DOMAIN-RELEVANT EMBEDDINGS FOR QUESTION SIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

The rate at which medical questions are asked online significantly exceeds the capacity of qualified people to answer them, leaving many questions unanswered or inadequately answered. Many of these questions are not unique, and reliable identification of similar questions would enable more efficient and effective question answering schema. While many research efforts have focused on the problem of general question similarity, these approaches do not generalize well to the medical domain, where medical expertise is often required to determine semantic similarity. In this paper, we show how a semi-supervised approach of pre-training a neural network on *medical question-answer pairs* is a particularly useful intermediate task for the ultimate goal of determining medical question similarity. While other pre-training tasks yield an accuracy below 78.7% on this task, our model achieves an accuracy of 82.6% with the same number of training examples, an accuracy of 80.0% with a much smaller training set, and an accuracy of 84.5% when the full corpus of medical question-answer data is used.

1 INTRODUCTION

With the ubiquity of the Internet and the emergence of medical question-answering websites such as ADAM (`www.adam.com`), WebMD (`www.webmd.com`), and HealthTap (`www.healthtap.com`), people are increasingly searching online for answers to their medical questions. However, the number of people asking medical questions online far exceeds the number of qualified experts – i.e doctors – answering them. One way to address this imbalance is to build a system that can automatically match *unanswered* questions with semantically similar *answered* questions, or mark them as priority if no similar answered questions exist. This approach uses doctor time more efficiently, reducing the number of unanswered questions and lowering the cost of providing online care.

Many of the individuals seeking medical advice online are otherwise reluctant to seek medical help due to cost, convenience, or embarrassment. For these patients, an accurate online system is critical because it may be the only medical advice they receive. Of course, some medical problems require in-person care, and an online system must indicate that. Other patients use the internet in addition to in-person care either to determine when an appointment is needed or to follow up after visits when they have lingering questions. For this second group, if the answers they see online do not match those given to them by their doctors, they are less likely to follow the advice of their doctors (Nosta, 2017), which can have serious consequences.

Coming up with an accurate algorithm for finding similar medical questions, however, is difficult. Simple heuristics such as word-overlap are ineffective because *Can a menstrual blood clot travel to your heart or lungs like other blood clots can?* and *Can clots from my period cause a stroke or embolism?* are similar questions with low overlap, but *Is candida retested after treatment* and *Is Chlamydia retested after treatment?* are critically different and only one word apart. Machine learning is a good candidate for such complex tasks, but requires labeled training data. As no widely available data for this particular task exists, we generate and release our own dataset of medical question pairs such as the ones shown in Table 1.

Given the recent success of pre-trained bi-directional transformer networks for natural language processing (NLP) outside the medical field (Peters et al., 2018; Devlin et al., 2018; Radford et al.; Yang et al., 2019; Liu et al., 2019), most research efforts in medical NLP have tried to apply general

Table 1: Examples from our medical question-pairs dataset

Question 1	Question 2	Label
After how many hour from drinking an antibiotic can I drink alcohol?	I have a party tonight and I took my last dose of Azithromycin this morning. Can I have a few drinks?	Similar
What specific exercises would help bursitis of the suprapatellar?	Can I take any medication for pain due to suprapatellar bursitis? Unable to exercise. :(	Different

language models to medical tasks (Ben Abacha et al., 2019). However, these models are not trained on medical information, and make errors that reflect this. In this work, we augment the features in these general language models using the depth of information that is stored within a *medical question-answer pair* to embed medical knowledge into the model. Our models pre-trained on this task outperform models pre-trained on out-of-domain question similarity with high statistical significance, and the results show promise of generalizing to other domains as well.

The task of question-answer matching was specifically chosen because it is closely related to that of question similarity; one component of whether or not two questions are semantically similar is whether or not the answer to one also answers the other. We show that the performance gains achieved by this particular task are not realized by other in-domain tasks, such as medical question-categorization and medical answer completion.

