BioNumQA-BERT: Answering Biomedical Questions Using Numerical Facts with a Deep Language Representation Model

Ye Wu
Department of Computer Science
The University of Hong Kong
Hong Kong, China
ywu@cs.hku.hk

Hing-Fung Ting
Department of Computer Science
The University of Hong Kong
Hong Kong, China
hfting@cs.hku.hk

Tak-Wah Lam
Department of Computer Science
The University of Hong Kong
Hong Kong, China
twlam@cs.hku.hk

Ruibang Luo
Department of Computer Science
The University of Hong Kong
Hong Kong, China
rbluo@cs.hku.hk

ABSTRACT

Biomedical question answering (QA) is playing an increasingly significant role in medical knowledge translation. However, current biomedical QA datasets and methods have limited capacity, as they commonly neglect the role of numerical facts in biomedical QA. In this paper, we constructed BioNumQA, a novel biomedical QA dataset that answers research questions using relevant numerical facts for biomedical QA model training and testing. To leverage the new dataset, we designed a new method called BioNumQA-BERT by introducing a novel numerical encoding scheme into the popular biomedical language model BioBERT to represent the numerical values in the input text. Our experiments show that BioNumQA-BERT significantly outperformed other state-of-art models, including DrQA, BERT and BioBERT (39.0% vs 29.5%, 31.3% and 33.2%, respectively, in strict accuracy). To improve the generalization ability of BioNumQA-BERT, we further pretrained it on a large biomedical text corpus and achieved 41.5% strict accuracy. BioNumQA and BioNumQA-BERT establish a new baseline for biomedical QA. The dataset, source codes and pretrained model of BioNumQA-BERT are available at https://github.com/LeaveYeah/BioNumQA-BERT.

KEYWORDS

Text Mining, Biomedical Question Answering, BERT, Numerical encoding

BCB '21, August 1–4, 2021, Virtual Event, USA

ACM ISBN 978-1-4503-7964-9/20/09…$15.00
https://doi.org/10.1145/3388440.3412457

ACM Reference format:

Ye Wu, Hing-Fung Ting, Tak-Wah Lam, and Ruibang Luo. 2021. BioNumQA-BERT: Answering Biomedical Questions Using Numerical Facts with a Deep Language Representation Model. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB'21). ACM-BCB,* August 1–4, 2021, Virtual Event, USA, 7 pages. https://doi.org/10.1145/3388440.3412457

1 Introduction

Biomedical literature is an important source of knowledge for developing better diagnoses and treatments [4, 5]. However, the enormous volume of biomedical literature poses a significant challenge for knowledge discovery and translation. MEDLINE [6], the public bibliographic database for biomedical literature, had over 26 million articles in 2020. Automatic question answering (QA) can save medical professionals a great deal of time and energy by linking their research questions to the most relevant biomedical research results [7]. Developing an intelligent QA system has become a long-term goal for medical artificial intelligence.

Recently, several datasets were introduced to biomedical QA [8-10]. These datasets consisted mostly of questions posed on a set of biomedical research articles, where the answer to every question is binary (yes or no), or a segment of text from the corresponding reading passage. BioASQ [11] and PubMedQA [12], are the two widely accepted human-annotated QA datasets for benchmarking biomedical natural language processing (NLP) [13]. BioASQ, the largest annotated biomedical QA dataset, is composed mainly of simple factoid questions. PubMedQA extends BioASQ by posing research-oriented questions. PubMedQA selects PubMed abstracts titled with yes/no questions, such as "Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?" [3]. The task is to answer such questions with yes/no/maybe from the corresponding abstracts without using the conclusion section. PubMedQA is more challenging than BioASQ, as it hides the

straightforward answer (the conclusion section) and requires reasoning over the context of the abstract to get the answer.

