BIOMRC: A Dataset for Biomedical Machine Reading Comprehension

Petros Stavropoulos[1,2], Dimitris Pappas[1,2], Ion Androutsopoulos[1], Ryan McDonald[3,1]

[1]Department of Informatics, Athens University of Economics and Business, Greece
[2]Institute for Language and Speech Processing, Research Center 'Athena', Greece
[3]Google Research
{pstav1993, pappasd, ion}@aueb.gr
ryanmcd@google.com

Abstract

We introduce BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments show that simple heuristics do not perform well on the new dataset, and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Non-expert human performance is also higher on the new dataset compared to BIOREAD, and biomedical experts perform even better. We also introduce a new BERT-based MRC model, the best version of which substantially outperforms all other methods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make the new dataset available in three different sizes, also releasing our code, and providing a leaderboard.

1 Introduction

Creating large corpora with human annotations is a demanding process in both time and resources. Research teams often turn to distantly supervised or unsupervised methods to extract training examples from textual data. In machine reading comprehension (MRC) (Hermann et al., 2015), a training instance can be automatically constructed by taking an unlabeled passage of multiple sentences, along with another smaller part of text, also unlabeled, usually the next sentence. Then a named entity of the smaller text is replaced by a placeholder. In this setting, MRC systems are trained (and evaluated for their ability) to read the passage and the smaller text, and guess the named entity that was replaced by the placeholder, which is typically one of the named entities of the passage. This kind of question answering (QA) is also known as cloze-type questions (Taylor, 1953). Several datasets have been created following this approach either using books (Hill et al., 2016; Bajgar et al., 2016) or news articles (Hermann et al., 2015). Datasets of this kind are noisier than MRC datasets containing human-authored questions and manually annotated passage spans that answer them (Rajpurkar et al., 2016, 2018; Nguyen et al., 2016). They require no human annotations, however, which is particularly important in biomedical question answering, where employing annotators with appropriate expertise is costly. For example, the BIOASQ QA dataset (Tsatsaronis et al., 2015) currently contains approximately 3k questions, much fewer than the 100k questions of a SQUAD (Rajpurkar et al., 2016), exactly because it relies on expert annotators.

To bypass the need for expert annotators and produce a biomedical MRC dataset large enough to train (or pre-train) deep learning models, Pappas et al. (2018) adopted the cloze-style questions approach. They used the full text of unlabeled biomedical articles from PUBMED CENTRAL,[1] and METAMAP (Aronson and Lang, 2010) to annotate the biomedical entities of the articles. They extracted sequences of 21 sentences from the articles. The first 20 sentences were used as a passage and the last sentence as a cloze-style question. A biomedical entity of the 'question' was replaced by a placeholder, and systems have to guess which biomedical entity of the passage can best fill the placeholder. This allowed Pappas et al. to produce a dataset, called BIOREAD, of approximately 16.4 million questions. As the same authors reported, however, the mean accuracy of three humans on a sample of 30 questions from BIOREAD was only 68%. Although this low score may be due to the fact that the three subjects were not biomedical experts, it is easy to see, by examining samples of BIOREAD, that many examples of the dataset do

[1] https://www.ncbi.nlm.nih.gov/pmc/

Proceedings of the BioNLP 2020 workshop, pages 140–149

'question' originating from caption: "figure 4 htert @entity6 and @entity4 XXXX cell invasion."
'question' originating from reference: "2004 , 17 , 250 257 .14967013 c samuni y. ; samuni u. ; goldstein s. the use of cyclic XXXX as hno scavengers ."
'passage' containing captions: "figure 2: distal UNK showing high insertion of rectum into common channel. figure 3: illustration of the cloacal malformation. figure 4: @entity5 showing UNK"

Table 1: Examples of noisy BIOREAD data. XXXX is the placeholder, and UNK is the 'unknown' token.

not make sense. Many instances contain passages or questions crossing article sections, or originating from the references sections of articles, or they include captions and footnotes (Table 1). Another source of noise is METAMAP, which often misses or mistakenly identifies biomedical entities (e.g., it often annotates 'to' as the country Togo).

In this paper, we introduce BIOMRC, a new dataset for biomedical MRC that can be viewed as an improved version of BIOREAD. To avoid crossing sections, extracting text from references, captions, tables etc., we use abstracts and titles of biomedical articles as passages and questions, respectively, which are clearly marked up in PUBMED data, instead of using the full text of the articles. Using titles and abstracts is a decision that favors precision over recall. Titles are likely to be related to their abstracts, which reduces the noise-to-signal ratio significantly and makes it less likely to generate irrelevant questions for a passage. We replace a biomedical entity in each title with a placeholder, and we require systems to guess the hidden entity by considering the entities of the abstract as candidate answers. Unlike BIOREAD, we use PUBTATOR (Wei et al., 2012), a repository that provides approximately 25 million abstracts and their corresponding titles from PUBMED, with multiple annotations.[2] We use DNORM's biomedical entity annotations, which are more accurate than METAMAP's (Leaman et al., 2013). We also perform several checks, discussed below, to discard passage-question instances that are too easy, and we show that the accuracy of experts and non-expert humans reaches 85% and 82%, respectively, on a sample of 30 instances for each annotator type, which is an indication that the new dataset is indeed less noisy, or at least that the task is more feasible for humans. Following Pappas et al. (2018), we release two versions of BIOMRC, LARGE and LITE, containing 812k and 100k instances respectively, for researchers with more or fewer resources, along with the 60 instances (TINY) humans answered. Random samples from BIOMRC LARGE where selected to create LITE and TINY. BIOMRC TINY is used only as a test set; it has no training and validation subsets.

