Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5532–5538
Marseille, 11–16 May 2020

Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering

Sarvesh Soni[1], **Kirk Roberts**[1]
[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston
{sarvesh.soni, kirk.roberts}@uth.tmc.edu

Abstract

We evaluate the performance of various Transformer language models, when pre-trained and fine-tuned on different combinations of open-domain, biomedical, and clinical corpora on two clinical question answering (QA) datasets (CliCR and emrQA). We perform our evaluations on the task of machine reading comprehension, which involves training the model to answer a question given an unstructured context paragraph. We conduct a total of 48 experiments on different combinations of the large open-domain and domain-specific corpora. We found that an initial fine-tuning on an open-domain dataset, SQuAD, consistently improves the clinical QA performance across all the model variants.

Keywords: clinical question answering, machine comprehension, transfer learning

1. Introduction

A tremendous amount of unstructured data is available in a patient's electronic health record (EHR) notes. Finding information quickly in those notes is important to ensure their appropriate use by clinicians. A very intuitive means of querying EHRs for this information is to use natural language questions (Ely et al., 2005; Patrick and Li, 2012), and thus automatic question answering (QA) methods provide an intuitive interface for finding information in EHRs. Most prior work in EHR QA has focused on analyzing questions, which would enable answering questions from the structured data in EHRs (Roberts and Demner-Fushman, 2015; Roberts et al., 2016a; Roberts and Demner-Fushman, 2016a; Roberts and Patra, 2017), or mapping questions to existing natural language processing (NLP) based information extraction models (Patrick and Li, 2012). Little work has been done, however, in developing EHR QA models that learn to answer questions directly from pairs of questions and their answers in unstructured notes. Such a task is also known as machine reading comprehension (MRC), but will be referred to as QA in this paper to avoid confusion.

Recently, many deep learning methods have been developed to do precisely this task for non-medical domains, and at human levels of accuracy. Given a large number of questions, short documents, and corresponding answers in those short documents, these techniques are able to learn end-to-end QA models with accuracies up to the 90% range (Chen et al., 2017; Seo et al., 2017; Hu et al., 2018; Yu et al., 2018; Wang et al., 2017; Kundu and Ng, 2018; Devlin et al., 2019; Yang et al., 2019). This requires a significant amount of data: as we describe in the next section, many of these open domain datasets have at least 100k question-answer pairs. Table 1 shows an instance of such a dataset from SQuAD (Rajpurkar et al., 2016).

Developing such a large dataset for clinical question answering may well be infeasible. The medical training required to fully understand EHR notes as well as the privacy limitations imposed by legal/ethical restrictions on clinical data both make crowdsourcing such a dataset impossible. Instead, transfer learning approaches seem to hold the most promise for improving QA in this important domain. This would enable learning a model from a large dataset and refining it on a much smaller clinical QA dataset. The two main types of transfer learning for QA as it pertains to this paper are (a) *pre-training*: learning a (generally transformer-based) language model on a large corpus so as to derive high-quality word embeddings that effectively encode the meaning of text, and (b) *initial fine-tuning*: starting with the pre-trained language model, then learning a QA model prior to the ultimate fine-tuning on the dataset of interest. For clinical QA, however, it is an open question as to which datasets to select for both of these tasks, and whether such additional learning is worth the computational costs.

This paper is an attempt to answer these two (entangled) questions by systematically exploring both the pre-training source (open domain, biomedical domain, clinical domain) and initial fine-tuning source for two clinical QA datasets: CliCR (Suster and Daelemans, 2018) and emrQA (Pampari et al., 2018). In total, 48 experiments are performed to evaluate the efficacy of different options. Since this level of experimentation is not always feasible, our goal in this paper is to develop some best practice guidelines for how best to pre-train and fine-tune for clinical QA. Hopefully this process also offers insights into other specialized domains as well.

The remainder of this paper is organized as follows. Section 2 outlines prior work regarding adaptation of knowledge to domain-specific tasks, as well as some background on clinical QA. Section 3 describes the corpora used in the experiments (both source and target). Section 4 describes the transformer-based language models used in the experiments, including any domain-specific pre-training. Section 5 provides our experimental details and evaluation methodology. Section 6 details our results. Section 7 overviews our results, summarizing our attempt to learn generalizable best practices from these experiments. Finally, Section 8 provides a conclusion.

