Biomedical Question Answering with SDNet

Lu Yang, Sophia Lu, and Erin Brown
Stanford University
{luy, sophialu, browne}@stanford.edu
Mentor: Suvadip Paul

Abstract

Motivation: Question answering (QA) systems usually rely on complex natural language processing approaches to understand the questions and find answers precisely. Advances in neural network models promise better performance in open-domains where large datasets are available, such as SQuAD 2.0, QuAC, and CoQA. However, these systems need to be exploited further in more restricted domains where construction of the training sets is costly. With the progress in natural language processing, extracting valuable information from biomedical literature has become a less daunting task. Recent developments in training a domain specific language representation models have paved the way for mining a large number of unannotated biomedical texts. **Aims:** We leverage BioBERT to effectively transfer the information from large-scale biomedical corpora and adapt the recently published SDNet neural architecture to tackle the BioASQ QA challenge [1, 2]. The original SDNet model was developed for general conversational QA tasks. We incorporate biomedical domain knowledge through BioBERT and domain-specific word2vec embeddings into the system by adapting on the proposed SDNet. Our approach requires some adaptation to the structure, but we are able to leverage the various attention mechanisms at its core. Ultimately, we work to develop a novel, contextual QA system specifically for biomedical-related text mining. **Results:** We have implemented a combined BioBERT-SDNet model to achieve competitive, though not state-of-the-art, results on biomedical question answering tasks. We expect that with some further work, this model could be a powerful tool for biomedical question answering in practice. To our knowledge, ours is the first usage of SDNet for non-conversational QA and the first integration of BioBERT with SDNet.

1 Introduction

Each year, an immense number of new biomedical documents are published. In 2017 alone, the National Library of Medicine reported that over 800,000 new citations were added to MEDLINE [3]. At such a rate, it is not feasible for medical professionals to keep up with recent developments. For this reason, natural language processing systems that can identify and return relevant information in a human readable form are essential.

The BioASQ organization seeks to address this problem via their yearly challenge on biomedical semantic indexing and QA [4]. The BioASQ challenge incorporates a number of different tasks, including semantic indexing (Task A), information retrieval (Task B Phase A), and question answering (Task B Phase B). We focus here on the question answering (QA) component, which in itself comprises four distinct types of questions including factoid, list, yes-no, and summary. For factoid, list, and yes-no questions, the challenge measures performance based on both exact answers and ideal answers, which are short rationales for the exact answers. We concentrate on obtaining exact answers for factoid questions, which have historically been more difficult than the yes-no questions and which are easily extensible into list questions. As of last year's BioASQ6b challenge, the best score obtained on factoid question exact answering was a strict accuracy of 0.24 [5], indicating that this is indeed an area of the challenge for which there remains significant room for improvement.

In this project, we leverage BioBERT, a domain specific language representation model pre-trained on large-scale biomedical corpora, and implement a modified SDNet model for exact factoid question answering in Task B, Phase B of the BioASQ challenge [2]. Our problem is formulated as follows: given a passage C, predict the answers $A_1, A_2, A_3, ..., A_k$ to the questions $Q_1, Q_2, Q_3, ..., Q_k$.

2 Related work

The OAQA Biomedical Question Answering System won the best-performing system in the factoid category of the BioASQ challenge two years in a row in 2015 and 2016 14 [6]. This system consisted of a number of distinct modules, including: key multi-word term extraction by frequency; biomedical named entity recognition and concept identification; synonym retrieval; part-of-speech tagging; candidate answer generation from identified concepts; and candidate answer relevance estimation. While successful, this model required great complexity, and has not kept pace with other more recent developments in terms of performance.

The DeepQA system of the 2017 BioASQ5b challenge introduced an extractive neural QA model, restricting the system to output substrings of the provided text snippets [7]. This system was based on FastQA [8], extended with biomedical word embeddings, pre-trained on SQuAD, and fine-tuned on the BioASQ training set. FastQA is a BiLSTM-based system which at the time of its publication achieved state-of-the-art results on the SQuAD challenge v1 and continues to rank on the v1 leaderboard today. As an ensemble system, DeepQA achieved state-of-the-art performance in 2017, but does not seem to have been entered into the challenge the following year.

The LabZhu team led the BioASQ leaderboard for exact-answer factoid questions in 2018 after performing well in the 2017 BioASQ5b challenge as well [7, 5]. They applied rule-based question type analysis and used the Stanford POS tagger along with PubTator, a tool for annotating biological entities and their relationships, for candidate answer generation. They also used word frequencies for candidate answer ranking. Notably, the LabZhu team's system outperformed the single model DeepQA in 2017, though it did not perform as well as the ensemble DeepQA model. Nevertheless, its exceptional performance shows that even in 2018, rule-based rather than neural QA methods still dominated for the challenge.

Given the success of the DeepQA system on the BioASQ challenge, it is a natural next step to explore the performance of current state-of-the-art SQuAD and related models for the BioASQ challenge, and this motivates our approach.

3 Approach

We implement a domain-specific adaptation of the recently published SDNet model for the BioASQ challenge. SDNet is a complex, innovative, and contextualized attention-based deep neural network that integrates the latest BERT contextual model with multiple attention mechanisms to achieve state-of-the-art results on the CoQA leaderboard. Whereas SDNet was designed specifically for conversational QA, we adapt it for single-turn QA. One distinct advantage of adapting a model designed for CoQA over one designed for SQuAD or similar challenges is that CoQA allows for free-form text answers, while SQuAD, for example, is purely extractive. While at present we have focused on extractive QA for the BioASQ challenge, the model we have adapted will naturally be able to tackle abstractive QA as well with some tuning. This capacity is highly desirable as it enables the generation of more natural-sounding answers and coherent accompanying rationales. Thus with some tuning our model can naturally be extended to summarization and ideal answer generation tasks that are both part of the BioASQ challenge and integral to any functional QA system that might be used in practice by biomedical professionals.