We tested on BIOMRC LITE the two deep learning MRC models that Pappas et al. (2018) had tested on BIOREAD LITE, namely Attention Sum Reader (AS-READER) (Kadlec et al., 2016) and Attention Over Attention Reader (AOA-READER) (Cui et al., 2017). Experimental results show that AS-READER and AOA-READER perform better on BIOMRC, with the accuracy of AOA-READER reaching 70% compared to the corresponding 52% accuracy of Pappas et al. (2018), which is a further indication that the new dataset is less noisy or that at least its task is more feasible. We also developed a new BERT-based (Devlin et al., 2019) MRC model, the best version of which (SCIBERT-MAX-READER) performs even better, with its accuracy reaching 80%. We encourage further research on biomedical MRC by making our code and data publicly available, and by creating an on-line leaderboard for BIOMRC.[3]

2 Dataset Construction

Using PUBTATOR, we gathered approx. 25 million abstracts and their titles. We discarded articles with titles shorter than 15 characters or longer than 60 tokens, articles without abstracts, or with abstracts shorter than 100 characters, or fewer than 10 sentences. We also removed articles with abstracts containing fewer than 5 entity annotations, or fewer than 2 or more than 20 distinct biomedical entity identifiers. (PUBTATOR assigns the same identifier to all the synonyms of a biomedical entity; e.g., 'hemorrhagic stroke' and 'stroke' have the same identifier 'MESH:D020521'.) We also discarded articles containing entities not linked to any of the ontologies used by PUBTATOR,[4] or entities linked to multiple ontologies (entities with multiple ids), or entities whose spans overlapped with those of other entities. We also removed articles with no entities in their titles, and articles with no entities shared by the title and abstract.[5]

[2] Like PUBMED, PUBTATOR is supported by NCBI. Consult: www.ncbi.nlm.nih.gov/research/pubtator/

[3] Our code, data, and information about the leaderboard will be available at http://nlp.cs.aueb.gr/publications.html.

[4] PUBTATOR uses the Open Biological and Biomedical Ontology (OBO) Foundry, which comprises over 60 ontologies.

[5] A further reason for using the title as the question is that the entities of the titles are typically mentioned in the abstract.

Passage	BACKGROUND: Most brain metastases arise from **@entity0** . Few studies compare the brain regions they involve, their numbers and intrinsic attributes. METHODS: Records of all @entity1 referred to Radiation Oncology for treatment of symptomatic brain metastases were obtained. Computed tomography (n = 56) or magnetic resonance imaging (n = 72) brain scans were reviewed. RESULTS: Data from 68 breast and 62 @entity2 @entity1 were compared. Brain metastases presented earlier in the course of the lung than of the **@entity0** @entity1 (p = 0.001). There were more metastases in the cerebral hemispheres of the breast than of the @entity2 @entity1 (p = 0.014). More **@entity0** @entity1 had cerebellar metastases (p = 0.001). The number of cerebral hemisphere metastases and presence of cerebellar metastases were positively correlated (p = 0.001). The prevalence of at least one @entity3 surrounded with > 2 cm of @entity4 was greater for the lung than for the breast @entity1 (p = 0.019). The @entity5 type, rather than the scanning method, correlated with differences between these variables. CONCLUSIONS: Brain metastases from lung occur earlier, are more @entity4 , but fewer in number than those from **@entity0** . Cerebellar brain metastases are more frequent in **@entity0** .
Candidates	**@entity0** : ['breast and lung cancer'] ; @entity1 : ['patients'] ; @entity2 : ['lung cancer'] ; @entity3 : ['metastasis'] ; @entity4 : ['edematous', 'edema'] ; @entity5 : ['primary tumor']
Question	Attributes of brain metastases from XXXX .
Answer	**@entity0** : ['breast and lung cancer']

Figure 1: Example passage-question instance of BIOMRC. The passage is the abstract of an article, with biomedical entities replaced by @entity$N$ pseudo-identifiers. The original entity names are shown in square brackets. Both 'edematous' and 'edema' are replaced by '@entity4', because PUBTATOR considers them synonyms. The question is the title of the article, with a biomedical entity replaced by XXXX. @entity0 is the correct answer.