Solving these QA tasks demands strong understanding of biomedical language to analyze the questions and corresponding contexts. Recently, deep contextualized language representation models have demonstrated powerful language understanding capability [14-16]. In particular, BERT [14], which applies bidirectional training of Transformer [17] encoders, achieved significant improvements on several NLP benchmarks [18-20]. Applying the attention mechanism [21], the Transformer encoder reads a sequence of words all at once, instead of sequentially from left to right. This characteristic allows BERT to learn the contextual representation of a word based on all of its surroundings (left and right). BERT obtains general language understanding capability through pretraining on large text corpora, such as Wikipedia articles. It is then applied to end tasks such as QA through a fine-tuning approach. In the biomedical domain, BioBERT [22] further pretrains BERT on a large corpus of PubMed articles to enhance its understanding of biomedical language. It outperforms BERT and achieves state-of-art performance on BioASQ and PubMedQA tasks.

Question: Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting?
Context:
OBJECTIVE: Recent studies have demonstrated that statins have pleiotropic effects, including anti-inflammatory effects and atrial fibrillation (AF) preventive effects [...]
RESULTS: The overall incidence of postoperative AF was 26%. *Postoperative AF was significantly lower in the Statin group compared with the Non-statin group (16% versus 33%, p=0.005).* Multivariate analysis demonstrated that independent predictors of AF [...]

BioNumQA Answer: Highlighted in red
PubMedQA Answer: Yes

Figure 1: An instance of BioNumQA curated from Sakamoto et al. [3]. The BioNumQA answer (highlighted in red) is compared to PubMedQA.

However, the current biomedical QA datasets and BERT-based methods are both limited, as they do not consider the important role of numerical facts in biomedical QA. Numerical facts are essential for comprehensive and informative answers to biomedical research

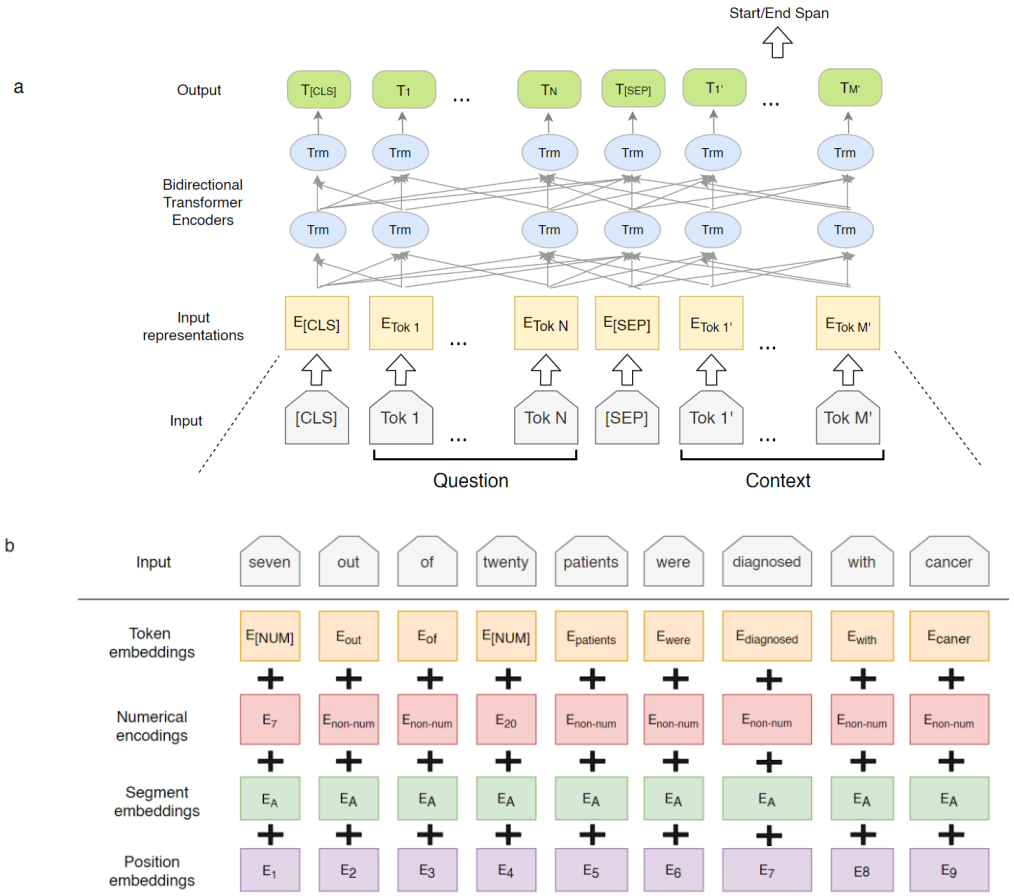

Figure 2: (a) Network structure of BioNumQA-BERT. (b) Input representation of BioNumQA-BERT for the sentence "seven out of twenty patients were diagnosed with cancer".