The main contributions of this paper are:

- We release a dataset of medical question pairs generated and labeled by doctors that is based upon real, patient-asked questions

- We prove that, particularly for medical NLP, domain matters: pre-training on a different task in the same domain outperforms pre-training on the same task in a different domain

- We show that the task of question-answer matching embeds relevant medical information for question similarity that is not captured by other in-domain tasks

2 RELATED WORK

2.1 PRE-TRAINED NETWORKS FOR GENERAL LANGUAGE UNDERSTANDING

NLP has undergone a transfer learning revolution in the past year, with several large pre-trained models earning state-of-the-art scores across many linguistic tasks. Two such models that we use in our own experiments are BERT (Devlin et al., 2018) and XLNet (Yang et al., 2019). These models have been trained on semi-supervised tasks such as predicting a word that has been masked out from a random position in a sentence, and predicting whether or not one sentence is likely to follow another. The corpus used to train BERT was exceptionally large (3.3 billion words), but all of the data came from BooksCorpus and Wikipedia. Talmor & Berant (2019) recently found that BERT generalizes better to other datasets drawn from Wikipedia than to tasks using other web snippets. This is consistent with our finding that pre-training domain makes a big difference.

2.2 DOUBLE FINETUNING FOR DOMAIN TRANSFER

To address the need for pre-trained models in particular domains, some researchers have recently re-trained BERT on different text corpora such as scientific papers (Beltagy et al., 2019), doctor's medical notes (Huang et al., 2019) and biomedical journal articles (Lee et al., 2019). However, re-training BERT on the masked-language and next-sentence prediction tasks for every new domain is unwieldy and time-consuming. We investigate whether the benefits of re-training on a new domain can also be realized by fine-tuning BERT on other in-domain tasks. Phang et al. (2018) see a boost with other tasks across less dramatic domain changes, where a different text corpus is used for the final task but not an entirely different technical vocabulary or domain.

2.3 Medical Question-Answering and Question-Entailment

Previous work on question similarity has investigated the importance of in-domain word embeddings, but the methods of determining question similarity have varied widely. Bogdanova et al. (2015) use a CNN that minimizes the mean squared error between two questions' vector representations. Lei et al. (2016) use an encoder and a question-pairs metric of cosine similarity. Finally, Gupta (2019) look at relative 'best' match out of a set instead of absolute similarity, as we are interested in. Abacha & Demner-Fushman (2016; 2019) very clearly describe the utility of medical question similarity as we have framed it; rather than training a model to answer every conceivable medical question correctly, we can train a model to determine if any existing questions in an FAQ mean the same thing as a new question and, if so, use the existing answer for the new question.

3 Data

If there were a large corpus of labeled similar medical questions, training on that would likely produce the best results. However, labeled training data is still one of the largest barriers to supervised learning, particularly in the medical field where it is expensive to get doctor time for hand-labeling data. Previous work has tried to overcome this using augmentation rules to generate similar question pairs automatically (Li et al., 2018), but this leads to an overly simplistic dataset in which negative question-pairs contain no overlapping keywords and positive question-pairs follow similar lexical structures. Another technique for generating training data is weak supervision (Ratner et al., 2017), but due to the nuances of determining medical similarity, generating labeling functions for this task is difficult. Another way to address a dearth of training data is to use transfer learning from a different but related task. This is the path we choose.

3.1 Datasets

Several large datasets exist that are relevant to our final task of medical question similarity.

Quora Question Pairs (QQP) is a labeled corpus of 363,871 question pairs from Quora, an online question-answer forum (Csernai, 2017). These question pairs cover a broad range of topics, most of which are not related to medicine. However, it is a well-known dataset containing labeled pairs of similar and dissimilar questions.

HealthTap is a medical question-answering website in which patients can have their questions answered by doctors. We use a publicly available crawl (durakkerem, 2018) with 1.6 million medical questions. Each question has corresponding long and short answers, doctor meta-data, category labels, and lists of related topics. We reduce this dataset to match the size of QQP for direct performance comparisons, but also run one experiment leveraging the full corpus.

WebMD is an online publisher of medical information including articles, videos, and frequently asked questions (FAQ). For a second medical question-answer dataset, we use a publicly available crawl (Nielsen, 2017) over the FAQ of WebMD with 46,872 question-answer pairs. We decrease the size of QQP and HealthTap to match this number before making direct performance comparisons.

3.2 Pre-Training Tasks

Most of our pre-training tasks come from restructuring the HealthTap and WebMD data.

Question Answer Pairs (QA) In order to correctly determine whether or not two questions are semantically similar, as is our ultimate goal, a network must be able to interpret the nuances of each question. Another task that requires such nuanced understanding is that of pairing questions with their correct answers. We isolate each true question-answer pair from the medical question-answering websites and label these as positive examples. We then take each question and pair it with a random answer *from the same main category or tag* and label these as negative examples. Finally, we train a classifier to label question-answer pairs as either positive or negative.