	BIOMRC LARGE				BIOMRC LITE				BIOMRC TINY		
	Training	**Development**	**Test**	**Total**	**Training**	**Development**	**Test**	**Total**	**Setting A**	**Setting B**	**Total**
Instances	700,000	50,000	62,707	812,707	87,500	6,250	6,250	100,000	30	30	60
Avg candidates	6.73	6.68	6.68	6.72	6.72	6.68	6.65	6.71	6.60	6.57	6.58
Max candidates	20	20	20	20	20	20	20	20	13	11	13
Min candidates	2	2	2	2	2	2	2	2	2	3	2
Avg abstract len.	253.79	257.41	253.70	254.01	253.78	257.32	255.56	254.11	248.13	264.37	256.25
Max abstract len.	543	516	511	543	519	500	510	519	371	386	386
Min abstract len.	57	89	77	57	60	109	103	60	147	154	147
Avg title len.	13.93	14.28	13.99	13.96	13.89	14.22	14.09	13.92	14.17	14.70	14.43
Max title len.	51	46	43	51	49	40	42	49	21	35	35
Min title len.	3	3	3	3	3	3	3	3	6	4	4

Table 2: Statistics of BIOMRC LARGE, LITE, TINY. The questions of the TINY version were answered by humans. All lengths are measured in tokens using a whitespace tokenizer.

Finally, to avoid making the dataset too easy for a system that would always select the entity with the most occurrences in the abstract, we removed a passage-question instance if the most frequent entity of its passage (abstract) was also the answer to the cloze-style question (title with placeholder); if multiple entities had the same top frequency in the passage, the instance was retained. We ended up with approx. 812k passage-question instances, which form BIOMRC LARGE, split into training, development, and test subsets (Table 2). The LITE and TINY versions of BIOMRC are subsets of LARGE.

In all versions of BIOMRC (LARGE, LITE, TINY), the entity identifiers of PUBTATOR are replaced by pseudo-identifiers of the form @entity$N$ (Fig. 1), as in the CNN and Daily Mail datasets (Hermann et al., 2015). We provide all BIOMRC versions in two forms, corresponding to what Pappas et al. (2018) call Settings A and B in BIOREAD.[6] In Setting A, each pseudo-identifier has a global scope, meaning that each biomedical entity has a unique pseudo-identifier in the whole dataset. This allows a system to learn information about the entity represented by a pseudo-identifier from all the occurrences of the pseudo-identifier in the training set. For example after seeing the same pseudo-identifier multiple times a model may learn that it stands for a drug, or that a particular pseudo-identifier tends to neighbor with specific words. Then, much like a language model, a system may guess the pseudo-identifier that should fill in the placeholder even without the passage, or at least it may infer a prior probability for each candidate answer. In contrast, Setting B uses a local scope, i.e., it restarts the numbering of the pseudo-identifiers (from @entity0) anew in each passage-question instance. This forces the models to rely only on information about the entities that can be inferred from the particular passage and question. This corresponds to a non-expert answering the question, who does not have any prior knowledge of the biomedical entities.

Table 2 provides statistics on BIOMRC. In TINY, we use 30 different passage-question instances in Settings A and B, because in both settings we asked the same humans to answer the questions, and we

[6] Pappas et al. (2018) actually call 'option a' and 'option b' our Setting B and A, respectively.

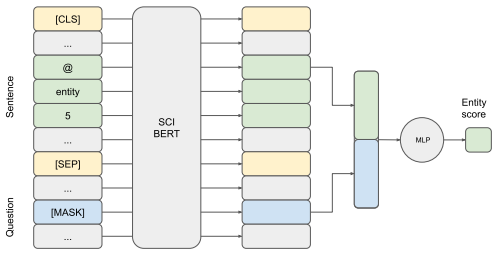

Figure 2: Illustration of our SCIBERT-based models. Each sentence of the passage is concatenated with the question and fed to SCIBERT. The top-level embedding produced by SCIBERT for the first sub-token of each candidate answer is concatenated with the top-level embedding of [MASK] (which replaces the placeholder XXXX) of the question, and they are fed to an MLP, which produces the score of the candidate answer. In SCIBERT-SUM-READER, the scores of multiple occurrences of the same candidate are summed, whereas SCIBERT-MAX-READER takes their maximum.

did not want them to remember instances from one setting to the other. In LARGE and LITE, the instances are the same across the two settings, apart from the numbering of the entity identifiers.

3 Experiments and Results

We experimented only on BIOMRC LITE and TINY, since we did not have the computational resources to train the neural models we considered on the LARGE version of BIOREAD. Pappas et al. (2018) also reported experimental results only on a LITE version of their BIOREAD dataset. We hope that others may be able to experiment on BIOMRC LARGE, and we make our code available, as already noted.

3.1 Methods

We experimented with the four basic baselines (BASE1–4) that Pappas et al. (2018) used in BIOREAD, the two neural MRC models used by the same authors, AS-READER (Kadlec et al., 2016) and AOA-READER (Cui et al., 2017), and a BERT-based (Devlin et al., 2019) model we developed.

