since, at the time of Synergy challenge Bio-ELECTRA models were not pretrained,

ELECTRA_Base [5] were fine-tuned using the combined SQuAD v1.1 and BioASQ 8b training data.

For factoid ideal answers, the highest Bio-AnswerFinder re-ranked sentence that contains the highest scored exact answer was selected. For list ideal answers, the sentence containing the most number of highest scored exact answers was selected among the top ten Bio-AnswerFinder re-ranked sentences.

9.1.2. Yes/No Questions

For yes/no questions, both binary and ternary classifiers described in Section 7 were used for different rounds. Similar to the factoid and list questions ELECTRA Base [5] models were used. Top 10 Bio-AnswerFinder selected sentences were passed to the binary classifier for the round 1 and yes/no decision was based on majority voting. The three-way ELECTRA Base based classifier for yes, no and neutral sentences was used using majority voting for rounds 2, 3 and 4. The highest re-ranked sentence from Bio-AnswerFinder was selected as the ideal answer.

9.1.3. Summary Questions

For summary questions, the extractive system described in Section 8.1 was used.

10. BioASQ 2021 9B Systems

Similar to the Synergy task, for BioASQ 9B Phase A task, Bio-AnswerFinder [9] was used with the Bio-ELECTRA++ based keyword classifier and Bio-ELECTRA Mid based re-ranker as described in Section 2. The iterative keyword query against Elasticsearch based document retrieval mechanism of the Bio-AnswerFinder was enhanced by a word embeddings based keyword synonym expansion mechanism for the batches 4 and 5. For each keyword selected by the Bio-ELECTRA++ based keyword classifier up to four most similar (by cosine similarity of GloVe word vectors) were added as synonyms to the Elasticsearch query which was iteratively refined until enough documents are returned. This approach was used for the challenge Task A system "bio-answerfinder-2".

For Task 9B Phase B, snippets provided by BioASQ challenge were first passed through Bio-AnswerFinder after bypassing retrieval section. The re-ranked candidate sentences were the input to the challenge subsystems.

10.1. Factoid and List Questions

For factoid and list questions, Bio-ELECTRA Base based answer span classifier and post-processing described in Section 4 was used. For ideal answers, the same mechanism as in the Synergy Task was used.

10.2. Yes/No Questions

For yes/no questions, the best performing Bio-ELECTRA Mid model based ternary yes/no/neutral classifier described in Section 7 was used. Final decision was made by score voting. For ideal

answers, also the same mechanism as in the Synergy Task was used.

10.3. Summary Questions

For the summary questions, both the extractive and abstractive systems described in the Section 8.1 and Section 8.2, respectively, were used. The abstractive summarization system was used for the BioASQ challenge Task B system "bio-answerfinder-2".

11. Discussion

Bio-AnswerFinder together with the extensions introduced in this paper is one of the few end-to-end systems participating BioASQ challenges that can handle Synergy and both phases of Task B for all question types.

For the Synergy task the systems described in Section 9, GloVe vectors used in the rounds 1 and 2 did not have any COVID-19 related terms since they were generated from 2017 PubMed abstracts. This had detrimental effects to the document retrieval and ranking which relies on GloVe vectors for keyword ranking for greedy iterative retrieval and wRWMD based ranking. Since documents with a feedback from previous rounds need to be excluded from the eligible document pool for the questions in the subsequent rounds, the detrimental effect from the first two rounds adversely affected the other rounds also. Also, because of a misunderstanding of the instructions for the Synergy challenge, only abstracts with a PubMed ID were indexed for search leaving out all preprint abstracts that make about the half of the COVID-19 corpus. This was not noticed until the Synergy version 2 challenge. Even after these setbacks, the system performance was decent based on official BioASQ Synergy Task results (on average, 12th out of the 23 individual systems on documents F_1 , 12th out of the 24 systems on snippets F_1 , 6th out of the 24 systems on yes/no overall F_1 , 6th out of the 24 systems on factoid MRR and 5th out of the 24 systems on list F_1). The GloVe embedding vector similarity based boolean search engine introduced for the round 4 to increase coverage over the iterative keyword query based document retrieval improved performance over the default retrieval based on the official Synergy Task test results.