2. Background

Open-domain tasks are usually more explored than domain-specific ones, mainly because of the ease of access to datasets and support from a wider research community. Thus open-domain methods and datasets are frequently applied to closed domains that lack sufficiently large datasets. Wiese et al. (2017) train a neural architecture on an open-domain QA dataset (SQuAD) and use domain-specific word embeddings to achieve state-of-the-art performance on the biomedical-domain QA task. Etchegoyhen et al. (2018) evaluate the advantages of applying domain-adapted statistical and neural techniques to the task of machine translation (MT). They experiment on the datasets from three different domains frequently managed by translation service providers and highlight the benefits of using domain adaptation. We aim to use various transfer learning techniques such as pre-training and initial fine-tuning to adapt language representations from open-domain transformer models.

Medical QA has been heavily studied (Athenikos and Han, 2010). The different types of medical QA can be best understood based on the various types of resources from which answers can be drawn (Roberts et al., 2016a). Some QA methods draw answers from general medical knowledge (Cairns et al., 2011), which is generally appropriate for consumers to learn about a topic (Roberts et al., 2014; Luo et al., 2015; Kilicoglu et al., 2018), or health professionals to refresh or deepen their knowledge (Yu et al., 2005; Yu and Cao, 2008; Zhang et al., 2018). Notably, the general knowledge questions that consumers and professionals ask are quite different (Roberts and Demner-Fushman, 2016b), resulting in the need for different QA systems. Other QA methods draw answers from the biomedical literature (Tsatsaronis et al., 2012), which is generally appropriate for researchers looking for recent discoveries (Demner-Fushman and Lin, 2007), or for clinicians looking for the latest evidence-based information (Demner-Fushman and Lin, 2007; Roberts et al., 2016b). A third set of QA methods draw answers from patient-specific sources. That is, the answer is not a solution for the patient (e.g., what is the best treatment for this patient?), but rather directly from the patient's records (e.g., what treatments have been given to this patient?). These answers are generally drawn from the EHR. However, little focus has been drawn toward adapting the fast growing developments in open-domain QA. To this effect, we evaluate the performance of several transformer language models (pre-trained on open-domain as well as biomedical and clinical corpora).

3. Corpora

Modeling deep learning methods usually requires a large amount of training data. Thus, we use 3 different large machine comprehension datasets for our analysis. In this section, we summarize the details of each of these datasets. We further describe any preprocessing steps applied on the datasets before the evaluations.

Passage: Dog intelligence is the ability of the dog to perceive information and retain it as knowledge for applying to solve problems. Dogs have been shown to learn by inference. A study with ***Rico*** showed that he knew the labels of over 200 different items. He inferred the names of novel items by exclusion learning and correctly retrieved those novel items immediately and also 4 weeks after the initial exposure. ...

Question: What is the name of the dog that could ID over 200 things?
Answer: ***Rico***

Table 1: An example of question-answer pair from the SQuAD dataset with an excerpt of the relevant paragraph. The answer is ***italicized*** in the passage.

Measure	Corpus		
	SQuAD	**CliCR**	**emrQA**
# of paragraphs	20,963	11,730	303
# of questions	98,169	74,743	73,111
Avg # of questions per para.	4.68	6.37	241.29
Avg paragraph length	134.78	1,389.76	1381.10
Avg question length	11.30	22.54	9.40
Avg answer length	3.35	1.98	1.88

Table 2: Descriptive statistics of the datasets used in evaluation. Length is in terms of tokens. **#** – Count. **Avg** – Average. **para** – paragraph.

3.1. SQuAD

SQuAD (Rajpurkar et al., 2016) is an open-domain question answering dataset constructed using Wikipedia articles through crowdsourcing. It consists of over 100,000 questions collected from crowdworkers against various excerpts from the Wikipedia articles. An instance from the dataset is presented in Table 1. Its large size and public availability prompted the construction of many machine reading comprehension models [1]. This dataset can be considered a good representative example of an open-domain QA task.

We do not apply any preprocessing techniques to the SQuAD dataset as it was already in the required format and comes with a pre-defined training set. The descriptive statistics for this dataset are presented in Table 2.