Basic baselines: BASE1, 2, 3 return the first, last, and the entity that occurs most frequently in the passage (or randomly one of the entities with the same highest frequency, if multiple exist), respectively. Since in BIOREAD the correct answer is never (by construction) the most frequent entity of the passage, unless there are multiple entities with the same highest frequency, BASE3 performs poorly. Hence, we also include a variant, BASE3+, which randomly selects one of the entities of the passage with the same highest frequency, if multiple exist, otherwise it selects the entity with the second highest frequency. BASE4 extracts all the token n-grams from the passage that include an entity identifier (@entity$N$), and all the n-grams from the question that include the placeholder (XXXX).[7] Then for each candidate answer (entity identifier), it counts the tokens shared between the n-grams that include the candidate and the n-grams that include the placeholder. The candidate with the most shared tokens is selected. These baselines are used to check that the questions cannot be answered by simplistic heuristics (Chen et al., 2016).

Neural baselines: We use the same implementations of AS-READER (Kadlec et al., 2016) and AOA-READER (Cui et al., 2017) as Pappas et al. (2018), who also provide short descriptions of these neural models, not provided here to save space. The hyper-parameters of both methods were tuned on the development set of BIOMRC LITE.

BERT-based model: We use SCIBERT (Beltagy et al., 2019), a pre-trained BERT (Devlin et al., 2019) model for scientific text. SCIBERT is pre-trained on 1.14 million articles from Semantic Scholar,[8] of which 82% (935k) are biomedical and the rest come from computer science. For each passage-question instance, we split the passage into sentences using NLTK (Bird et al., 2009). For each sentence, we concatenate it (using BERT's [SEP] token) with the question, after replacing the XXXX with BERT's [MASK] token, and we feed the concatenation to SCIBERT (Fig. 2). We collect SCIBERT's top-level vector representations of the entity identifiers (@entity$N$) of the sentence and [MASK].[9] For each entity of the sentence, we concatenate its top-level representation with that of [MASK], and we feed them to a Multi-Layer Perceptron (MLP) to obtain a score for the particular entity (candidate answer). We thus obtain a score for all the entities of the passage. If an entity occurs multiple times in the passage, we take the sum or the maximum of the scores of its occurrences. In both cases, a softmax is then applied to the scores of all the entities, and the entity with the maximum score is selected as the answer. We call

[7] We tried $n = 2, \ldots, 6$ and use $n = 3$, which gave the best accuracy on the development set of BIOMRC LARGE.

[8] https://www.semanticscholar.org/

[9] BERT's tokenizer splits the entity identifiers into sub-tokens (Devlin et al., 2019). We use the first one. The top-level token representations of BERT are context-aware, and it is common to use the first or last sub-token of each named-entity.

Method	BIOMRC Lite – Setting A							BIOMRC Lite – Setting B						
	Train Acc	Dev Acc	Test Acc	Train Time	All Params	Word Embeds	Entity Embeds	Train Acc	Dev Acc	Test Acc	Train Time	All Params	Word Embeds	Entity Embeds
BASE1	37.58	36.38	37.63	0	0	0	0	37.58	36.38	37.63	0	0	0	0
BASE2	22.50	23.10	21.73	0	0	0	0	22.50	23.10	21.73	0	0	0	0
BASE3	10.03	10.02	10.53	0	0	0	0	10.03	10.02	10.53	0	0	0	0
BASE3+	44.05	43.28	44.29	0	0	0	0	44.05	43.28	44.29	0	0	0	0
BASE4	56.48	57.36	56.50	0	0	0	0	56.48	57.36	56.50	0	0	0	0
AS-READER	**84.63**	62.29	62.38	18 x 0.92 hr	12.87M	12.69M	1.59M	79.64	66.19	66.19	18 x 0.65 hr	6.82M	6.66M	0.60k
AOA-READER	82.51	70.00	69.87	29 x 2.10 hr	12.87M	12.69M	1.59M	**84.62**	71.63	71.57	36 x 1.82 hr	6.82M	6.66M	0.60k
SCIBERT-SUM-READER	71.74	71.73	71.28	11 x 4.38 hr	**154k**	**0**	**0**	68.92	68.64	68.24	6 x 4.38 hr	**154k**	**0**	**0**
SCIBERT-MAX-READER	81.38	**80.06**	**79.97**	19 x 4.38 hr	**154k**	**0**	**0**	81.43	**80.21**	**79.10**	15 x 4.38 hr	**154k**	**0**	**0**

Table 3: Training, development, test accuracy (%) on BIOMRC LITE in Settings A (global scope of entity identifiers) and B (local scope), training times (epochs $\times$ time per epoch), and number of trainable parameters (total, word embedding parameters, entity identifier embedding parameters). In the lower zone (neural methods), the difference from each accuracy score to the next best is statistically significant ($p < 0.02$). We used singe-tailed Approximate Randomization (Dror et al., 2018), randomly swapping the answers to 50% of the questions for 10k iterations.

this model SCIBERT-SUM-READER or SCIBERT-MAX-READER, depending on how it aggregates the scores of multiple occurrences of the same entity.