Answer Completion (AA) One task that has been known to generalize well is that of next-sentence prediction, which is one of two tasks used to train the BERT model. To mimic this task, we take each answer from HealthTap and split it into two parts: the first two sentences (start), and the remaining

sentences (end). We then take each answer start and end that came from the same original question and label these pairs as positives. We also pair each answer start with a different end from the same main category and label these as negatives. This is therefore a binary classification task in which the model tries to predict whether an answer start is completed by the given answer end.

Question Categorization (QC) We take the questions from HealthTap, pair them up with their main-category labels and call these positive examples. We then pair each question with a random other category and call this a negative example. There are 227 main categories represented, such as abdominal pain, acid reflux, acne, adhd, alcohol, etc. The model is trained to classify category matches and mismatches, rather than predict to which of the classes each example belongs.

A small number of questions from HealthTap are used in the question pairs dataset and thus withheld from the above tasks to reduce bias.

4   MEDICAL QUESTION SIMILARITY DATASET RELEASE

There is no existing dataset that we know of for medical question similarity. Therefore, one contribution of this paper is that we have generated such a dataset and are releasing it. Although our task is related to that of recognizing question entailment (RQE), for which there is a small dataset available (MED, 2019), it is different in two key ways. First, our metric of similarity is symmetric, but in question entailment, it is possible to have asymmetric entailment, if one question is more specific than the other. Second, our questions are all patient-asked, which means they use less technical language, include more misspellings, and span a different range of topics than doctor-asked questions. Because of these differences we decide to generate a dataset that is specific to our needs.

We have doctors hand-generate 3,000 medical question pairs. We explicitly choose doctors for this task because determining whether or not two medical questions are the same requires medical training that crowd-sourced workers rarely have. We present doctors with a list of patient-asked questions from HealthTap, and for each provided question, ask them to:

1. Rewrite the original question in a different way while maintaining the same intent. Restructure the syntax as much as possible and change medical details that would not impact your response (ex. 'I'm a 22-y-o female' could become 'My 26 year old daughter').

2. Come up with a related but dissimilar question for which the answer to the original question would be WRONG OR IRRELEVANT. Use similar key words.

The first instruction generates a *positive* question pair (match) and the second generates a *negative* question pair (mismatch). With the above instructions, we intentionally frame the task such that positive question pairs can look very different by superficial metrics, and negative question pairs can conversely look very similar. This ensures that the task is not trivial.

Table 2: Examples from our data generation setup. Columns 1 and 2 are used to generate similar question pairs while columns 1 and 3 generate dissimilar question pairs.

Original Question	Similar Question	Different Question
If I had hepatitis a, does that mean I can't drink alcohol for a certain number of weeks afterwards?	How soon can I drink alcohol after being tested positive for Hep A?	Can Hep A spread via sharing cigarettes or alcohol bottles of an infected person?
Am I over weight (192.9) for my age (39)?	I am a 39 y/o male currently weighing about 193 lbs. Do you think I am overweight?	What diet is good for losing weight? Keto or vegan?
What specific exercises would help bursitis of the suprapatellar?	Hey doc! My doctor diagnosed me with suprapatellar bursitis. Are there any exercises that I can do at home?	Can I take any medication for pain due to suprapatellar bursitis? Unable to exercise. :(

We anticipate that each doctor interprets these instructions slightly differently, so no doctor providing data in the train set generates any data in the test set. This should reduce bias. To obtain an oracle score, we have doctors hand-label question pairs that a different doctor generated. The accuracy of the second doctor with respect to the labels intended by the first is used as an oracle and is 87.6% in our test set of 836 question pairs. See Table 2 for example questions from our curated data.

5 PROBLEM SETUP

Our ultimate goal is to be able to determine whether two medical questions mean the same thing. Our hypothesis is that by taking an existing language model with complex word embeddings and training it on a large corpus for a similar medical task, we can embed medical knowledge into an otherwise generic language model. Our approach uses transfer learning from a bi-directional transformer network to get the most out of our small medical question pairs dataset.