3.2. CliCR

CliCR (Suster and Daelemans, 2018) is a cloze style medical QA dataset built from clinical case reports. It contains over 100,000 gap filling queries constructed using the learning points (which summarizes the report contents) associated with each case report. Precisely, they replace medical entities in the learning points with blanks to construct the queries. A clinical case report details information regarding the signs and symptoms of a medical condition

[1] https://rajpurkar.github.io/SQuAD-explorer

Passage: Summary *Obturator hernia* (OH) is an uncommon cause of bowel obstruction and described in elderly females in the literature. The treatment has traditionally been laparotomy because of an acute nature of the condition. However, because of old age and comorbidities that OH is associated with, general anaesthesia may need to be avoided. In the current case, a transinguinal preperitoneal approach and ...

Original query: _____ is a rare but relevant differential diagnosis of a bowel obstruction?

Question: What is a rare but relevant differential diagnosis of a bowel obstruction?
Answer: *Obturator hernia*

Table 3: An example of question-answer pair from the CliCR dataset along with the original query and an excerpt of the relevant case report. The answer is *italicized* in the note.

Passage: RECORD #XXXXX ... the patient had atrial fibrillation which was treated and controlled pharmacologically, and also the patient was treated with *prophylactic anticoagulation* with Coumadin. The patient went to the Operating Room on 11/23 , had a coronary artery bypass graft x three with a saphenous vein graft to the LAD , first branch of the obtuse marginal and the posterior descending artery. ...

Question: Why was the patient prescribed coumadin?
Answer: *prophylactic anticoagulation*

Table 4: An example of question-answer pair from the emrQA dataset with an excerpt of the relevant de-identified clinical note. The answer is *italicized* in the note.

(usually rare or unusual) along with a discussion on its diagnosis and treatment for a patient. The contents of these reports are similar to the discharge summaries present in electronic health records, and hence can be considered a proxy for EHR data. However, it should be noted that case reports are much cleaner than EHR text, particularly lacking many of the abbreviations, telegraphic grammar, and document structure seen in EHR notes.

Preprocessing We perform several preprocessing steps on the dataset for mapping it to a more challenging and SQuAD-like format. First, we strip off all the medical entity annotations from the queries and the reports making the task more difficult, as no candidate entities are marked in the context paragraph. Note that the task explained in the original paper expects a candidate set of answers while the models we used for evaluations directly extract answers from a given context paragraph. Second, we filter the list of queries to the ones with answers present in the associated report. In cases where more than one unique answers are found in the passage (e.g., synonyms – as the authors of the original paper automatically add related terms to the answer set using a medical tool) we use only one of the answers available for a query. Last, for mapping the gap filling queries to fully formed questions (like SQuAD) we map the blanks in the queries to *"what"*, making the queries interrogative. An instance of such a transformation is given in Table 3.

Split There is a pre-defined training-testing-development split for the CliCR dataset. Thus, we simply combine the questions from their training and development sets to form the training set for our evaluations and use their testing split as it is. Table 2 shows the descriptive statistics of the refined dataset. Note that the average question length of the CliCR dataset (22.54) is longer than that of the SQuAD dataset (11.30).

3.3. emrQA

emrQA (Pampari et al., 2018) is a large medical QA dataset automatically constructed from the i2b2 challenge datasets. Specifically, they utilize the existing NLP annotations to populate pre-defined question and paraphrase templates and associate them with the corresponding clinical notes. So while this is a large corpus, it is not necessarily composed of a realistic distribution of questions. We make use of the question-answer pairs (a total of 400,000 in the original corpus) from the dataset where each question can be answered using an associated clinical note (e.g., see Table 4). Clinical notes such as progress notes and discharge summaries are part of EHR data and contain vital patient information in the form of natural language. Thus, this dataset is a good representative of a clinical QA task.

Preprocessing We mainly preprocess the dataset for mapping it to a SQuAD-like format. The questions in this dataset broadly fall into 3 categories on the basis of their answer type – empty, single, or complex. We follow the authors of this dataset and exclude the questions with empty answers from our evaluations as such pairs are more representative of a class prediction task rather than QA. We include all the questions with a single answer provided the answer can be found in the associated note. Last, from the questions with complex answers (i.e., requiring multiple entities from the passage to accurately answer a question), we only include the ones having a unique answer which is present in the associated paragraph.

Split As the emrQA dataset does not have any pre-defined training or testing splits, we split the dataset into training and testing sets after the preprocessing step. We divide the dataset at context paragraph level, i.e., all the questions belonging to a context remain in only one of the sets. This is done to avoid any model bias that could cause due to the presence of a context in both train and test sets. We split the

dataset into training and testing sets in the ratio of 90:10. More details about the dataset can be found in Table 2. We note that the average paragraph length in CliCR (1,389.76) and emrQA (1,381.10) datasets are much longer compared to that in the general-domain SQuAD dataset (134.78).