SCIBERT-SUM-READER is closer to AS-READER and AOA-READER, which also sum the scores of multiple occurrences of the same entity. This summing aggregation, however, favors entities with several occurrences in the passage, even if the scores of all the occurrences are low. Our experiments indicate that SCIBERT-MAX-READER performs better. In all cases, we only update the parameters of the MLP during training, keeping the parameters of SCIBERT frozen to their pre-trained values to speed up training. With more computing resources, it may be possible to improve the scores of SCIBERT-MAX-READER (and SCIBERT-SUM-READER) further by fine-tuning SCIBERT on BIOMRC training data.

3.2 Results on BIOMRC LITE

Table 3 reports the accuracy of all methods on BIOMRC LITE for Settings A and B. In both settings, all the neural models clearly outperform all the basic baselines, with BASE3 (most frequent entity of the passage) performing worst and BASE3+ performing much better, as expected. In both settings, SCIBERT-MAX-READER clearly outperforms all the other methods on both the development and test sets. The performance of SCIBERT-SUM-READER is approximately ten percentage points worse than SCIBERT-MAX-READER's on the development and test sets of both settings, indicating that the superior results of SCIBERT-MAX-READER are to a large extent due to the different aggregation function (max instead of sum) it uses to combine the scores of multiple occurrences of a candidate answer, not to the extensive pre-training of SCIBERT. AOA-READER, which does not employ any pre-training, is competitive to SCIBERT-SUM-READER in Setting A, and performs better than SCIBERT-SUM-READER in Setting B, which again casts doubts on the value of SCIBERT's extensive pre-training. We expect, however, that the performance of the SCIBERT-based models, could be improved further by fine-tuning SCIBERT's parameters.

The performance of SCIBERT-SUM-READER is slightly better in Setting A than in Setting B, which might suggest that the model manages to capture global properties of the entity pseudo-identifiers from the entire training set. However, the performance of SCIBERT-MAX-READER is almost the same across the two settings, which contradicts the previous hypothesis. Furthermore, the development and test performance of AS-READER and AOA-READER is higher in Setting B than A, indicating that these two models do not capture global properties of entities well, performing better when forced to consider only the information of the particular passage-question instance. Overall, we see no strong evidence that the models we considered are able to learn global properties of the entities.

In both Settings A and B, AOA-READER performs better than AS-READER, which was expected since it uses a more elaborate attention mechanism, at the expense of taking longer to train (Table 3).[10] The two SCIBERT-based models are also competitive in terms of training time, because we only train the MLP (154k parameters) on top of SCIBERT, keeping the parameters of SCIBERT frozen.

The trainable parameters of AS-READER and AOA-READER are almost double in Setting A compared to Setting B. To some extent, this difference is due to the fact that for both models we learn a word embedding for each @entity$N$ pseudo-identifier, and in Setting A the numbering of the identifiers is not reset for each passage-question

[10] We trained all models for a maximum of 40 epochs, using early stopping on the dev. set, with patience of 3 epochs.

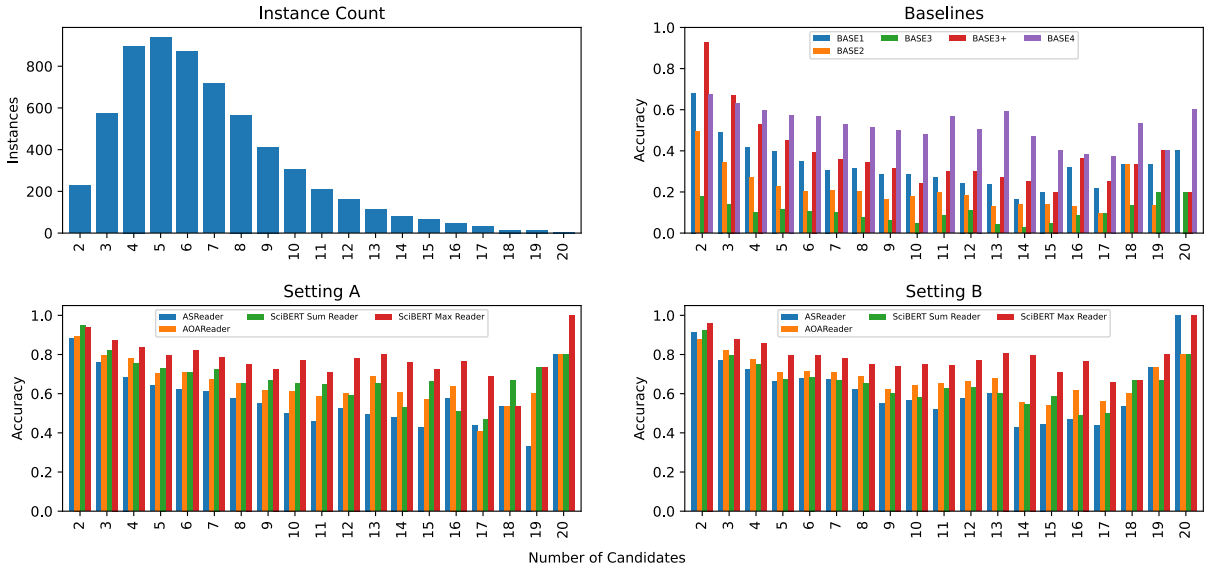

Figure 3: More detailed statistics and results on the development subset of BIOMRC LITE. Number of passage-question instances with 2, 3, ..., 20 candidate answers (top left). Accuracy (%) of the basic baselines (top right). Accuracy (%) of the neural models in Settings A (bottom left) and B (bottom right).