We combine the SDNet model architecture with BioBERT, another recently published development. BioBERT is a pre-trained biomedical language representation model for biomedical text mining based on the BERT architecture and trained over large-scale biomedical corpora, including PubMed abstracts, PMC full-text articles, as well as general-domain corpora including English Wikipedia and BooksCorpus [2]. While BERT has achieved competitive performance compared to previous state-of-the-art models on many biomedical-domain text mining tasks, BioBERT has been shown to perform significantly better, highlighting the importance of domain-specific tailoring. While testing on the BioASQ4b challenge factoid question set, for example, Lee et. al. found that BioBERT achieved an absolute improvement of 9.73% in strict accuracy over BERT and 15.89% over the previous state-of-the-art [2]. For this task, BioBERT was fine-tuned using the BERT model designed for SQuAD [9].

3.1 Encoding layer

For our model, we use BioBERT rather than the standard BERT as discussed above. Because this procedure could require up to 20 days on 8 V100 GPUs, we utilize the publicly available pre-trained weights of BioBERT to fine-tune BERT architecture on our training datasets. These are fed into the encoding layer of the SDNet architecture.

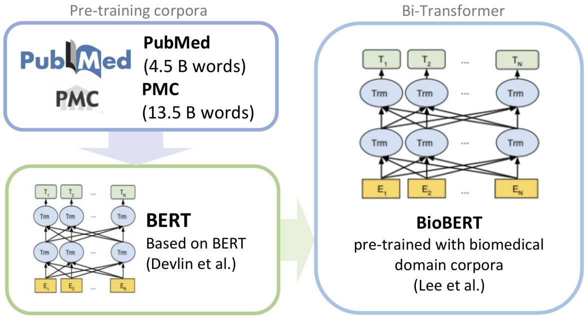


Figure 1: Overview of the pre-training and fine-tuning of BioBERT. Reproduced from [2].

SDNet employs both standard GloVe embedding and BERT contextualized embedding. In the SDNet publication, the BERT contextualized embedding is derived from a weighted sum of all L hidden states of the transformer with locked internal weights rather than the final hidden state alone as proposed in the original BERT publication [9]. In detail, suppose that a word w has BPE tokenization $w = \{b_1, b_2, \ldots, b_s\}$, and let $\boldsymbol{h}_t^l$ denote the lth hidden state for the tth BPE token b_t, $1 \le l \le L$ and $1 \le t \le s$. Then the contextualized embedding for w is given by the per-layer weighted sum of average BERT embedding $\text{BERT}_w = \sum_{l=1}^{L} (\alpha_l / s) \sum_{t=1}^{s} \boldsymbol{h}_t^l$ with weights $\alpha_1, \ldots, \alpha_l$, where these weights are hyperparameters that may be tuned.

3.2 Integration layer

The integration layer of SDNet features four different attention mechanisms along with the core BiLSTM layers.

3.2.1 Obtaining word input vectors with word-level inter-attention

First, question-to-context attention is conducted based on the standard embedding. Suppose there are m context words and n question words with embeddings $\boldsymbol{h}_1^C, \ldots, \boldsymbol{h}_m^C \in \mathbb{R}^d$ and $\boldsymbol{h}_1^Q, \ldots, \boldsymbol{h}_n^Q \in \mathbb{R}^d$ respectively, where d is the dimension of the standard embedding. First, the similarity matrix $\boldsymbol{S} \in \mathbb{R}^{m \times n}$ is computed, where each entry is given by

$$\boldsymbol{S}_{ij} = \text{ReLU}(\boldsymbol{U}\boldsymbol{h}_i^C)\boldsymbol{D}\ \text{ReLU}(\boldsymbol{U}\boldsymbol{h}_j^Q) \tag{1}$$

where $\boldsymbol{D} \in \mathbb{R}^{k \times k}$ and $\boldsymbol{U} \in \mathbb{R}^{d \times k}$ for attention hidden size k. From the similarity matrix, the attention score distribution $\boldsymbol{a}$ is computed as $\boldsymbol{a}_{ij} = \text{softmax}\,(\boldsymbol{S}_{ij})$. Then the question-to-context attended vectors are given by $\hat{\boldsymbol{h}}_i^C = \sum_{j=1}^{n} \boldsymbol{a}_{ij} \boldsymbol{h}_j^Q$.

For notional simplicity, we will refer to the attention function defined here as simply $\text{Attn}\,(\mathbf{A}, \mathbf{B}, \mathbf{C})$ moving forward. SDNet also incorporates a feature vector $\boldsymbol{f}_w$ into the input vector for each context word per the methodology developed in DrQA [10]. This vector includes a 12-dimensional part-of-speech embedding, an 8-dimensional named entity recognition embedding, a 3-dimensional exact matching vector indicating whether each context word appears in the question in original, lowercase, or lemma form, and a normalized term frequency entry. Using this feature vector, the input vector for each context word $\tilde{\boldsymbol{w}}$ becomes

$$\tilde{\boldsymbol{w}}_i^C = \left[\text{GloVe}_{w_i^C}; \text{BERT}_{w_i^C}; \hat{\boldsymbol{h}}_i^C; \boldsymbol{f}_{w_i^C}\right],$$

and our question word input vector remains simply

$$\tilde{\boldsymbol{w}}_i^Q = \left[\text{GloVe}_{w_i^Q}; \text{BERT}_{w_i^Q}\right]. \tag{2}$$

3.2.2 Contextualized understanding of context and question

After obtaining input word vectors, the SDNet model uses two BiLSTMs to form the contextualized understanding for context and question:

$$\boldsymbol{h}_1^{C,k}, \dots, \boldsymbol{h}_m^{C,k} = \text{BiLSTM}(\boldsymbol{h}_1^{C,k-1}, \dots, \boldsymbol{h}_m^{C,k-1}), \tag{3}$$

$$\boldsymbol{h}_1^{Q,k}, \dots, \boldsymbol{h}_m^{Q,k} = \text{BiLSTM}(\boldsymbol{h}_1^{Q,k-1}, \dots, \boldsymbol{h}_m^{Q,k-1}), \tag{4}$$

$$\boldsymbol{h}_i^{C,0} = \tilde{\boldsymbol{w}}_i^C, \;\; \boldsymbol{h}_i^{Q,0} = \tilde{\boldsymbol{w}}_i^Q \tag{5}$$

for $1 \leq k \leq K$ where K is the number of RNN layers. In the implementation published, SDNet implements variational dropout for the input vector to each layer of the RNN, which we do also.