4. Models

We select a variety of models to analyze the effect of using different pre-training datasets on clinical QA. Specifically, we use 4 variants of Transformer language models pre-trained on different open-domain, biomedical, and clinical datasets. A Transformer relies on self-attention mechanisms to learn the input and output representations instead of using recurrent layers (allowing them to be trained in a more parallel fashion) (Vaswani et al., 2017). We run our evaluations using the BERT (Devlin et al., 2019), BioBERT (Lee et al., 2019), Clinical BERT (Alsentzer et al., 2019), and XLNet (Yang et al., 2019) models.

BERT (Bidirectional Encoder Representations from Transformers) uses masked language models to pre-train a deep bidirectional Transformer that allows for just fine-tuning the pre-trained model parameters on the downstream tasks (the QA task in our case). It is pre-trained on large datasets such as BooksCorpus (800M words) and English Wikipedia (2,500M words). We use the cased $\text{BERT}_{\text{BASE}}$ variant of the model (12 layers, 110M parameters) which is pre-trained for 1M steps. We choose this model variant for a fair comparison with the other models included in this study (as 2 of the other models are built upon cased $\text{BERT}_{\text{BASE}}$).

BioBERT (BERT for Biomedical Text Mining) is initiated with the same $\text{BERT}_{\text{BASE}}$ model as described above and further trained on biomedical data. We use the BioBERT model trained on PubMed abstracts (4500M words) for 1M steps as it is the recommended model variant from the authors.

Clinical BERT is initialized from a BioBERT variant pre-trained on PubMed abstracts (for 200K) and PMC articles (for 270K), and further trained on MIMIC III notes. We use the model variant trained on discharge summaries from the MIMIC database for 100,000 steps (again, it is the recommended variant of the model from the authors).

XLNet is an autoregressive transformer model which uses permutation language modeling and overcomes the pretrain-finetune discrepancy suffered by the BERT architecture (hence better modeling the bidirectional contexts). The pre-training corpus includes all the datasets from BERT along with 3 additional datasets – Giga5, ClueWeb 2012-B, and Common Crawl. We choose the cased $\text{XLNet}_{\text{BASE}}$ variant for our evaluations as it is similar to the $\text{BERT}_{\text{BASE}}$ model in terms of the architectural parameters (but trained on a larger corpus).

The aforementioned models are good representatives for open-domain as well as the biomedical and clinical domain tasks. The different corpora used for pre-training these models are summarized in Figure 1.

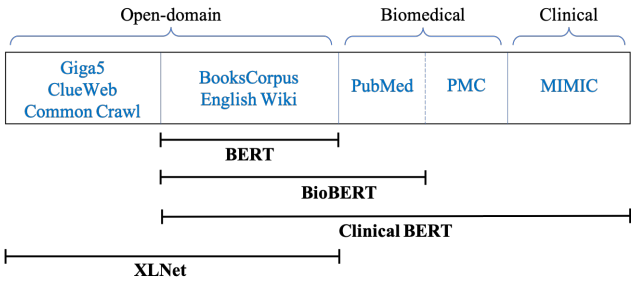

Figure 1: Datasets for pre-training the models used in our evaluation.

5. Evaluation

We perform an array of experiments to analyze the effect of different fine-tuning datasets on clinical QA performance. Specifically, we run the above models by fine-tuning on different combinations of the included datasets and evaluate their performance on the held out test sets from CliCR and emrQA datasets.

Fine-tuning on one dataset In this fine-tuning variant, we tune all models on exactly one dataset for 2 epochs. We choose the number of epochs following the hyperparameter recommendations for these models. An epoch corresponds to passing the entire train set through a model once. All these tuned models are then tested on both the CliCR and the emrQA test splits. E.g., we fine-tune on SQuAD or CliCR or emrQA for 2 epochs and predict on CliCR or emrQA test set.

Fine-tuning on two datasets The models' performance on a dataset can vary when one fine-tunes on a different dataset before fine-tuning it on the target dataset itself. Hence, in this variant, before fine-tuning the models on each of the 2 domain-specific datasets we first fine-tune the model on a different dataset for 1 epoch. After this initial fine-tuning for 1 epoch, we fine-tune the models normally on each of the domain-specific datasets for 2 epochs. Hence, a model is fine-tuned for a total of 3 epochs. E.g., we first fine-tune a model on SQuAD or emrQA for 1 epoch, then further fine-tune it on CliCR for 2 epochs (hence fine-tune for a total of 3 epochs), and finally predict on the CliCR dataset.