instance, leading to many more pseudo-identifiers (31.77k pseudo-identifiers in the vocabulary of Setting A vs. only 20 in Setting B); this accounts for a difference of 1.59M parameters.[11] The rest of the difference in total parameters (from Setting A to B) is due to the fact that we tuned the hyper-parameters of each model separately for each setting (A, B), on the corresponding development set. Hyper-parameter tuning was performed separately for each model in each setting, but led to the same numbers of trainable parameters for AS-READER and AOA-READER, because the trainable parameters are dominated by the parameters of the word embeddings. Note that the hyper-parameters of the two SCIBERT-based models (of their MLPs) were very minimally tuned, hence these models may perform even better with more extensive tuning.

AOA-READER was also better than AS-READER in the experiments of Pappas et al. (2018) on a LITE version of their BIOREAD dataset, but the development and test accuracy of AOA-READER in Setting A of BIOREAD was reported to be only 52.41% and 51.19%, respectively (cf. Table 3); in Setting B, it was 50.44% and 49.94%, respectively. The much higher scores of AOA-READER (and AS-READER) on BIOMRC LITE are an indication that the new dataset is less noisy, or that the task is at least more feasible for machines. The results of Pappas et al. (2018) were slightly higher in Setting A than in Setting B, suggesting that AOA-READER was able to benefit from the global scope of entity identifiers, unlike our findings in BIOMRC.[12]

Figure 3 shows how many passage-question instances of the development subset of BIOMRC LITE have 2, 3, ..., 20 candidate answers (top left), and the corresponding accuracy of the basic baselines (top right), and the neural models (bottom). BASE3+ is the best basic baseline for 2 and 3 candidates, and for 2 candidates it is competitive to the neural models. Overall, however, BASE4 is clearly the best basic baseline, but it is outperformed by all neural models in almost all cases, as in Table 3. SCIBERT-MAX-READER is again the best system in both settings, almost always outperforming the other systems. AS-READER is the worst neural model in almost all cases. AOA-READER is competitive to SCIBERT-SUM-READER in Setting A, and slightly better overall than SCIBERT-SUM-READER in Setting B, as can be seen in Table 3.

3.3 Results on BIOMRC TINY

Pappas et al. (2018) asked humans (non-experts) to answer 30 questions from BIOREAD in Setting A, and 30 other questions in Setting B. We mirrored their experiment by providing 30 questions (from

[11] Hyper-parameter tuning led to 50- and 30-dimensional word embeddings in Settings A, B, respectively. AS-READER and AOA-READER learn word embeddings from the training set, without using pre-trained embeddings.

[12] For AS-READER, Pappas et al. (2018) report results only for Setting B: 37.90% development and 42.01% test accuracy on BIOREAD LITE. They did not consider BERT-based models.

Passage	The study enrolled 53 @entity1 (29 males, 24 females) with **@entity1576** aged 15-88 years. Most of them were 59 years of age and younger. In 1/3 of the @entity1 the diseases started with symptoms of @entity1729, in 2/3 of them–with pulmonary affection. @entity55 was diagnosed in 50 @entity1 (94.3%), acute @entity3617 –in 3 @entity1. ECG changes were registered in about half of the examinees who had no cardiac complaints. 25 of them had alterations in the end part of the ventricular ECG complex; rhythm and conduction disturbances occurred rarely. Mycoplasmosis @entity1 suffering from @entity741 (@entity741) had stable ECG changes while in those free of @entity741 the changes were short. @entity296 foci were absent. @entity299 comparison in @entity1 with **@entity1576** and in other @entity1729 has found that cardiovascular system suffers less in acute mycoplasmosis. These data are useful in differential diagnosis of @entity296 .
Candidates	@entity1 : ['patients'] ; **@entity1576** : ['respiratory mycoplasmosis'] ; @entity1729 : ['acute respiratory infections', 'acute respiratory viral infection'] ; @entity55 : ['Pneumonia'] ; @entity3617 : ['bronchitis'] ; @entity741 : ['IHD', 'ischemic heart disease'] ; @entity296 : ['myocardial infections', 'Myocardial necrosis'] ; @entity299 : ['Cardiac damage'] .
Question	Cardio-vascular system condition in XXXX .
Expert Human Answers	annotator1: **@entity1576**; annotator2: **@entity1576**.
Non-expert Human Answers	annotator1: @entity296; annotator2: @entity296; annotator3: **@entity1576**.
Systems' Answers	AS-READER: @entity1729; AOA-READER: @entity296; SCIBERT-SUM-READER: **@entity1576**.