We start with the architecture and weights from BERT (Devlin et al., 2018) and perform a double finetune; we first finetune on an intermediate task and then we finetune on the final task of medical question pairs. We do this for four different intermediate tasks: quora question similarity (QQP), medical question answering (QA), medical answer completion (AA), and medical question classification (QC) (Figure 1). For a baseline we skip the intermediate finetune and directly train BERT on our small medical question-pairs dataset.
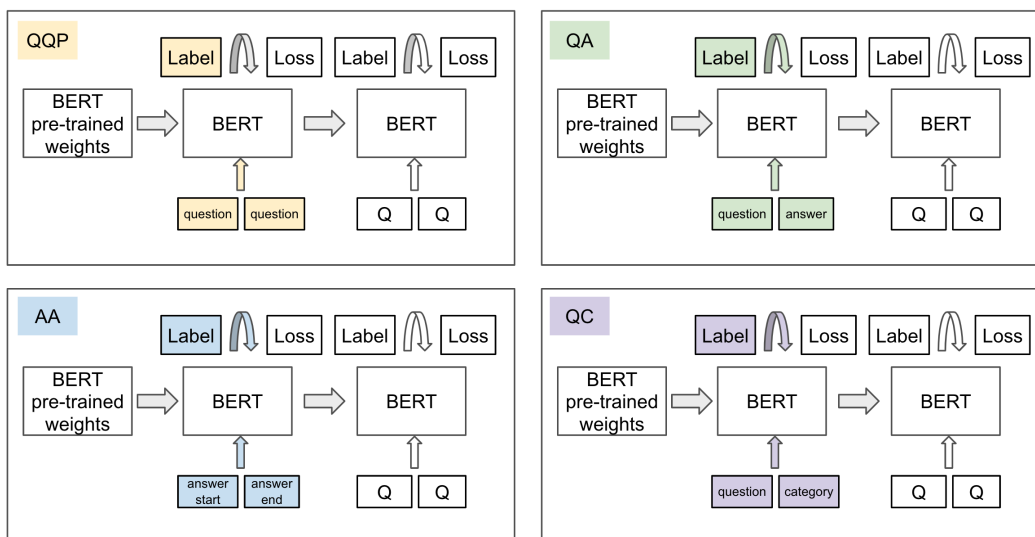

Figure 1: We perform a double finetune from BERT to an intermediate task to our medical question-similarity task for four different intermediate tasks: quora question-question pairs (top left), medical question-answer pairs (top right), medical answer-answer pairs (bottom left), and medical question-category pairs (bottom right)

For each intermediate task, we train the network for 5 epochs (Liu et al., 2019) with 364 thousand training examples to ensure that differences in performance are not due to different dataset sizes. We then finetune each of these intermediate-task-models on a small number of labeled, medical-question pairs until convergence. A maximum sentence length of 200 tokens, learning rate of 2e-5, and batch size of 16 is used for all models. Each model is trained on two parallel NVIDIA Tesla V100 GPUs. All experiments are done with 5 different random train/validation splits to generate error bars representing one standard deviation in accuracy. We use accuracy of each model as our quantitative metric for comparison and a paired t-test to measure statistical significance.

To compare against previous state-of-the-art (SOTA) models in the medical field, we also finetune the BioBERT (Lee et al., 2019), SciBERT (Beltagy et al., 2019), and ClinicalBERT (Huang et al., 2019) models on our final task, three BERT models that have been finetuned once already on the

original BERT tasks but with different text corpora. We also perform an ablation over pre-trained model architecture and reproduce our results starting with the XLNet model instead of BERT.

To get a better qualitative understanding of performance, we perform error analysis. We define a consistent error as one that is made by at least 4 of the 5 models trained on different train/validation splits. Similarly, we consider a model as getting an example consistently correct if it does so on at least 4 of the 5 models trained on different train/validation splits. By investigating the question pairs that a model-type gets consistently wrong, we can form hypotheses about why the model may have failed on that specific example. Then, by making small changes to the input until the models label those examples correctly, we can validate or disprove these hypotheses.

6 RESULTS

6.1 DOMAIN MATTERS

Table 3: Three sets of results comparing use of an in-domain question-answer task (QA) to an out-of-domain question-similarity task (QQP) for pre-training

Model	XLNet	BERT	BERT
Intermediate Train Set Size	364k	364k	27k
Baseline (No intermediate)	77.7% ±2.1%	78.5% ±1.3%	78.5% ±1.3%
Quora Question Pairs (QQP)	78.2% ±0.2%	78.2% ±0.8%	77.9% ±0.4%
HealthTap (QA)	**82.6%** ±0.8%	**81.6%** ±0.8%	78.3% ±0.7%
WebMD (QA)	– –	– –	**79.2%** ±1.2%

Here we investigate whether domain of the training corpus matters more than task-similarity when choosing an intermediate training step for the medical question similarity task. Accuracy on the final task (medical question similarity) is our quantitative proxy for performance.