Fine-tuning on three datasets We take it a step further and initially fine-tune on each of the other 2 datasets for 1 epoch before fine-tuning on a medical dataset. E.g., we fine-tune on SQuAD for 1 epoch, then fine-tune on emrQA for 1 epoch, and finally fine-tune on the CliCR dataset for 2 epochs (a total of 4 epochs). Lastly, we predict on the CliCR test set to compute the performance.

We use a standard set of metrics for evaluating the performance of the models on the QA task – exact match (EM) and F1 score. EM is a stricter metric among the two and matches whole answer phrases (i.e., the predicted and the ground truth answer should match exactly). F1 is a word-level match and calculates the similarity between the predicted and the ground truth answers. E.g., if the exact answer is *"severe hypertension"*, but the system predicts *"hypertension"*, EM gives no credit, while F1 would be 0.67.

Fine-tuned on	BERT		BioBERT		Clinical BERT		XLNet	
	EM	F1	EM	F1	EM	F1	EM	F1
emrQA	8.78	13.62	5.79	9.14	4.95	7.69	5.71	8.87
SQuAD	28.06	37.59	34.81	45.12	31.06	40.83	33.23	41.88
CliCR	31.76	40.19	43.17	54.16	37.37	47.5	34.61	44.49
emrQA ⇒ CliCR	35.29	44.6	42.44	53.48	38.74	48.91	36.49	46.33
SQuAD ⇒ CliCR	**35.95**	**45.7**	**43.41**	**54.21**	**40.88**	**51.81**	38.18	47.98
SQuAD ⇒ emrQA ⇒ CliCR	34.89	44.4	42.69	53.44	40.14	50.49	**38.52**	**48.67**

Table 5: Model performances on the CliCR test set when fine-tuned on different combinations of the datasets. **EM** – Exact Match. **F1** – F1 score.

Fine-tuned on	BERT		BioBERT		Clinical BERT		XLNet	
	EM	F1	EM	F1	EM	F1	EM	F1
CliCR	4.64	6.36	2.55	3.77	3.12	4.85	3.81	5.65
SQuAD	7.03	18.45	13.45	26.92	22.12	34.91	18.88	33.22
emrQA	60.03	64.65	67.41	72.22	65.33	71.09	69.18	74.58
CliCR ⇒ emrQA	61.63	65.73	64.93	69.9	67.29	73.38	64.94	71.44
SQuAD ⇒ emrQA	63.93	69.38	69.16	73.58	68.64	74.91	**69.41**	**75.1**
SQuAD ⇒ CliCR ⇒ emrQA	**65.32**	**70.53**	**69.25**	**75.56**	**70.56**	**75.39**	67.1	72.57

Table 6: Model performances on the emrQA test set when fine-tuned on different combinations of the datasets. **EM** – Exact Match. **F1** – F1 score.

5.1. Experimental Setup

We use the recommended hyperparameters for fine-tuning all the model variants. Namely, the maximum sequence length is 384; document stride is 128; maximum query length is 128. For the BERT-based models, we use a learning rate of 3 x 10^{-5}. Furthermore, for the XLNet model, the learning rate is 2 x 10^{-5} and the Adam epsilon is 1 x 10^{-5}. All the experiments are performed on an NVidia Tesla V100 GPU (32G).

6. Results

The prediction results of the various model variants, pre-trained and fine-tuned on different dataset combinations, on our target clinical datasets CliCR and emrQA are presented in Table 5 and Table 6, respectively. We observe that all the BERT-based model variants (BERT, BioBERT, and Clinical BERT) perform better in terms of both EM and F1 metrics when more than one dataset is involved in the fine-tuning process. For evaluation on the CliCR dataset, the sequential fine-tuning combination of SQuAD and CliCR datasets achieves better prediction results for all the models as compared to fine-tuning combination of emrQA and CliCR (shown in fourth and fifth rows in Table 5). A similar observation is noted in the prediction results on the emrQA dataset in Table 6 where the fine-tuning combination of SQuAD and emrQA improves performance as compared to combining CliCR and emrQA.