Figure 4: Example from BIOMRC TINY. In Setting A, humans see both the pseudo-identifiers (@entity$N$) and the original names of the biomedical entities (shown in square brackets). Systems see only the pseudo-identifiers, but the pseudo-identifiers have global scope over all instances, which allows the systems, at least in principle, to learn entity properties from the entire training set. In Setting B, humans no longer see the original names of the entities, and systems see only the pseudo-identifiers with local scope (numbering reset per passage-question instance).

BIOMRC LITE) to three non-experts (graduate CS students) in Setting A, and 30 other questions in Setting B. We also showed the same questions of each setting to two biomedical experts. As in the experiment of Pappas et al. (2018), in Setting A both the experts and non-experts were also provided with the original names of the biomedical entities (entity names before replacing them with @entity$N$ pseudo-identifiers) to allow them to use prior knowledge; see the top three zones of Fig. 4 for an example. By contrast, in Setting B the original names of the entities were hidden.

Table 4 reports the human and system accuracy scores on BIOMRC TINY. Both experts and non-experts perform better in Setting A, where they can use prior knowledge about the biomedical entities. The gap between experts and non-experts is three points larger in Setting B than in Setting A, presumably because experts can better deduce properties of the entities from the local context. Turning to the system scores, SCIBERT-MAX-READER is again the best system, but again much of its performance is due to the max-aggregation of the scores of multiple occurrences of entities. With sum-aggregation, SCIBERT-SUM-READER obtains exactly the same scores as AOA-READER, which again performs better than AS-READER. (AOA-READER and SCIBERT-SUM-READER make different mistakes, but their scores just happen to be identical because of the small size of TINY.) Unlike our results on BIOMRC LITE, we now see all systems performing better in Setting A compared to Setting B, which suggests they do benefit from the global scope of entity identifiers. Also, SCIBERT-MAX-READER performs better than both experts and non-experts in Setting A, and better than non-experts in Setting B. However, BIOMRC TINY contains only 30 instances in each setting, and hence the results of Table 4 are less reliable than those from BIOMRC LITE (Table 3).

In the corresponding experiments of Pappas et al. (2018), which were conducted in Setting B only, the average accuracy of the (non-expert) humans was 68.01%, but the humans were also allowed not to answer (when clueless), and unanswered questions were excluded from accuracy. On average, they did not answer 21.11% of the questions, hence their accuracy drops to 46.90% if unanswered questions are counted as errors. In our experiment, the humans were also allowed not to answer (when clueless), but we counted unanswered questions as errors, which we believe better reflects human performance. Non-experts answered all questions in Setting A, and did not answer 13.33% (4/30) of the questions on average in Setting B. The decrease in the questions non-experts did not answer (from 21.11% to 13.33%) in Setting B (the only one considered in BIOREAD) again suggests that the new dataset is less noisy, or at least that the task is more feasible for humans, even when the names of the entities are hidden. Experts did not answer 2.5% (0.75/30) and 1.67% (0.5/30) of the questions on average in Settings A and B, respectively.

Inter-annotator agreement was also higher for experts than non-experts in our experiment, in both

Method	Setting A	Setting B
Experts (Avg)	**85.00**	**61.67**
Non-Experts (Avg)	81.67	55.56
AS-READER	66.67	46.67
AOA-READER	70.00	56.67
SCIBERT-SUM-READER	70.00	56.67
SCIBERT-MAX-READER	**90.00**	**60.00**

Table 4: Accuracy (%) on BIOMRC TINY. Best human and system scores shown in bold.

Annotators (Setting)	Kappa
Experts (A)	70.23
Non Experts (A)	65.61
Experts (B)	72.30
Non Experts (B)	47.22

Table 5: Human agreement (Cohen's Kappa, %) on BIOMRC TINY. Avg. pairwise scores for non-experts.

Settings A and B (Table 5). In Setting B, the agreement of non-experts was particularly low (47.22%), possibly because without entity names they had to rely more on the text of the passage and question, which they had trouble understanding. By contrast, the agreement of experts was slightly higher in Setting B than Setting A, possibly because without prior knowledge about the entities, which may differ across experts, they had to rely to a larger extent on the particular text of the passage and question.

4 Related work

Several biomedical MRC datasets exist, but have orders of magnitude fewer questions than BIOMRC (Ben Abacha and Demner-Fushman, 2019) or are not suitable for a cloze-style MRC task (Pampari et al., 2018; Ben Abacha et al., 2019; Zhang et al., 2018). The closest dataset to ours is CLICR (Šuster and Daelemans, 2018), a biomedical MRC dataset with cloze-type questions created using full-text articles from BMJ case reports.[13] CLICR contains 100k passage-question instances, the same number as BIOMRC LITE, but much fewer than the 812.7k instances of BIOMRC LARGE. Šuster et al. used CLAMP (Soysal et al., 2017) to detect biomedical entities and link them to concepts of the UMLS Metathesaurus (Lindberg et al., 1993). Cloze-style questions were created from the 'learning points' (summaries of important information) of the reports, by replacing biomedical entities with placeholders. Šuster et al. experimented with the Stanford Reader (Chen et al., 2017) and the Gated-Attention Reader (Dhingra et al., 2017), which perform worse than AOA-READER (Cui et al., 2017).