Domain Similarity vs Task Similarity We finetune BERT on the intermediate tasks of Quora question pairs (QQP) and HealthTap question answer pairs (QA) before finetuning on the final task to compare performance. We find that the QA model performs better than the QQP model by 2.4% to 4.5%, depending on size of the final training set (Figure 2). Conducting a paired t-test over the 5 data splits used for each experiment, the p-value is always less than 0.0006, so this difference is very statistically significant. We thus see with high confidence that models trained on a related in-domain task (medical question-answering) outperform models trained on the same question-similarity task but an out-of-domain corpus (quora question pairs). Furthermore, when the full corpus of question-answer pairs from HealthTap is used, the performance climbs all the way to 84.5% ±0.7%.

Results hold across models The same trends hold when the BERT base model is replaced with XLNet, with a p-value of 0.0001 (Table 3).

Results hold across datasets We repeat our experiments with a question-answer dataset from WebMD and restrict the HealthTap and QQP dataset sizes for fair comparison. We find that the QA model again outperforms the QQP model by a statistically significant margin (p-value 0.049) and that the WebMD model even outperforms the HealthTap model with the same amount of data (Table 3). Our findings therefore hold across multiple in-domain datasets.

6.2 NOT ALL IN-DOMAIN TASKS EMBED RELEVANT MEDICAL KNOWLEDGE

We investigate further the extent to which task matters for an in-domain corpus in two different ways. We start by using the same HealthTap data and forming different tasks from the questions therein, and then we compare our models against intermediate models trained by other researchers.

To test the extent to which any in-domain task would boost the performance of an out-of-domain model, we design two additional tasks using the HealthTap data: answer completion (AA) and question categorization (QC). As before, we use accuracy on the final question-similarity task as our proxy for performance and keep the test set constant across all models. We follow the same protocol
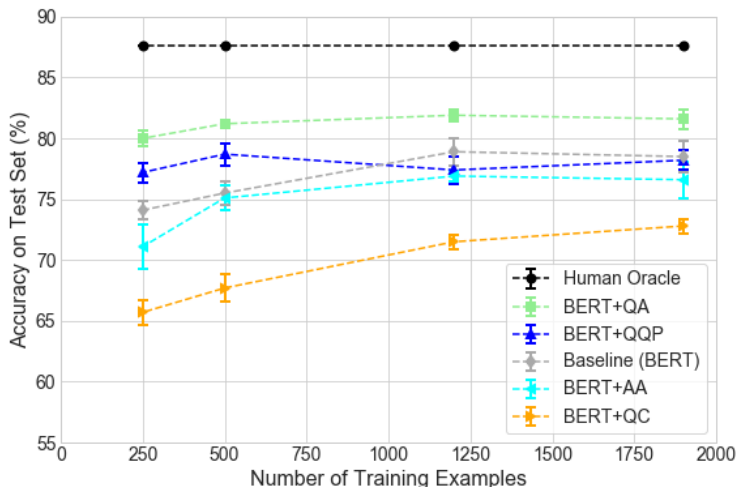
Figure 2: The intermediate task of training on question-answer pairs (BERT+QA) reliably outperforms other intermediate tasks: Quora question pairs (BERT+QQP), medical answer completion (BERT+AA), and medical question categorization (BERT+QC). Differences are exacerbated with fewer training examples. Error bars represent one standard deviation across different data splits.

as above, finetuning BERT on the intermediate task before finetuning further on the final task. We find that both of these tasks actually perform worse than the baseline BERT model, making the word embeddings less useful for understanding the subtler differences between two questions (Figure 2). We conclude that, while domain does matter a lot, many tasks are not well-suited to encoding the proper domain information from the in-domain corpus.

Comparison to Medical SOTA Models To benchmark ourselves against existing medical models, we compare our finetuned models to BioBERT, SciBERT, and ClinicalBERT. Each of these models has finetuned the original BERT weights on a medically relevant corpus using the original BERT tasks. We take each of these off-the-shelf models and finetune them on our final task dataset as we do with our own intermediate-task models. Only BioBERT outperforms the original BERT model, and the differences in performance are not statistically significant. We hypothesize that this is because technical literature and doctor notes each have their own vocabularies that, while more medical in nature that Wikipedia articles, are still quite distinct from those of medical question-answer forums.