questions. Datasets such as BioASQ and PubMedQA collect mostly short and simple answers to questions, which are often insufficient for medical professionals. For example, if medical practitioners ask the question "Is the Pfizer vaccine effective for COVID-19?", they would not be satisfied with a simple "yes". Instead, they are seeking numerical facts, such as the effective rate and p-value for vaccine experiments. Further information, such as the number and age of participants, and severity distribution in the vaccine and placebo groups would also be important for their decision making. An intelligent biomedical QA system should provide the necessary empirical evidence to the users and leave the final judgement to them.

BioBERT and other BERT-based methods do not effectively capture numerical facts, as they cannot properly represent numerical values in the input texts. In BERT-based models, numbers are represented in the same way as words, whose sematic information is coded in the word embeddings, which are obtained through the pretraining process according to each word's surrounding context in the text corpora. Therefore, the numbers are represented not by their value or magnitude, but by their surrounding context. It is difficult for BioBERT to distinguish "0.01" and "0.02" as they are used in similar contexts. Designed experiments by Naik et al. and Wallace et al. [23, 24] also show that BERT-based models are inadequate to capture the magnitude of numbers in the text. Effectively extracting numerical facts from the literature requires proper representation of numerical values.

For this paper, we constructed a novel biomedical QA dataset, BioNumQA, which answers questions using the relevant numerical facts in the corresponding context. Figure 1 shows an instance of BioNumQA. Unlike PubMedQA, the answer of BioNumQA is based on numerical evidence, which is more informative for medical professionals. To fully leverage the numerical power of BioNumQA, we designed BioNumQA-BERT, a biomedical language representation model specially optimized to work with numerical values. Building upon BioBERT, we improved the numerical reasoning capability by adding extra numerical encoding to represent the value of numbers in the text. Our experiments show that BioNumQA-BERT outperforms BioBERT significantly in both strict accuracy (39.0% vs. 33.2%) and lenient accuracy (66.2% vs. 61.5% for the top 3 answers and 78.7% vs. 45.3% for the top 5 answers). To further improve BioNumQA-BERT, we pretrained it on a large-scale biomedical corpus, resulting in a 2.5% performance increase (from 39.0% to 41.5%) in strict accuracy. The dataset, source codes and pretrained model of BioNumQA-BERT are available at https://github.com/LeaveYeah/BioNumQA-BERT.

2 Materials and Methods

2.1 Definition and Criteria of BioNumQA

The task of BioNumQA is to answer biomedical questions with the most relevant numerical facts from corresponding articles. Each BioNumQA instance comprises (a) a biomedical research question, (b) an article, and (c) an annotated answer which is a segment of text from the context. Each annotated answer comprises one or more complete sentences. In the example shown in Fig. 1, there is only one annotated answer to the question. However, there are cases with multiple alternative answers to a question. We set the following criteria for multiple answers:

(1) *Each fact can separately answer the question*. If for a question, q, both fact a and fact b can answer the question, then the answer to q, can be either a, b, or a combination of a and b, i.e., $a + b$.
(2) *One fact can supplement another to answer the question*. If for a question, q, fact a can answer the question and fact b cannot answer the question but can supplement a, then the answer to q can be either a or $a + b$.

For detailed examples, refer to the Supplementary Notes.

2.2 Collection of BioNumQA Dataset

We reused 600 questions originally answered with yes/no from BioASQ [11] and PubMedQA [12], and reannotated them according to the BioNumQA criteria. We selected yes/no questions because straightforward yes/no answers for these questions are not sufficient for medical professionals. We provide more informative and comprehensive answers by annotating the corresponding numerical facts.

Questions are answered from the corresponding PubMed abstracts. Each abstract is structured with a clear results section, which describes the experiment and numerical findings. As with PubMedQA, the conclusion sections, which provide straightforward answers to the questions, are excluded from the input to reduce noise in the input.