The QA dataset of BIOASQ (Tsatsaronis et al., 2015) contains questions written by biomedical experts. The gold answers comprise multiple relevant documents per question, relevant snippets from the documents, exact answers in the form of entities, as well as reference summaries, written by the experts. Creating data of this kind, however, requires significant expertise and time. In the eight years of BIOASQ, only 3,243 questions and gold answers have been created. It would be particularly interesting to explore if larger automatically generated datasets like BIOMRC and CLICR could be used to pre-train models, which could then be fine-tuned for human-generated QA or MRC datasets.

Outside the biomedical domain, several cloze-style open-domain MRC datasets have been created automatically (Hill et al., 2016; Hermann et al., 2015; Dunn et al., 2017; Bajgar et al., 2016), but have been criticized of containing questions that can be answered by simple heuristics like our basic baselines (Chen et al., 2016). There are also several large open-domain MRC datasets annotated by humans (Kwiatkowski et al., 2019; Rajpurkar et al., 2016, 2018; Trischler et al., 2017; Nguyen et al., 2016; Lai et al., 2017). To our knowledge the biggest human annotated corpus is Google's Natural Questions dataset (Kwiatkowski et al., 2019), with approximately 300k human annotated examples. Datasets of this kind require extensive annotation effort, which for open-domain datasets is usually crowd-sourced. Crowd-sourcing, however, is much more difficult for biomedical datasets, because of the required expertise of the annotators.

5 Conclusions and Future Work

We introduced BIOMRC, a large-scale cloze-style biomedical MRC dataset. Care was taken to reduce noise, compared to the previous BIOREAD dataset of Pappas et al. (2018). Experiments showed that BIOMRC's questions cannot be answered well by simple heuristics, and that two neural MRC models that had been tested on BIOREAD perform much better on BIOMRC, indicating that the new dataset is indeed less noisy or at least that its task is more feasible. Human performance was also higher on a sample of BIOMRC compared to BIOREAD, and biomedical experts performed even better. We also developed a new BERT-based model, the best version of which outperformed all other meth-

[13] https://casereports.bmj.com/

ods tested, reaching or surpassing the accuracy of biomedical experts in some experiments. We make BIOMRC available in three different sizes, also releasing our code, and providing a leaderboard.

We plan to tune more extensively the BERT-based model to further improve its efficiency, and to investigate if some of its techniques (mostly its max-aggregation, but also using sub-tokens) can also benefit the other neural models we considered. We also plan to experiment with other MRC models that recently performed particularly well on open-domain MRC datasets (Zhang et al., 2020). Finally, we aim to explore if pre-training neural models on BIOREAD is beneficial in human-generated biomedical datasets (Tsatsaronis et al., 2015).

Acknowledgments

We are most grateful to I. Almirantis, S. Kotitsas, V. Kougia, A. Nentidis, S. Xenouleas, who participated in the human evaluation with BIOMRC TINY.

References

Alan R Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2016. Embracing data abundance: BookTest Dataset for Reading Comprehension. *CoRR*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: Pretrained Language Model for Scientific Text. In *EMNLP*.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20:511.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379, Florence, Italy.

Steven Bird, Loper Edward, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to Answer Open-Domain Questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-Attention Neural Networks for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 593–602, Vancouver, Canada.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. 2017. Gated-Attention Readers for Text Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1832–1846, Vancouver, Canada.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia.

Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Güney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine. *CoRR*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, page 1693–1701, Cambridge, MA, USA.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *CoRR*.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text Understanding with the Attention Sum Reader Network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918, Berlin, Germany.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai,

Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Donald A. B. Lindberg, Betsy L. Humphreys, and Alexa T. McCray. 1993. The Unified Medical Language System. *Yearbook of medical informatics*, 1:41–51.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *ArXiv*.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium.

Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou. 2018. BioRead: A New Dataset for Biomedical Reading Comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2017. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Simon Šuster and Walter Daelemans. 2018. CliCR: a Dataset of Clinical Case Reports for Machine Reading Comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, New Orleans, Louisiana.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool for Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. NewsQA: A Machine Comprehension Dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada.

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, and G. Paliouras. 2015. An Overview of the BioASQ Large-Scale Biomedical Semantic Indexing and Question Answering Competition. *BMC Bioinformatics*, 16(138).

Chih-Hsuan Wei, Bethany R. Harris, Donghui Li, Tanya Z. Berardini, Eva Huala, Hung-Yu Kao, and Zhiyong Lu. 2012. Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, 2012.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018. Medical Exam Question Answering with Large-scale Reading Comprehension. *ArXiv*.

Zhuosheng Zhang, Jun jie Yang, and Hai Zhao. 2020. Retrospective Reader for Machine Reading Comprehension. *ArXiv*.