Next, one more layer of RNN is applied for each question word:

$$bfh_1^{Q,K+1}, \dots, \boldsymbol{h}_n^{Q,K+1} = \text{BiLSTM}(\boldsymbol{h}_1^Q, \dots, h_n^Q), \qquad \boldsymbol{h}_i^Q = \left[\boldsymbol{h}_i Q, 1; \dots; \boldsymbol{h}_i^{Q,K}\right]. \tag{6}$$

Self-attention is then applied on the question to obtain the final question representation

$$\{\mathbf{u}_i^Q\}_{i=1}^n = \text{Attn}\left(\{\boldsymbol{h}_i^{Q,K+1}\}_{i=1}^n, \{\boldsymbol{h}_i^{Q,K+1}\}_{i=1}^n, \{\boldsymbol{h}_i^{Q,K+1}\}_{i=1}^n\right). \tag{7}$$

Next, multi-level inter-attention is applied from question to context based on all layers of generated representations. The history-of-word concept from FusionNet [11] is employed to boost computational efficiency, where context and question history-of-word vectors are denoted respectively as HoW_i^C and HoW_i^Q From each RNN layer output of question to context, $K+1$ times of multilevel attention are conducted in terms of these history-of-word vectors which we denote HoW_i^C and HoW_i^Q for context and question respectively:

$$\{\boldsymbol{m}_i^{(k),C}\}_{i=1}^m = \text{Attn}\left(\{\text{HoW}_i^C\}_{i=1}^m, \{\text{HoW}_i^Q\}_{i=1}^n, \{\text{HoW}_i^Q\}_{i=1}^n,\right) \tag{8}$$

with $1 \leq k \leq K+1$. Then, an additional RNN layer is applied in order to obtain the contextualized representation $\boldsymbol{v}_i^C$ for each context word:

$$\boldsymbol{v}_1^C, \dots, \boldsymbol{v}_m^C = \text{BiLSTM}(\boldsymbol{y}_1^C, \dots, \boldsymbol{y}_n^C), \tag{9}$$

where $\boldsymbol{y}_i^C = [\boldsymbol{h}_i^{C,1}; \dots; \boldsymbol{h}_i^{C,k}; \boldsymbol{m}_i^{(1),C}; \dots; \boldsymbol{m}_i^{(K+1),C}]$. Self-attention on context is applied as the final attention layer, using the history-of-word concept as above:

$$\mathbf{s}_i^C = [\text{GloVe}_{w_i^C}; \text{BERT}_{w_i^C}; \boldsymbol{h}_i^{C,1}; \dots; \boldsymbol{h}_i^{C,k}; \boldsymbol{m}_i^{(1),Q}; \dots; \boldsymbol{m}_i^{(K+1),Q}; \boldsymbol{v}_i^C] \tag{10}$$

and

$$\{\tilde{\boldsymbol{v}}_i^C\}_{i=1}^m = \text{Attn}\left(\{\mathbf{s}_i^C\}_{i=1}^m, \{\mathbf{s}_i^C\}_{i=1}^m, \{\boldsymbol{v}_i^C\}_{i=1}^m,\right). \tag{11}$$

The final context representation is obtained by applying an additional RNN layer:

$$\{\mathbf{u}_i^C\}_{i=1}^n = \text{BiLSTM}\left([\boldsymbol{v}_1^C; \tilde{\boldsymbol{v}}_1^C], \dots, [\boldsymbol{v}_m^C; \tilde{\boldsymbol{v}}_m^C]\right). \tag{12}$$

3.3 SDNet output layer

In the output layer, for parametrized vector $\boldsymbol{w}$ we condense the question representation into a single vector $\mathbf{u}^Q = \sum_{i=1}^{n} \beta_i \mathbf{u}_i^Q$, where $\beta_i = \text{softmax}(\boldsymbol{w}^T \mathbf{u}_i^Q)$. We then generate predicted answers to questions in the standard way. Details regarding output generation are included in the appendix 8.1.

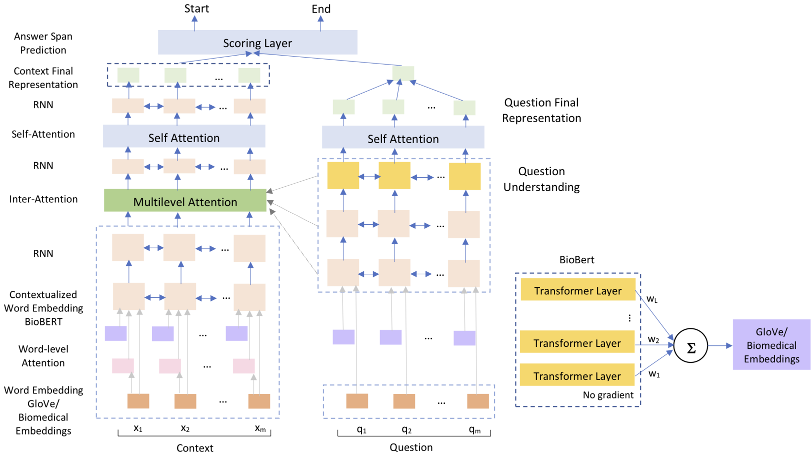


Figure 2: Overview of the pre-training and fine-tuning of BioBERT. Reproduced from [2].