2.3 Evaluation Metrics

For each question, a QA model is expected to return a list of up to 5 answers ordered in decreasing level of confidence. As with BioASQ [11], we use strict accuracy (SAcc) and lenient accuracy (LAcc) to evaluate the performance of the QA model. Strict accuracy counts a question as correctly answered if the first answer on the list matches the annotated answer. In contrast, lenient accuracy counts a question as correctly answered if one of the top 3 or top 5 answers on the list matches the annotated answer. Strict accuracy and lenient accuracy are calculated as a percentage of correctly answered questions among all of the questions.

Each experiment on BioNumQA was conducted with 10-fold cross-validation, i.e., repeated 10 times with 90% of the data (randomly selected) for training and 10% of the data for validation.

2.3 Structure of BioNumQA-BERT

BioNumQA-BERT is based on the successful implementation of BioBERT [22] for the BioASQ task [11]. BioNumQA-BERT adopts a multi-layer bidirectional Transformer encoder architecture similar to BioBERT, except for the input representations. For the detailed implementation of Transformer, refer to the description in Rajpurkar et al. [17]. We describe the application of BioNumQA-BERT on the BioNumQA task as follows (Figure 2a):

Given a question $q = (Tok\ 1, Tok\ 2, \dots, Tok\ N)$ asking for an answer from context $c = (Tok\ 1', Tok\ 2', \dots, Tok\ M')$, we formulate the input as a packed sequence $x =$

$([CLS], Tok\ 1, Tok\ 2, \dots, Tok\ N, [SEP], Tok\ 1', Tok\ 2', \dots, Tok\ M')$, where $[CLS]$ denotes the start of the sequence and $[SEP]$ separates q and c. Let InputR(·) be the input representation layer (see section E for a detailed description). We first obtain the input representation as $l = InputR(x) \in \mathbb{R}^{r_h * |x|}$, where $|x|$ is the length of the input sequence and r_h is the size of the hidden dimension. Let MultiTrm(·) be the multi-layer Transformer encoders. We then obtain the hidden representation as $T = MultiTrm(l) \in \mathbb{R}^{r_h * |x|}$. Finally, we apply a SoftMax layer to calculate each word's probability of being the start or end of the answer. We introduce two trainable vectors, a start vector $S \in \mathbb{R}^{r_h}$ and an end vector $E \in \mathbb{R}^{r_h}$ during finetuning. The probability of word i being the start of the answer span is computed as $P_i = \frac{e^{S \cdot T_i}}{\sum_k e^{S \cdot T_k}}$. The probability for the end of the answer is computed in an analogous formula. The probability of position i and j being the start and end of the answer, respectively, can thus be calculated as $P_{i,j} = \frac{e^{S \cdot T_i}}{\sum_k e^{S \cdot T_k}} * \frac{e^{E \cdot T_j}}{\sum_k e^{E \cdot T_k}} \propto S \cdot T_i + E \cdot T_j$. Therefore, we define the score for a candidate answer span from position i to position j as $S \cdot T_i + E \cdot T_j$. The answer span with the maximum score where $j \geq i$ is used as a prediction. The training loss is the sum of the log-likelihoods of the correct start and end positions.

We implemented BioNumQA-BERT using PyTorch [25]. We adopted the $\text{BERT}_{\text{BASE}}$ parameter setting with 768 hidden dimensions, 12 attention heads in transformers and 12 layers. The total number of parameters was 110M. In each experiment, BioNumQA-BERT was initialized with the pretrained weights of BioBERT-v1.0 [22].

2.4 Input representations of BioNumQA-BERT

The original input representations of BioBERT [22] comprised token embeddings, position embeddings, and segment embeddings: token embeddings represent the semantic meanings of each word; position embeddings represent the position of each token in a sequence; and segment embeddings represent the roles, e.g., <Question, Context>, for a pair of word sequences. In BioNumQA-BERT, numerical encodings are added to the input representations to encode the value of the input numbers, as shown in Fig. 2b.