We note that in the case of fine-tuning using a single dataset, the best performance is obtained when the datasets are the same for both fine-tuning and prediction. For instance, while evaluating the results on emrQA, the highest EM and F1 scores are achieved when the models are fine-tuned on the emrQA itself versus when fine-tuned individually on CliCR or SQuAD (compare the third row with the first two rows in Table 6).

Another noteworthy finding is the difference in the performance trend of XLNet on different fine-tuning variations. When XLNet is used to test on the CliCR dataset, incorporating intermediate fine-tuning on emrQA helps in achieving the highest performance among all the fine-tuning combinations (see the last row in Table 5). Whereas when the XLNet is tested on the emrQA dataset, incorporating intermediate fine-tuning on CliCR does not result in performance improvement over the combination of SQuAD and emrQA (compare the fifth and sixth rows in Table 6).

Among all the fine-tuning and model variations, the highest EM (70.56) achieved for the emrQA dataset is for 'SQuAD ⇒ CliCR ⇒ emrQA' fine-tuning variation utilizing Clinical BERT. For prediction on CliCR, the highest EM (43.41) is achieved through the sequential fine-tuning of 'SQuAD ⇒ CliCR' utilizing the BioBERT model variant.

7. Discussion

This work investigates the impact of pre-training and fine-tuning various Transformer-based language models on different dataset combinations when applied to question answering task for two clinical datasets. The aim of this paper is not to achieve state-of-the-art results on the included clinical datasets (emrQA and CliCR). However, our best performing models surpass the performances reported

in both the CliCR (Suster and Daelemans, 2018) and the emrQA (Pampari et al., 2018) papers. Furthermore, we use a stricter version of the QA task for CliCR (no entities are required a priori by the models in contrast to the original paper) and apply a stricter evaluation metric for emrQA (use standard definitions of EM and F1 in contrast to their lenient versions employed in Pampari et al. (2018)).

It can be noted that fine-tuning on another medical dataset (different than the one for which the task is predicted) consistently performs worse than fine-tuning on the SQuAD dataset. In other words, fine-tuning on an open-domain dataset performs better than fine-tuning on another dataset in the same domain. This shows a gap in the availability of a well-generalizable clinical QA dataset.

The results also reveal that fine-tuning on SQuAD for an epoch before fine-tuning on any of the medical datasets improves their performance almost all the time. This can be seen as a characteristic of the good quality of this manually constructed large dataset.

Though we did not run each of the above models multiple times (because training these models is a resource-intensive job), we note that there is a clear trend in almost all the variants across all the models. The pattern in the performance improvement by applying the different fine-tuning variations is similar for most of the models which indicates that the performances are representative of the real world scenarios and are not merely by chance.

8. Conclusion

We evaluated the performance of different Transformer language models when pre-trained and fine-tuned on different combinations of the open-domain and domain-specific datasets. We experimented with different fine-tuning combinations using single as well as multiple datasets. We performed a total of 48 experiments running various combinations of the models. We found that the initial fine-tuning helps in improving the performance in majority of the cases. Also, we achieved better results than the currently available best results for the included clinical datasets (CliCR and emrQA).

9. Acknowledgments

This work was supported by the U.S. National Library of Medicine, National Institutes of Health (NIH), under award R00LM012104; the Cancer Prevention and Research Institute of Texas (CPRIT), under award RP170668; as well as the Bridges Family Doctoral Fellowship Award.

10. Bibliographical References

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., and Savova, G. K. (2011). The MiPACQ Clinical Question Answering System. *AMIA Annual Symposium Proceedings*, 2011:171–180.

Chen, D., Fisch, A., Weston, J., and Bordes, A. (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. Association for Computational Linguistics.

Demner-Fushman, D. and Lin, J. (2007). Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ely, J. W., Osheroff, J. A., Chambliss, M. L., Ebell, M. H., and Rosenbaum, M. E. (2005). Answering Physicians' Clinical Questions: Obstacles and Potential Solutions. *Journal of the American Medical Informatics Association*, 12(2):217–224.

Etchegoyhen, T., Fernández Torné, A., Azpeitia, A., Martínez Garcia, E., and Matamala, A. (2018). Evaluating Domain Adaptation for Machine Translation Across Scenarios. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Hu, M., Peng, Y., Huang, Z., Qiu, X., Wei, F., and Zhou, M. (2018). Reinforced Mnemonic Reader for Machine Reading Comprehension. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 4099–4106. AAAI Press.