7 ERROR ANALYSIS

From looking at the question pairs that our models get wrong, we can form hypotheses about why each example is mislabeled. We can then augment each question pair to add or remove one challenging aspect at a time and observe whether or not those changes result in a different label. With this method, we can prove or disprove our hypotheses. The augmented questions are not added to our test set and do not contribute to our quantitative performance metrics; they are only created for the sake of probing and understanding the network.

Consider the example in Table 4. In order to label this example correctly as it is written in row 1, the model has to understand the syntax of the question and know that 4'8" in this context represents poor growth. Changing the second question to what is written in row 2 prompts the QA model to label it correctly, indicating that one thing the QA model was misunderstanding was the question's word order. Additionally, changing the phrase *I am 4'8"* with *I have not grown* as shown in row 3 is enough to help the out-of-domain models label it correctly. So, while numerical reasoning was the difficult part of that question pair for the other models, the question answer model was actually able to identify 4'8" as a short height. This supports the claim that pre-training on the medical task of

Table 4: Example question pair that was augmented to reveal which aspects of the question pair the network failed to understand

	Question 1		Question 2			Label
	Baseline	QQP	QA	AA	QC	
1	Can coarctation of the aorta cause poor growth in height?		"I am 4' 8"". My mom said that I have a birth defect (coarctation of aorta). Are the two related?"			same
	wrong	wrong	wrong	wrong	wrong	
2	Can coarctation of the aorta cause poor growth in height?		My mom said that I have a birth defect (coarctation of aorta). Could that be the reason I am 4'8"?			same
	wrong	mixed	correct	wrong	mixed	
3	Can coarctation of the aorta cause poor growth in height?		My mom said that I have a birth defect (coarctation of aorta). Could that be the reason I have not grown?			same
	correct	correct	correct	correct	correct	

question answering embedded medically relevant information into the model that the out-of-domain models are still missing. More examples of errors we analyzed can be found in Appendix A.

Misspellings, capitalization We find that differences in spelling and capitalization do not cause a significant number of errors in any model, although they are present in many questions.

8 DISCUSSION: GENERALIZING TO OTHER (EXPERT) DOMAINS

To understand the broader applicability of our findings, we apply our approach to a non-medical domain: the AskUbuntu question-answer pairs from Lei et al. (2016). As before, we avoid making the pre-training task artificially easy by creating negatives from related questions. This time, since there are no category labels, we index all of the data with Elasticsearch[1]. For the question similarity task, the authors have released a candidate set of pairs that were human labeled as similar or dissimilar.

Without any pre-training (baseline), we observe an accuracy of $65.3\% \pm 1.2\%$ on the question similarity task. Pre-training on QQP leads to a significant reduction in accuracy to $62.3\% \pm 2.1\%$ indicating that an out-of-domain pretraining task can actually hurt performance. When the QA task is used for intermediate pre-training, the results improve to $66.6\% \pm 0.9\%$. While this improvement may not be statistically significant, it is consistent with the main premise of our work that related tasks in the same domain can help performance. We believe that the low accuracy on this task, as well as the small inter-model performance gains, may be due to the exceptionally long question lengths, some of which are truncated by the models during tokenization. In the future, we would explore ways to reduce the length of these questions before feeding them into the model.

9 CONCLUSIONS AND FUTURE WORK

In this work, we release a medical question-pairs dataset and show that the semi-supervised approach of pre-training on in-domain question-answer matching (QA) is particularly useful for the difficult task of duplicate question recognition. Although the QA model outperforms the out-of-domain same-task QQP model, there are a few examples where the QQP model seems to have learned information that is missing from the QA model (see Appendix A). In the future, we can further explore whether these two models learned independently useful information from their pre-training tasks. If they did, then we hope to be able to combine these features into one model with multi-task learning. An additional benefit of the error analysis is that we have a better understanding of the types of mistakes that even our best model is making. It is therefore now easier to use weak supervision and augmentation rules to supplement our datasets to increase the number of training examples in those difficult regions of the data. With both of these changes, we expect to be able to bump up accuracy on this task by several more percentage points.