4 Experiments

4.1 Data

We used historic data available from the BioASQ4b and BioASQ5b challenges to produce training, dev, and test sets. The training set for BioASQ4b contained 217 factoid questions, while the BioASQ5b training set contained 322. Included in each of the provided datasets are a set of questions, with type identified, and snippets from PubMed that are relevant to the specific question [4]. We created training datasets from the provided data in both SQuAD and CoQA formats. All of the information required for SQuAD format input was provided by the BioASQ challenge and so the SQuAD format was straightforward to produce. However, after processing the raw provided data into suitable formats, our training datasets still posed significant challenges in terms of both small size and inconsistancy in well-posedness. We found that though the factoid QA task is posed as one of extractive question answering, in many cases the golden answer provided cannot be found as-is in the text. We were able to identify a number of prominent subsets of problems causing mismatch between context and answer, each of which and their respective fixes are detailed with examples in the Appendix 8.3. Note that no alterations were made to test sets.

A conversation from the CoQA datasets contains a question, an answer, and a rationale that supports the answer [12]. As the SDNet model requires input of CoQA format, we had to generate text spans to accompany the provided golden answers. To do so we selected the span of text consisting of the answer text itself, which we could do after ensuring that all questions had an appropriately extractive answer, and the following span of length equal to three times the length of the golden answer. As our questions are non-conversational, we simply set all turn numbers to one. An example of processed question-answer pair from BioASQ datasets and its corresponding CoQA format is provided in the Appendix 8.2.

4.2 Evaluation Method

Exact answer factoid questions are evaluated based on *mean reciprocal rank (MRR)*, *strict accuracy* and *lenient accuracy*, in accordance with the standard BioASQ challenge evaluation methods [13]. Note that MRR is used as the deciding metric in ranking model performance for BioASQ. We also compute, in some cases, F_1 scores to get a better idea of our model performance. We provide the F_1 scores calculated across the entire test set and the F_1 scores of those questions for which the model output partially aligned with the golden answers respectively to better analyze model performance.

4.3 Experimental details

4.3.1 Baseline model

To provide an additional baseline, we ran the baseline code provided for the default project on a training dataset comprising 633 factoid questions from the BioASQ5b training set and the BioASQ4b test and training sets. In place of the default GloVe word vectors, we substituted a freely available pre-trained set of domain-specific word vectors [14]. We also ran the baseline code with the biomedical-domain word vectors on SQuAD data to ensure correctness and to determine how domain-specific vectors impact performance.

4.3.2 BioBERT

To show viability of BioBERT on the BioASQ challenge, we implemented our BioBERT based on the BERT repository by Google [15] and used pre-trained weights for biomedical text mining tasks released by *biobert-pretrained* repository [16, 17]. We fine-tuned the BioBERT model on the BioASQ4b and BioASQ5b training sets respectively and evaluated the model on each of the provided test sets. Hyperparameter optimization was performed on sampled datasets. The optimal hyperparameter values were determined to be a batch size of 12, a learning rate of 3e-5, and 50 epochs. The dropout probability was kept as 0.1. We also used a max sequence length of 384 and a doc stride of 128. The model size was adopted from $\text{BERT}_{\text{BASE}}$, with 12 Transformer blocks, a hidden size of 768, and 12 self-attention heads. Note that as we used WordPiece vocabulary with size of 28996 provided by Google, any new words in biomedical corpus can be represented with subwords (for instance, diphtheria-tetanus → dip ##ht ##her ##ia - te ##tan ##us).

4.3.3 SDNet

As in our BioBERT implementation, we obtained the *BioBERT v1.0 (+ PubMed 200K + PMC 270K)* version of the pre-trained BioBERT weights from the Naver GitHub [16]. As these weights are available as TensorFlow checkpoint files, we then converted them to PyTorch format for use with SDNet. Microsoft recently published the SDNet model from [1] to GitHub [18], and we employed it to construct our own SDNet model, with integration to BioBERT, modifications to basic configuration variables and input formatting. We used 300-dim Glove embedding [19] and the transformer output from BioBERT as contextualized embedding for each word in context and question.

We first evaluated our BioBERT-SDNet model on sampled CoQA conversational question answering datasets across five different domains. Table 3 summarizes F1 scores of our model and demonstrates the viability of our implementation. The BioBERT-SDNet model was then trained and evaluated with BioASQ train/dev sets in CoQA format. We initiated this with GloVE word embeddings of size 300. We used a mini-batch size of 32, a total of 30 epochs, and a learning rate of 0.0001. Specifically our model employs sizes as following: 2 layers of RNN context encoder, 2 layers of RNN question encoder, and a query self-attention hidden size of 300. In our 4b training data, 1196 out of 63640 words are out-of-vocabulary (OOV) (1.8793%), while 2095 out of 95793 words are OOV (2.1870%) in our 5b training data. Figure 7 in Appendix 8.5 demonstrates a successful implementation of BioBERT-SDNet model with loss convergence on the validation datasets over 30 epochs.

4.4 Results

The default baseline achieved an average strict accuracy of 0.14 over the BioASQ5b test sets, slightly surpassing the BioASQ5b-baseline strict accuracy of 0.13, and an average F_1 score of 16.12. Detailed results on the default baseline evaluated on BioASQ5b test sets are shown in Table 1. We also tested the baseline model with biomedical embeddings on the SQuAD dataset and achieved an F_1 score of 78.02 and an exact match score of 68.40 on the dev set.

Test Sets	**1**	**2**	**3**	**4**	**5**	**Average**
F1 (all)	9.21	21.30	23.44	14.44	12.22	**16.12**
SAcc	0.21	0.17	0.14	0.08	0.10	**0.14**

Table 1: Results of the default baseline model evaluated on BioASQ5b test sets.