First, we replace the numbers in the input tokens with a uniform token, "[NUM]". Since token embeddings cannot properly capture the different magnitudes of numbers, we normalize the representations of numbers in the token embeddings. The numbers in the text, in both word and digit form, are detected using regular expression and Named Entity Recognition modules of Scispacy [26]. As shown in the example sentence in Figure 2a, "seven out of twenty patients are diagnosed with cancer", the numbers "seven" and "twenty" are detected in the text and replaced with "[NUM]" in the token embeddings.

Then, we add "numerical encodings" to the input embeddings to inject the information about each number's magnitude. Since there are infinite values of numbers, we cannot represent each value with a trainable embedding similar to other input representations. We need to use an "encoding" function to map a numerical value to the input vector space. We found that more than 99.9% of the numbers in biomedical literature lie between 10^{-5} and 10^{5}, so to solve BioNumQA, we focused on numerical encodings within this range. Numbers beyond this range were mapped to the upper or lower limits. For other applications, this range can be adjusted accordingly.

In this paper, we experimented with two different schemes for numerical encodings, single-dimension (SD) and multi-dimension (MD) encoding, described as follows:

(1) *SD numerical encoding*. This represents a numerical value using a single number in the input vector. It refers to the idea of scientific notation of numbers. We first convert a numerical value to the format $C * 10^n$, where $0 < C < 1$ and n is an integer range from -4 to 5. We use the first 10 dimensions in the input vector to represent $10^{-4}, 10^{-3}, \dots, 10^{5}$. A numerical value is represented by C in the corresponding 10^n dimension, as shown in Figure 3a.

(2) *MD numerical encoding*. This represents a numerical value using multiple numbers in the input vector. We referred to the idea of positional encoding in the Transformer [15]. But instead of encoding relative positions like the Transformer, we encode the absolute value of numbers. The function of MD encoding is as follows:

$$NE(num, i) = \begin{cases} \frac{num}{100{,}000^{2i/d_h}}, if\ 0 < \frac{num}{100{,}000^{2i/d_h}} < 1 \\ 0, \ otherwise \end{cases}$$

where num is the numerical value, and i is the dimension, and d_h is the total number of input dimensions. The distribution of MD encoding for numbers ranging from 10^{-5} to 10^{5} is shown in Figure 3b. We can see that a larger value is encoded by numbers in the higher dimensions. There is a linear relationship between the logarithm of the encoded value and the non-zero dimensions it represents in the encoding.

0	0	0	0	0	0.9	0	0	0	0	0
10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}	10^{5}	[...]

Figure 3a: Single-dimension encoding of number 9.

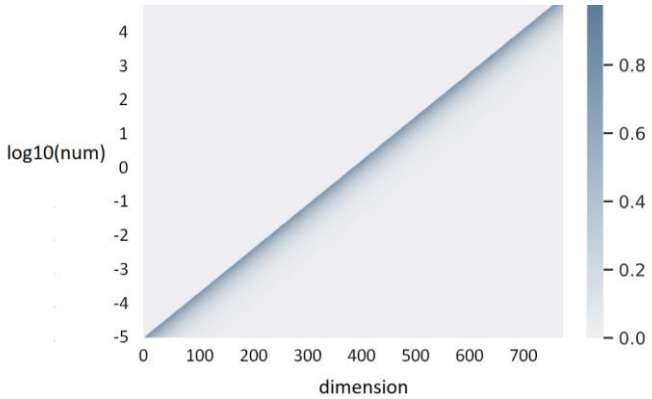

Figure 3b: Heatmap for MD numerical encoding

2.5 Pretraining of BioNumQA-BERT

Pretraining is used to increase generalization capability and adapt BERT-based models with domain knowledge for specific end tasks [27]. The limited data of BioNumQA may not be enough for the Transformer encoders to fully learn the contextualized representations of numerical encodings. Therefore, we introduced a pretraining step for BioNumQA-BERT to further increase its reasoning capability over numerical facts.

To pretrain BioNumQA-BERT, we adopted the masked language model training objective from BERT [14]. The masked language model predicts randomly masked words in a sequence, so it can be used for learning bidirectional representations from our newly input numerical representations. We initialized BioNumQA-BERT with the weights of BioBERT [22], which was already pretrained on Wikipedia articles and PubMed articles for both general and biomedical language understanding. To further pretrain BioNumQA-BERT, we used PubMed abstracts structured similar to the contexts of BioNumQA. The text corpus consisted of 4.3 million PubMed abstracts with 1.7 billion tokens in total, downloaded from PubMed Central open access subset [28] in 11/2020. More settings and hyperparameters are available in the Supplementary Notes.