In the BioASQ 9B Phase A, the introduced system was the best system on document retrieval in four out of the five test batches based on F_1 score and second on the remaining batch. In snippets, the system was second best in two batches and third in three batches. The keyword synonym expansion approach described in Section 10 ('bio-answerfinder-2') used in batches 4 and 5 had slightly worse performance on document retrieval. For snippets, the results were more mixed, the expansion approach had better performance than the original system in batch 4 while performing worse in batch 5.

In the 9B Phase B, the introduced systems were second for yes/no questions in test batches 2 and 3. For the list questions, the performance was better than last year. While factoid question performance was decent, there is room for improvement. However, based on the factoid question error analysis for last year's submissions described in the next section, the observed near-miss issue is suspected this year also. This will be investigated once the gold standard annotations will be available. For the ideal questions, only automatic ROUGE scores are available. Based on both ROUGE-2 (F1) and ROUGE-SU4 (F1) scores, the introduced system was the best scoring

system for the test batches 1 and 2. Despite the terse nature of its results (usually a single sentence), T5 [10] based abstractive summarization system 'bio-answerfinder-1' seems to work well outperforming extractive summarization in batches 2 and 4.

In BioASQ 8B, Bio-AnswerFinder won the second place for human evaluated ideal answers in three test batches. Since Bio-AnswerFinder was mainly designed as a practical knowledge discovery tool for biomedical researchers who prefer ideal answers (answer with evidence and context), this result was very encouraging validation towards the main design goal of Bio-AnswerFinder.

11.1. Error Analysis for Factoid Questions

Based on the analysis of the BioASQ 8b factoid question 'bio-answerfinder' system submissions against the ground truth answers in the BioASQ 9B training data, it was identified that about 53% of the errors can be attributed to near misses, i.e. singular/plural differences, differences in stop words (e.g. articles), single special character differences, acronym versus its expansion, and other transliterations or paraphrasings. Another common issue is the provided ground truth being a paraphrased sentence, more akin to an ideal answer than factoid answer, not occurring in any of the supplied documents for the question. Representative near miss error of different types are shown in Table 8. For examples 4 and 5, the ground truth does not exist in the provided phrases. Even though these were errors for the automatic evaluation, for a human user the predicted answers would be the correct answers. QA systems are designed for human usage and while automatic evaluation provides fast, systematic evaluation of QA systems, more than 50% near miss emphasizes importance of human evaluation for QA systems even though they are more costly than automatic evaluation.

12. Conclusions

In this paper, extensions to an end-to-end biomedical QA system, Bio-AnswerFinder [9] for BioASQ biomedical question answering challenge were introduced. To this end, three ELECTRA [5] discriminative language representation models were pretrained from scratch on PubMed abstracts and PMC open access papers. Based on performance comparison against numerous other language representation models including BioBERT, the introduced Bio-ELECTRA models had shown superior performance for the classifiers used in the Bio-AnswerFinder sub-systems. The resulting system(s) had shown very good performance in BioASQ 9B Phase A and good performance for yes/no questions and ideal answers in Phase B based on the official automatic evaluation results. In the future, sensitivity of some subsystems such as keyword ranking and weighted relaxed WMD measure based answer candidate ranking to the out-of-vocabulary terms will be addressed. Based on the insights of the in-depth analysis of questions that cannot be properly answered from BioASQ Synergy and 9B Tasks, Bio-AnswerFinder will be further improved with the eventual goal of answering all answerable biomedical domain questions.

Table 8

Examples of near miss factoid errors for BioASQ 8b bio-answerfinder system

No	Type	Description
1	Question	What gene is mutated in Huntington’s Disease patients?
	Ground Truth	HTT gene encoding the protein huntingtin
	Prediction	HTT gene huntingtin (HTT) gene huntingtin gene
2	Question	Which diagnostic test is approved for coronavirus infection screening?
	Ground Truth	real-time reverse transcription-PCR
	Prediction	rRT-PCR
3	Question	How large is a lncRNAs?
	Ground Truth	>200 nucleotides
	Prediction	more than 200 nucleotides
4	Question	When was vaxchora first licensed by the FDA?
	Ground Truth	10 June 2016
	Prediction	June 10, 2016
5	Question	What is the LINCS Program?
	Ground Truth	NIH-funded program to generate a library of integrated, network-based, cellular signatures
	Prediction	NIH Common Fund Library of Integrated Network-based Cellular Signatures Library of Integrated Network-based Cellular Signatures

13. Software and Data Availability

Bio-AnswerFinder source code and documentation is available on GitHub (<https://github.com/scicrunch/bio-answerfinder>). The datasets used for Bio-ELECTRA model evaluations and Bio-ELECTRA++ source code are available on Github (https://github.com/SciCrunch/bio_electra). The small Bio-ELECTRA++ models are available on Zenodo (<https://doi.org/10.5281/zenodo.3971235>). The mid and base sized pre-trained Bio-ELECTRA models are available on Zenodo (<https://doi.org/10.5281/zenodo.4699034>).