Our BioBERT model achieved an average strict accuracy (SAcc), lenient accuracy (LAcc), and mean reciprocal rank (MRR) of 0.15, 0.25, and 0.19 on BioASQ4b challenge, outperforming the published BioASQ4b-baselines for factoid questions which were 0.07, 0.16, and 0.11 respectively. For BioASQ5b, we obtained an average SAcc, LAcc, and MRR of 0.26, 0.40, and 0.31 which were better than those of the published BioASQ5b-baslines, 0.13, 0.27, and 0.19 respectively.

Metrics	**Test Sets**										**Avg**	
	4b1	**4b2**	**4b3**	**4b4**	**4b5**	**5b1**	**5b2**	**5b3**	**5b4**	**5b5**	**4b**	**5b**
SAcc	0.15	0.06	0.12	0.2	0.12	0.28	0.23	0.42	0.18	0.2	**0.15**	**0.26**
LAcc	0.31	0.26	0.19	0.32	0.18	0.44	0.35	0.5	0.33	0.37	**0.25**	**0.40**
MRR	0.21	0.13	0.14	0.31	0.14	0.34	0.27	0.46	0.23	0.26	**0.19**	**0.31**

Table 2: Results of BioBERT testing evaluated on BioASQ4b/5b test sets for factoid questions. SAcc, LAcc, and MRR are reported.

To evaluate our BioBERT-SDNet, we also trained on CoQA data and evaluated on the CoQA dev set. The resulting F_1 scores were calculated and categorized by domain names in Table 3. Our BioBERT-SDNet model achieved an average strict accuracy (SAcc), lenient accuracy (LAcc), and mean reciprocal rank (MRR) of 0.13, 0.20, and 0.16 on BioASQ4b challenge, outperforming the published BioASQ4b-baselines for factoid questions. For BioASQ5b, we obtained an average SAcc, LAcc, and MRR of 0.18, 0.24, and 0.19 which were also above the published BioASQ5b-baslines. F_1 scores of all the tests were calculated for measuring both the F_1 score of the entire test set and the F_1 score of the test set members with at least a partial match, as shown in Tables 5 and 6.

	Children's stories	**Literature**	**Mid/high school**	**News**	**Wikipedia**	**Overall**
BioBERT-SDNet	**72.0**	**66.5**	**67.7**	**71.3**	**75.4**	**70.5**
BERT-SDNet	75.4	73.9	77.1	80.3	83.1	78.0

Table 3: F_1 scores of our BioBERT-SDNet predictions on CoQA dataset, compared against the results for single model SDNet with BERT as published in [1].

Metrics	**Test Sets**										**Avg**	
	4b1	**4b2**	**4b3**	**4b4**	**4b5**	**5b1**	**5b2**	**5b3**	**5b4**	**5b5**	**4b**	**5b**
SAcc	0.10	0.13	0.04	0.26	0.12	0.28	0.19	0.19	0.09	0.14	**0.13**	**0.18**
LAcc	0.15	0.16	0.14	0.34	0.18	0.35	0.23	0.25	0.16	0.20	**0.20**	**0.24**
MRR	0.12	0.13	0.11	0.30	0.16	0.31	0.19	0.20	0.12	0.13	**0.16**	**0.19**

Table 4: Results of BioBERT-SDNet testing evaluated on BioASQ4b/5b test sets for factoid questions. SAcc, LAcc, and MRR are reported.

Metrics	**Test Sets**										**Avg**	
	4b1	**4b2**	**4b3**	**4b4**	**4b5**	**5b1**	**5b2**	**5b3**	**5b4**	**5b5**	**4b**	**5b**
F_1 all	34.9	7.19	16.0	35.9	23.0	39.1	37.1	50.7	27.4	32.8	**23.4**	**37.4**
F_1 partial	71.7	55.7	69.2	79.6	65.2	69.9	71.9	77.6	75.3	67.5	**68.3**	**72.4**

Table 5: F_1 scores of BioBERT predictions evaluated on BioASQ4b/5b test sets.

Metrics	**Test Sets**										**Avg**	
	4b1	**4b2**	**4b3**	**4b4**	**4b5**	**5b1**	**5b2**	**5b3**	**5b4**	**5b5**	**4b**	**5b**
F_1 all	22.0	24.1	13.2	28.6	21.4	39.9	32.7	39.0	27.8	30.6	**21.9**	**34.0**
F_1 partial	61.3	49.8	57.3	80.6	60.8	71.2	63.3	63.3	61.2	71.5	**61.9**	**66.1**

Table 6: F_1 scores of BioBERT-SDNet predictions evaluated on BioASQ4b/5b test sets.

5 Analysis

Model comparison. Both the BioBERT and the BioBERT-SDNet models outperformed the published baselines. BioBERT produced average F_1 scores of 23.4 and 37.4 on the BioASQ4b and BioASQ5b test sets respectively, while BioBERT-SDNet produced average F_1 scores of 21.9 and 34.0. In general, the BioBERT model outperformed the BioBERT-SDNet model on every quantitative metric. It may be the case that SDNet as designed is primarily beneficial to conversational QA in which previous questions and answers play an important role in context. However, the individual methods that SDNet incorporates intuitively seem to be desirable for the task. In particular, the feature vector that SDNet utilizes includes a part-of-speech embedding, a named entity recognition embedding, an exact matching vector, and a normalized term frequency vector, all aspects that have been common to many of the most successful BioASQ challenge models in the past.

Upon performing additional studies on the model outputs, we found that the distinction in quality was not so clear. BioBERT produced answers that completely deviated from the golden answers 48.57% of the time, while BioBERT-SDNet produced answers with zero alignment for only 42.86% of the questions. Furthermore, in many cases, BioBERT-SDNet answers were qualitatively preferable. We compared the answers generated by the two models and observed that while BioBERT generally produced answers that were lexically similar to the golden answers, oftentimes these responses were not semantically appropriate or syntactically complete. In contrast, BioBERT-SDNet more consistently yielded answers that, even if incorrect, fit well with the question and golden answer structure. We provide question/answer examples in the appendix 8.5 that illustrate these distinctions. This qualitative difference observed might be the result of the inclusion of the feature vector in the model which helps ensure that the output answer has the desired part of speech and named entity status. As SDNet is a complex architecture with many internal mechanisms, an ablative study testing these various mechanisms on non-conversational QA tasks could help determine which part or parts of the model are perhaps better suited to conversational QA and may be removed from the model, and which provide the qualitative improvements mentioned here.