3 Results

3.1 Comparison between BioNumQA-BERT and existing methods

We compared BioNumQA-BERT with three state-of-art methods for question answering: DrQA [29], BERT [14], and BioBERT [22]. DrQA uses a multi-layer bidirectional long short-term memory network (LSTM) [30] to detect answers from contexts. BERT and BioBERT both apply training of multi-layer Transformer encoders with the same parameter setting ($BERT_{BASE}$). BioBERT was pretrained on PubMed articles to adapt it to the biomedical domain. All methods were directly fine-tuned on the BioNumQA dataset.

The experiment results are shown in Table 1. BioBERT outperformed DrQA and BERT by a significant margin in lenient accuracy (more than 5%). This indicates that biomedical domain knowledge helps BioBERT pinpoint a general area for the correct answer. In contrast, BioBERT did not show a significant advantage over BERT in strict accuracy (33.2% vs. 31.3%).

BioNumQA-BERT outperformed BioBERT in both strict and lenient accuracy. BioNumQA-BERT's advantage was most significant on strict accuracy (39.0% vs. 33.2%). This indicates that the injection of numerical information helps BioNumQA-BERT better locate the exact numerical facts to answer the questions. As shown in the example in Figure 4, BioNumQA-BERT accurately found a numerical fact connected with the question, while BioBERT pointed to a loosely related fact. With the introduction of numerical encodings, BioNumQA-BERT is empowered with stronger numerical reasoning capability. The results show that BioNumQA-BERT with SD and MD numerical encodings achieved a similar performance. This indicates that the two encoding schemes are equally effective in encoding numerical information.

Table 1: Performance comparison of models

Model	SAcc	LAcc-top 3	LAcc-top 5
DrQA	29.5%	55.8%	68.8%
BERT	31.3%	55.5%	69.5%
BioBERT	33.2%	61.5%	75.2%
BioNumQA-BERT w/ S.D.	38.8%	65.8%	**78.7%**
BioNumQA-BERT w/ M.D.	**39.0%**	**66.2%**	78.3%

SAcc: Strict Accuracy; LAcc: Lenient Accuracy; SD: Singe-dimension numerical encoding; MD: dimension numerical encoding

Question: Does a geriatric oncology consultation modify the cancer treatment plan for elderly patients?
Context:
[..] *Of the 93 patients with an initial treatment decision, the treatment plan was modified for 38.7% of cases after this assessment. Only body mass index and the absence of depressive symptoms were associated with a modification of the treatment plan.* [...]
BioNumQA-BERT Answer: Highlighted in Blue
BioBERT Answer: Highlighted in Green

Figure 4: Example comparing BioNumQA-BERT and BioBERT answers.

3.2 Comparison of BioNumQA-BERT models with different pretraining settings

We compared the performance of BioNumQA-BERT with and without pretraining. We pretrained BioNumQA-BERT with each numerical encoding on a text corpus of 4.3M PubMed abstracts for 300k steps. It took 4 NIVIDA RTX 2080Ti GPU cards about 140 hours to complete each pretraining session.

The results are shown in Table 2. Pretraining significantly improved the performance of BioNumQA-BERT, especially in strict accuracy (2.5% increase for both SD and MD). This indicates that pretraining with large scale data is effective in improving the generalization capability of numerical encodings. The MD encoding provides a small advantage over SD encoding. We plotted the performance of BioNumQA-BERT w/ MD under different pretraining steps (Figure 5). The performance of BioNumQA-BERT did not immediately increase after pretraining, and it decrease slightly after 100k steps. This indicates that the model did not yet reach a stabilized status. The model performance increased afterwards with a slower speed. Further increasing the pretraining steps may not achieve a significant improvement.