14. Acknowledgments

This work was supported by the NIDDK Information Network (dkNET; <http://dknet.org>) via NIH’s National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) award U24DK097771. I would like also to thank Google TensorFlow Research Cloud (TFRC) program for providing me with free TPUs which allowed me to pretrain Bio-ELECTRA models.

References

- [1] 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.
- [2] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, Q. V. Le, Xlnet: Generalized autoregressive pretraining for language understanding, in: *Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., 2019, pp. 5753–5763.
 - [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.
 - [4] 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 (2019) 1234–1240. URL: <https://doi.org/10.1093/bioinformatics/btz682>. doi:10.1093/bioinformatics/btz682.
 - [5] 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.
 - [6] 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, 2020, pp. 104–112. URL: <https://www.aclweb.org/anthology/2020.sdp-1.12>. doi:10.18653/v1/2020.sdp-1.12.
 - [7] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, Y. Almirantis, J. Pavlopoulos, N. Baskiotis, P. Gallinari, T. Artieres, A. Ngonga, N. Heino, E. Gaussier, L. Barrio-Alvers, M. Schroeder, I. Androutsopoulos, G. Paliouras, An overview of the bioasq large-scale biomedical semantic indexing and question answering competition, *BMC Bioinformatics* 16 (2015) 138. URL: <http://www.biomedcentral.com/content/pdf/s12859-015-0564-6.pdf>. doi:10.1186/s12859-015-0564-6.
 - [8] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, Overview of bioasq 8a and 8b: Results of the eighth edition of the bioasq tasks a and b, in: *Proceedings of the 8th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, 2020. URL: http://ceur-ws.org/Vol-2696/paper_164.pdf.
 - [9] I. B. Ozyurt, A. Bandrowski, J. S. Grethe, Bio-AnswerFinder: a system to find answers to questions from biomedical texts, *Database* 2020 (2020). URL: <https://doi.org/10.1093/database/baz137>. doi:10.1093/database/baz137, baz137.
 - [10] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, 2020. arXiv:1910.10683.
 - [11] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (1997) 1735–1780. URL: <https://doi.org/10.1162/neco.1997.9.8.1735>. doi:10.1162/neco.1997.9.8.1735.
 - [12] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, G. Hullender, Learning to rank using gradient descent, in: *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, Association for Computing Machinery, New York, NY, USA, 2005, p. 89–96. URL: <https://doi.org/10.1145/1102351.1102363>. doi:10.1145/1102351.1102363.

- [13] I. B. Ozyurt, J. Grethe, Iterative document retrieval via deep learning approaches for biomedical question answering, in: 2019 15th International Conference on eScience (eScience), 2019, pp. 533–538. doi:10.1109/eScience.2019.00072.
- [14] M. Kusner, Y. Sun, N. Kolkin, K. Weinberger, From word embeddings to document distances, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, PMLR, Lille, France, 2015, pp. 957–966.
- [15] R. Sennrich, B. Haddow, A. Birch, Neural machine translation of rare words with subword units, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Berlin, Germany, 2016, pp. 1715–1725. URL: <https://www.aclweb.org/anthology/P16-1162>. doi:10.18653/v1/P16-1162.
- [16] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: <https://www.aclweb.org/anthology/D16-1264>. doi:10.18653/v1/D16-1264.
- [17] R. Nogueira, Z. Jiang, R. Pradeep, J. Lin, Document ranking with a pretrained sequence-to-sequence model, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 708–718. URL: <https://www.aclweb.org/anthology/2020.findings-emnlp.63>. doi:10.18653/v1/2020.findings-emnlp.63.
- [18] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: <https://www.aclweb.org/anthology/D14-1162>. doi:10.3115/v1/D14-1162.