Limitations within the dataset. We note that both the BioBERT and BioBERT-SDNet models perform worse on BioASQ4b than 5b. The difference in training data size is a likely culprit. It is notable that when testing the default baseline model on SQuAD challenge data, the F_1 is much higher than the F_1 score achieved on the BioASQ test sets with the same model. Since the question/answer structure between SQuAD and BioASQ is very similar, it is likely that this difference in performance results from the small training size for BioASQ data rather than any fundamental difference in model applicability between the two tasks. As observed in past leaderboard results, neural models generally have not performed as well on BioASQ as rule-based models, and this is likely due in large part to the increased dependency neural models have on training data over rule-based models. In lieu of an expansion of the BioASQ challenge training data, pre-training on a similar dataset such as SQuAD may help alleviate this problem.

6 Conclusion

We have achieved competitive results on biomedical question answering tasks with our implementation of a combined BioBERT-SDNet model. In comparing with the recently published BioBERT, we found that while BioBERT-SDNet does not perform quite as well quantitatively, it offers some significant qualitative improvements in answer generation that mark it as a promising area for future study. It is particularly noteworthy that the model was able to achieve competitive results while training only on very small datasets. Though beyond the scope of this effort, pre-training on large datasets such as SQuAD has yielded large performance gains for similar models in the literature (e.g. [2]). Therefore, an immediate next step in improving our model would be to conduct pre-training. BioASQ recently released the training data for the 2019 challenge, and we will test our model on this new dataset, the largest yet. After pre-training on SQuAD and experimenting further with parameter tuning, we will submit to the challenge to compare against state-of-the-art models. We expect that with this further work, the model can be a powerful tool for biomedical question answering in practice.

7 Acknowledgements

We thank our project mentor Suvadip Paul for his tremendous help and inspiring advice. We thank Xiaoxue Zang for her input into our research direction. We thank Professor Chris Manning and the CS224n TAs for their instructions throughout the course.

References

[1] C. Zhu, M. Zeng, and X. Huang, "Sdnet: Contextualized attention-based deep network for conversational question answering," *arXiv preprint arXiv:1812.03593*, 2018.

[2] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *CoRR*, vol. abs/1901.08746, 2019.

[3] N. L. of Medicine, "Citations added to medline by fiscal year," 2018. [Online; accessed 12-February-2019].

[4] G. Tsatsaronis, G. Balikas, P. Malakasiotis, I. Partalas, M. Zschunke, M. R. Alvers, D. Weissenborn, A. Krithara, S. Petridis, D. Polychronopoulos, *et al.*, "An overview of the bioasq large-scale biomedical semantic indexing and question answering competition," *BMC bioinformatics*, vol. 16, no. 1, p. 138, 2015.

[5] A. Nentidis, A. Krithara, K. Bougiatiotis, G. Paliouras, and I. Kakadiaris, "Results of the sixth edition of the bioasq challenge," pp. 1–10, 2018.

[6] K. Chandu, A. Naik, A. Chandrasekar, Z. Yang, N. Gupta, and E. Nyberg, "Tackling biomedical text summarization: Oaqa at bioasq 5b," in *BioNLP 2017*, pp. 58–66, Association for Computational Linguistics, 2017.

[7] A. Nentidis, K. Bougiatiotis, A. Krithara, G. Paliouras, and I. Kakadiaris, "Results of the fifth edition of the bioasq challenge," *BioNLP 2017*, pp. 48–57, 2017.

[8] D. Weissenborn, G. Wiese, and L. Seiffe, "Fastqa: A simple and efficient neural architecture for question answering," *CoRR*, vol. abs/1703.04816, 2017.

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

[10] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

[11] H.-Y. Huang, C. Zhu, Y. Shen, and W. Chen, "Fusionnet: Fusing via fully-aware attention with application to machine comprehension," in *International Conference on Learning Representations*, 2018.

[12] S. Reddy, D. Chen, and C. D. Manning, "Coqa: A conversational question answering challenge," *arXiv preprint arXiv:1808.07042*, 2018.

[13] Evaluation-Measures, "Evaluation measures for bioasq yearly challenge." [Online; https://github.com/BioASQ/Evaluation-Measures].

[14] H. M. T. S. Sampo Pyysalo, Filip Ginter and S. Ananiadou, "Distributional semantics resources for biomedical text processing," *LBM*, 2013.

[15] Google, "Bert." [Online].

[16] Microsoft, "Naver." [Online; https://github.com/naver/biobert-pretrained].

[17] dmis lab, "Biobert." [Online; https://github.com/dmis-lab/biobert].

[18] Microsoft, "Sdnet." [Online; https://github.com/Microsoft/SDNet].

[19] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

8 Appendix

8.1 Generating model outputs for factoid questions

We generate predicted answer span by computing the probability $\boldsymbol{p}_i^S$ that the answer span should start at the ith context word:

$$\boldsymbol{p}_i^S = \text{softmax}\left((\mathbf{u}^Q)^T \mathbf{W}_S \mathbf{u}_i^C\right)$$

where $\mathbf{W}_S$ is a parametrized matrix. We apply a GRU to obtain

$$\boldsymbol{t}^Q = \text{GRU}\left(\mathbf{u}^Q, \sum_{i=1}^{m} P_i^S \mathbf{u}_i^C\right)$$

and from this compute the probability $\boldsymbol{p}_i^E$ that the answer span should end at the ith context word:

$$\boldsymbol{p}_i^E = \text{softmax}\left((\boldsymbol{t}^Q)^T \mathbf{W}_E \mathbf{u}_i^C\right)$$

where $\mathbf{W}_E$ is another parametrized matrix. We return the five highest probability spans as per the BioASQ challenge.