Kilicoglu, H., Ben Abacha, A., Mrabet, Y., Shooshan, S. E., Rodriguez, L., Masterton, K., and Demner-Fushman, D. (2018). Semantic annotation of consumer health questions. *BMC Bioinformatics*, 19(1):34.

Kundu, S. and Ng, H. T. (2018). A Nil-Aware Answer Extraction Framework for Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4243–4252. Association for Computational Linguistics.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, pages 1–7.

Luo, J., Zhang, G.-Q., Wentz, S., Cui, L., and Xu, R. (2015). SimQ: Real-Time Retrieval of Similar Consumer Health Questions. *Journal of Medical Internet Research*, 17(2):e43.

Pampari, A., Raghavan, P., Liang, J., and Peng, J. (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. Association for Computational Linguistics.

Patrick, J. and Li, M. (2012). An ontology for clinical questions about the contents of patient notes. *Journal of Biomedical Informatics*, 45(2):292–306.

Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (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. Association for Computational Linguistics.

Roberts, K. and Demner-Fushman, D. (2015). Toward a Natural Language Interface for EHR Questions. In *AMIA Joint Summits on Translational Science Proceedings. AMIA Summit on Translational Science*, volume 2015, pages 157–161. American Medical Informatics Association.

Roberts, K. and Demner-Fushman, D. (2016a). Annotating logical forms for EHR questions. In *Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016*, pages 3772–3778. NIH Public Access.

Roberts, K. and Demner-Fushman, D. (2016b). Interactive use of online health resources: A comparison of consumer and professional questions. *Journal of the American Medical Informatics Association*, 23(4):802–811.

Roberts, K. and Patra, B. G. (2017). A Semantic Parsing Method for Mapping Clinical Questions to Logical Forms. In *AMIA Annual Symposium Proceedings*, volume 2017. American Medical Informatics Association.

Roberts, K., Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2014). Automatically Classifying Question Types for Consumer Health Questions. *AMIA Annual Symposium Proceedings*, 2014:1018–1027.

Roberts, K., Rodriguez, L., Shooshan, S. E., and Demner-Fushman, D. (2016a). Resource Classification for Medical Questions. In *AMIA Annual Symposium Proceedings*, volume 2016, pages 1040–1049. American Medical Informatics Association.

Roberts, K., Simpson, M., Demner-Fushman, D., Voorhees, E., and Hersh, W. (2016b). State-of-the-art in biomedical literature retrieval for clinical cases: A survey of the TREC 2014 CDS track. *Information Retrieval Journal*, 19(1-2):113–148.

Seo, M., Kembhavi, A., Farhadi, A., and Hajishirzi, H. (2017). Bidirectional Attention Flow for Machine Comprehension. In *Proceedings of the 5th International Conference on Learning Representations*.

Suster, S. and Daelemans, W. (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*, pages 1551–1563. Association for Computational Linguistics.

Tsatsaronis, G., Schroeder, M., Paliouras, G., Almirantis, Y., Androutsopoulos, I., Gaussier, E., Gallinari, P., Artieres, T., Alvers, M. R., Zschunke, M., et al. (2012). Bioasq: A challenge on large-scale biomedical semantic indexing and question answering. In *2012 AAAI Fall Symposium Series*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Wang, W., Yang, N., Wei, F., Chang, B., and Zhou, M. (2017). Gated Self-Matching Networks for Reading Comprehension and Question Answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198. Association for Computational Linguistics.

Wiese, G., Weissenborn, D., and Neves, M. (2017). Neural Domain Adaptation for Biomedical Question Answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 281–289. Association for Computational Linguistics.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*.

Yu, H. and Cao, Y.-g. (2008). Automatically Extracting Information Needs from Ad Hoc Clinical Questions. *AMIA Annual Symposium Proceedings*, 2008:96–100.

Yu, H., Sable, C., and Zhu, H. R. (2005). Classifying medical questions based on an evidence taxonomy. In *Proceedings of the AAAI 2005 Workshop on Question Answering in Restricted Domains*.

Yu, A. W., Dohan, D., Luong, M.-T., Zhao, R., Chen, K., Norouzi, M., and Le, Q. V. (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. In *ICLR*.

Zhang, X., Wu, J., He, Z., Liu, X., and Su, Y. (2018). Medical exam question answering with large-scale reading comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.