Table 2: Performance of BioNumQA-BERT with Pretraining

Model	Pretraining	SAcc	LAcc-top 3	LAcc-top 5
BioNumQA-BERT w/ SD	No	38.8%	65.8%	78.7%
	Yes	41.3%	66.7%	79.2%
BioNumQA-BERT w/ MD	No	39.0%	66.2%	78.3%
	Yes	**41.5%**	**67.3%**	**79.3%**

SAcc: Strict Accuracy; LAcc: Lenient Accuracy; SD: Singe-dimension numerical encoding; MD: Multi-dimension numerical encoding

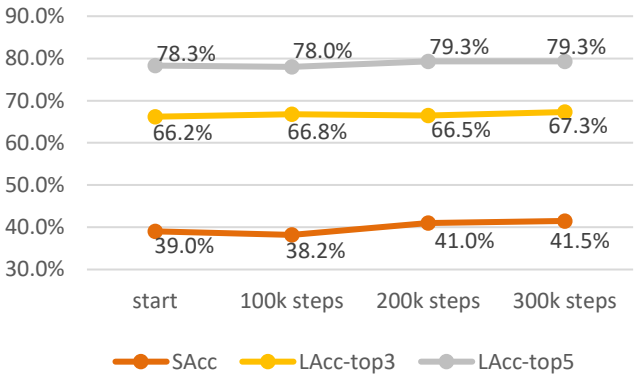

Figure 5: Performance of BioNumQA-BERT with different pretraining steps

4 Conclusion

In this paper, we constructed a novel dataset, BioNumQA, for answering biomedical research questions with numerical facts. Compared with existing biomedical QA datasets, BioNumQA provides more informative and comprehensive answers to research questions. To leverage the power of BioNumQA, we improved the numerical reasoning capability of BioBERT by adding a numerical encoding scheme. Our experiments showed that our method, BioNumQA-BERT, outperformed BioBERT by a significant margin, thus establishing a new baseline for biomedical QA. We further improved the performance of BioNumQA-BERT by pretraining it on a large biomedical corpus of PubMed abstracts. We demonstrated BioNumQA-BERT's numerical power using the biomedical QA task, but this idea should be appliable to all other numerical-sensitive biomedical text-mining tasks.

ACKNOWLEDGMENTS

R. L. was supported by the ECS (grant number 27204518) and TRS (grant number T21-705/20-N) of the HKSAR government. H. T. was supported by HKU's URC fund (grant number 17208019).

REFERENCES

[1] Floch, N. R., Hinder, R. A., Klingler, P. J., Branton, S. A., Seelig, M. H., Bammer, T. and Filipi, C. J. Is laparoscopic reoperation for failed antireflux surgery feasible? *Arch Surg*, 134, 7 (Jul 1999), 733-737.
[2] Baune, M. B. and Aljeesh, Y. [Is pain a clinically relevant problem in general adult psychiatry? A clinical epidemiological cross-sectional study in patients with psychiatric disorders]. *Schmerz*, 18, 1 (Feb 2004), 28-37.
[3] Sakamoto, H., Watanabe, Y. and Satou, M. Do preoperative statins reduce atrial fibrillation after coronary artery bypass grafting? *Ann Thorac Cardiovasc Surg*, 17, 4 (2011), 376-382.
[4] Gonzalez, G. H., Tahsin, T., Goodale, B. C., Greene, A. C. and Greene, C. S. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Brief Bioinform*, 17, 1 (Jan 2016), 33-42.
[5] Wu, Y., Luo, R., Leung, H. C. M., Ting, H.-F. and Lam, T.-W. *RENET: A Deep Learning Approach for Extracting Gene-Disease Associations from Literature*. Springer International Publishing, City, 2019.
[6] Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., Skripchenko, Y., Wang, J., Ye, J., Trawick, B. W., Pruitt, K. D. and Sherry, S. T. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, 49, D1 (2021), D10-D17.
[7] Athenikos, S. J. and Han, H. Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99, 1 (2010/07/01/ 2010), 1-24.
[8] Pappas, D., Androutsopoulos, I. and Papageorgiou, H. *BioRead: A New Dataset for Biomedical Reading Comprehension*. European Language Resources Association (ELRA), City, 2018.
[9] Kim, S., Park, D., Choi, Y., Lee, K., Kim, B., Jeon, M., Kim, J., Tan, A. C. and Kang, J. A Pilot Study of Biomedical Text Comprehension using an Attention-Based Deep Neural Reader: Design and Experimental Analysis. *JMIR Med Inform*, 6, 1 (Jan 5 2018), e2.
[10] Esteva, A., Kale, A., Paulus, R., Hashimoto, K., Yin, W., Radev, D. and Socher, R. Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. *arXiv preprint arXiv:2006.09595* (2020).
[11] Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artiéres, T., Ngomo, A.-C. N., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I. and Paliouras, G. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16, 1 (2015/04/30 2015), 138.
[12] Jin, Q., Dhingra, B., Liu, Z., Cohen, W. and Lu, X. *PubMedQA: A Dataset for Biomedical Research Question Answering*. Association for Computational Linguistics, City, 2019.
[13] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J. and Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779* (2020).
[14] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Association for Computational Linguistics, City, 2019.
[15] Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. *Deep Contextualized Word Representations*. Association for Computational Linguistics, City, 2018.
[16] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[17] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. Attention is all you need. *arXiv preprint arXiv:1706.03762* (2017).
[18] Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W., Salakhutdinov, R. and Manning, C. D. *HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering*. Association for Computational Linguistics, City, 2018.
[19] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q. and Petrov, S. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7 (mar 2019), 452-466.
[20] Rajpurkar, P., Zhang, J., Lopyrev, K. and Liang, P. *SQuAD: 100,000+ Questions for Machine Comprehension of Text*. Association for Computational Linguistics, City, 2016.
[21] Bahdanau, D., Cho, K. and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).
[22] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H. and Kang, J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36, 4 (2019), 1234-1240.
[23] Naik, A., Ravichander, A., Rose, C. and Hovy, E. *Exploring Numeracy in Word Embeddings*. Association for Computational Linguistics, City, 2019.
[24] Wallace, E., Wang, Y., Li, S., Singh, S. and Gardner, M. *Do NLP Models Know Numbers? Probing Numeracy in Embeddings*. Association for Computational Linguistics, City, 2019.