8.2 BioASQ to CoQA conversion example

BioASQ format

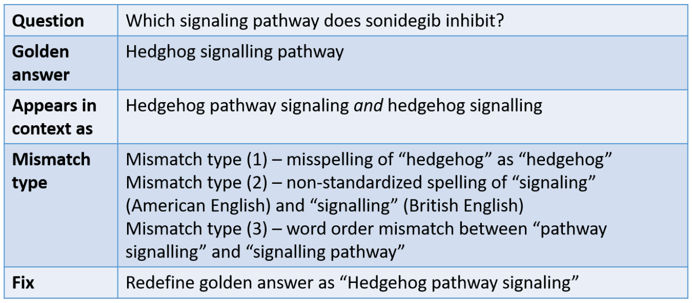
```
{"question":
"What is the inheritance pattern of Li\u2013Fraumeni syndrome?",
"answers": [{"text": "autosomal dominant", "answer_start": 474}]}
```

BioASQ in CoQA format

```
"questions": [
{
"input_text":
"What is the inheritance pattern of Li\u2013Fraumeni syndrome?",
"turn_id": 1
}
],
"answers": [
{
"input_text": "autosomal dominant",
"span_text": "autosomal dominant pattern of inheritance. Familial co",
"span_start": 474,
"span_end": 528,
"turn_id": 1
}
```

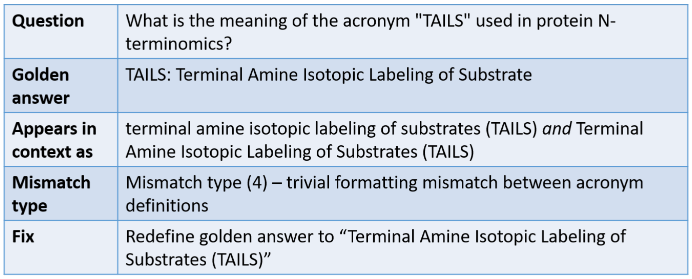
8.3 Context-answer mismatching

Examples of question-answer pairs that have context answer mismatches, as shown in Figure 3-6. We define mismatch types including (1) misspellings, (2) non-standardized spellings, (3) word order mismatch, (4) formatting mismatch, (5) inference required, and (6) unsuitable answer provided due to over-abstraction.



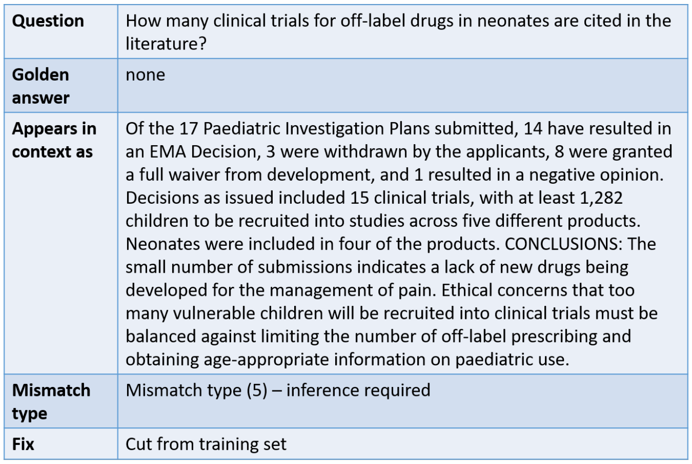
Question	Which signaling pathway does sonidegib inhibit?
Golden answer	Hedghog signalling pathway
Appears in context as	Hedgehog pathway signaling *and* hedgehog signalling
Mismatch type	Mismatch type (1) – misspelling of “hedgehog” as “hedgehog” Mismatch type (2) – non-standardized spelling of “signaling” (American English) and “signalling” (British English) Mismatch type (3) – word order mismatch between “pathway signalling” and “signalling pathway”
Fix	Redefine golden answer as “Hedgehog pathway signaling”

Figure 3: Example question from the BioASQ7b training dataset demonstrating context-answer mismatching resulting in infeasibility for extractive QA. This example, and others like it, was corrected by hand and added back to the training set.



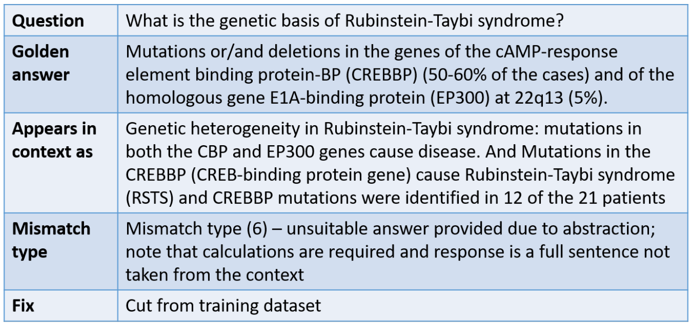
Question	What is the meaning of the acronym "TAILS" used in protein N-terminomics?
Golden answer	TAILS: Terminal Amine Isotopic Labeling of Substrate
Appears in context as	terminal amine isotopic labeling of substrates (TAILS) *and* Terminal Amine Isotopic Labeling of Substrates (TAILS)
Mismatch type	Mismatch type (4) – trivial formatting mismatch between acronym definitions
Fix	Redefine golden answer to “Terminal Amine Isotopic Labeling of Substrates (TAILS)”

Figure 4: Example question from the BioASQ7b training dataset demonstrating context-answer mismatching due to formatting inconsistencies resulting in infeasibility for extractive QA. This example, and others like it, was corrected by hand and added back to the training set.



Question	How many clinical trials for off-label drugs in neonates are cited in the literature?
Golden answer	none
Appears in context as	Of the 17 Paediatric Investigation Plans submitted, 14 have resulted in an EMA Decision, 3 were withdrawn by the applicants, 8 were granted a full waiver from development, and 1 resulted in a negative opinion. Decisions as issued included 15 clinical trials, with at least 1,282 children to be recruited into studies across five different products. Neonates were included in four of the products. CONCLUSIONS: The small number of submissions indicates a lack of new drugs being developed for the management of pain. Ethical concerns that too many vulnerable children will be recruited into clinical trials must be balanced against limiting the number of off-label prescribing and obtaining age-appropriate information on paediatric use.
Mismatch type	Mismatch type (5) – inference required
Fix	Cut from training set

Figure 5: Example question from the BioASQ7b training dataset demonstrating context-answer mismatching resulting in infeasibility for extractive QA as inference is required. This example, and others like it, was cut from the training set.



Question	What is the genetic basis of Rubinstein-Taybi syndrome?
Golden answer	Mutations or/and deletions in the genes of the cAMP-response element binding protein-BP (CREBBP) (50-60% of the cases) and of the homologous gene E1A-binding protein (EP300) at 22q13 (5%).
Appears in context as	Genetic heterogeneity in Rubinstein-Taybi syndrome: mutations in both the CBP and EP300 genes cause disease. And Mutations in the CREBBP (CREB-binding protein gene) cause Rubinstein-Taybi syndrome (RSTS) and CREBBP mutations were identified in 12 of the 21 patients
Mismatch type	Mismatch type (6) – unsuitable answer provided due to abstraction; note that calculations are required and response is a full sentence not taken from the context
Fix	Cut from training dataset

Figure 6: Example question from the BioASQ7b training dataset demonstrating context-answer mismatching resulting in infeasibility for extractive QA as the golden answer provided requires significant abstraction. This example, and others like it, was cut from the training set.

8.4 Qualitative analysis of predictions

Question 1: Where are Paneth cells located?
Golden: in the intestinal crypt base columnar cells
BioBERT-SDNet: Intestinal stem cells
BioBERT: The intestinal epithelium is a classic example of a rapidly self-renewing tissue fueled by dedicated resident stem cells.

Question 2: Which disease is treated with Eliglustat?
Golden: Gaucher's disease type 1
BioBERT-SDNet: Gaucher's disease type 1
BioBERT: glucosylceramide synthase

We notice that even though answers provided by BioBERT-SDNet did not match exactly with the golden answer in **Question 1**, the answer does capture the information included in the golden answer, and moreover is semantically and syntactically appropriate. In contrast, the output of BioBERT does not make sense though it is lexically similar to the golden answer and has significant overlap in word content. In **Question 2**, BioBERT-SDNet output an exact match to the golden answer while BioBERT yielded a clearly wrong answer. We have found that BioBERT-SDNet usually produces exact match when the golden answers are short, and captures partially the golden answer when the ideal answer is long.

8.5 SDNet loss

Loss converging over 30 epochs for BioASQ 4b and 5b challenges.

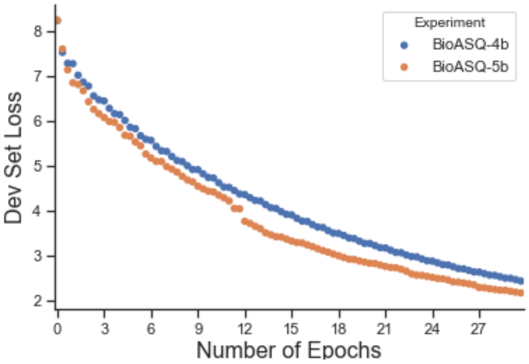


Figure 7: BioBERT-SDNet Loss on BioASQ dev sets over training epochs.

8.6 Further results reference tables

	Test 1			Test 2			Test 3			Test 4			Test 5		
	S	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**
BioASQ baseline	0.10	0.15	0.12	0.06	0.16	0.11	0.12	0.19	0.14	0.03	0.19	0.08	0.06	0.12	0.08
OAQA	-	-	-	-	-	-	0.23	0.27	0.24	0.29	0.39	0.33	0.21	0.29	0.29
LabZhu	-	-	-	0.19	0.26	0.23	0.19	0.27	0.22	0.10	0.19	0.14	0.18	0.33	0.25
BioBERT	0.15	0.31	0.21	0.06	0.26	0.13	0.12	0.19	0.14	0.29	0.32	0.31	0.12	0.18	0.14
SDNet	0.10	0.15	0.12	0.13	0.16	0.13	0.04	0.14	0.22	0.26	0.34	0.30	0.12	0.18	0.15

Table 7: Results evaluated on BioASQ4b test sets for factoid questions. Strict accuracy S, lenient accuracy L, and mean reciprocal rank M are reported. Results can be accessed at the BioASQ webpage.[1]

	Test 1			Test 2			Test 3			Test 4			Test 5		
	S	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**
BioASQ Baseline	0.28	0.40	0.33	0.16	0.35	0.22	0.11	0.27	0.19	0.03	0.12	0.07	0.06	0.20	0.12
DeepQA Ensemble	0.56	0.68	0.60	0.38	0.52	0.44	0.31	0.62	0.42	0.33	0.55	0.42	0.26	0.51	0.35
LabZhu	0.40	0.44	0.42	0.32	0.48	0.38	0.35	0.42	0.38	0.27	0.45	0.35	0.40	0.51	0.45
Default Baseline	0.21	-	-	0.17	-	-	0.14	-	-	0.08	-	-	0.10	-	-
BioBERT	0.28	0.44	0.34	0.23	0.35	0.27	0.42	0.50	0.46	0.18	0.33	0.23	0.20	0.37	0.26
SDNet	0.28	0.35	0.31	0.19	0.23	0.19	0.19	0.25	0.20	0.09	0.16	0.12	0.14	0.20	0.13

Table 8: Results evaluated on BioASQ5b test sets for factoid questions. Strict accuracy S, lenient accuracy L, and mean reciprocal rank M are reported. Results can be accessed at the BioASQ webpage.[2]