[25] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N. and Antiga, L. Pytorch: An imperative style, high-performance deep learning library. *arXiv preprint arXiv:1912.01703* (2019).
[26] Neumann, M., King, D., Beltagy, I. and Ammar, W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *ArXiv*, abs/1902.07669 (2019).
[27] Xu, H., Liu, B., Shu, L. and Yu, P. S. BERT post-training for review reading comprehension and aspect-based sentiment analysis. *arXiv preprint arXiv:1904.02232* (2019).
[28] Roberts, R. J. PubMed Central: The GenBank of the published literature. *Proc Natl Acad Sci U S A*, 98, 2 (Jan 16 2001), 381-382.
[29] Chen, D., Fisch, A., Weston, J. and Bordes, A. *Reading Wikipedia to Answer Open-Domain Questions*. Association for Computational Linguistics, City, 2017.
[30] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Comput.*, 9, 8 (1997), 1735–1780.

Supplementary Note

Examples of multiple answers

(1) Each fact can separately answer the question

Question: Is laparoscopic reoperation for failed antireflux surgery feasible?

Context:
[..] (a) The well-being score (1 best; 10, worst) was 8.6+/-2.1 before and 2.9+/-2.4 after surgery (P<.001). (b) Thirty-one (89%) of 35 patients were satisfied with their decision to have reoperation [...]

BioNumQA Answer: (a) or (b) or (a)+(b)

Supplementary Figure 1: An instance of BioNumQA curated from Floch et al. [1].

(2) One fact can supplement another to answer the question

Question: Is pain a clinically relevant problem in general adult psychiatry?
Context:
[..] (a) The point prevalence of pain was about 50%, the 6-month prevalence 75.5% and the 12-month prevalence 76.5%. (b) The patients' most frequent complaints were low back pain, headache and shoulder and neck pain. [...]

BioNumQA Answer: (a) or (a)+(b)

Supplementary Figure 2: An instance of BioNumQA curated from Baune and Aljeesh [2].

Hyperparameters for BioNumQA-BERT

Supplementary Table 3: Hyperparameters of BioNumQA-BERT in experiments

Hyperparameters	Fine-tuning	Pretraining
Learning rate	5e-5	1e-4
Training epochs	4	-
Training batch size	32	48
Max length	512	512