---

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html

REFERENCES

Mediqa 2019 - recognizing question entailment (rqe), 2019. URL `https://www.aicrowd.com/challenges/mediqa-2019-recognizing-question-entailment-rqe`.

Asma Ben Abacha and Dina Demner-Fushman. Recognizing question entailment for medical question answering, 2016.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering, 2019.

Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: Pretrained contextualized embeddings for scientific text, 2019.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 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*, pp. 370–379, Florence, Italy, August 2019. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W19-5039`.

Dasha Bogdanova, Cícero dos Santos, Luciano Barbosa, and Bianca Zadrozny. Detecting semantically equivalent questions in online user forums. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 123–131, Beijing, China, July 2015. Association for Computational Linguistics. doi: 10.18653/v1/K15-1013. URL `https://www.aclweb.org/anthology/K15-1013`.

Kornel Csernai. Quora, Jan 2017. URL `https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs`.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

durakkerem. Medical question answer datasets, 2018. URL `https://github.com/durakkerem/Medical-Question-Answer-Datasets`.

Santosh Gupta. Docproduct. `https://github.com/re-search/DocProduct`, 2019.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission, 2019.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining, 2019.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Llus Mrquez. Semi-supervised question retrieval with gated convolutions. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016. doi: 10.18653/v1/n16-1153. URL `http://dx.doi.org/10.18653/v1/n16-1153`.

Yaliang Li, Liuyi Yao, Nan Du, Jing Gao, Qi Li, Chuishi Meng, Chenwei Zhang, and Wei Fan. Finding similar medical questions from question answering websites, 2018.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/p19-1441. URL `http://dx.doi.org/10.18653/v1/p19-1441`.

Lasse Regin Nielsen. Medical question answer data, 2017. URL `https://github.com/LasseRegin/medical-question-answer-data`.

John Nosta. The internet is making us lose trust in our doctors, May 2017. URL `https://www.forbes.com/sites/johnnosta/2017/05/08/the-internet-is-making-us-lose-trust-in-our-doctors/#51f429cf65fa`.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018. doi: 10.18653/v1/n18-1202. URL `http://dx.doi.org/10.18653/v1/n18-1202`.

Jason Phang, Thibault Fvry, and Samuel R. Bowman. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher R. Snorkel. *Proceedings of the VLDB Endowment*, 11(3):269282, Nov 2017. ISSN 2150-8097. doi: 10.14778/3157794.3157797. URL `http://dx.doi.org/10.14778/3157794.3157797`.

Alon Talmor and Jonathan Berant. Multiqa: An empirical investigation of generalization and transfer in reading comprehension, 2019.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2019.

A   APPENDIX: ERRORS

Table 5: Examples that were consistently labeled wrong by at least one model type. Patterns reveal the key differences in what is learned by each intermediate task

Question 1		Question 2			Label
Baseline	**QQP**	**QA**	**AA**	**QC**	
What to do if I have hypertension, depression of lost love, completely broke. what to do?		I recently lost my job and the love of my life. The RN who saw me yesterday said I have high blood pressure. I feel depressed, how do I get better?			same
wrong	wrong	correct	correct	wrong	
How many tylenol (acetaminophen) 3's are equal to a percocet?		I am out of Percocet, how many Tylenol pills should I take to give a similar amount of pain relief?			same
wrong	wrong	correct	wrong	wrong	
How do u know the days u ovulate if your period changes every month and how do I increase my chances of becomming pregnant?		How can one track one's ovulation with irregular periods?			same
wrong	wrong	correct	wrong	wrong	
How does a doughnut get broken down in the digestive system if it has essentially no nutrients?		Do doughnuts have any nutrient value or are they pure junk?			different
wrong	correct	wrong	wrong	wrong	
Is it contraindicated to insert a urinary catheter into a patient who has a myocardial infarction?		In case of urinary incontinence in a patient with myocardial infarction, what are the options to help the patient pass urine except a urinary catheter?			different
correct	correct	correct	wrong	wrong	
Can Celiac Disease mimic MS even after being treated with a strict gluten-free diet? Could my untreated Celiac have caused my brain lesions?		Does a strict gluten free diet help in complete control of Celiac disease?			different
correct	correct	correct	wrong	wrong	
At a doctor's visit, I hit my head against a box on the wall containing hazardous materials (used syringes, used needles...). Will I get infected?		Today morning, I had an appointment with the doctor. After that, I saw the nurse for a shot. As I got up, her box syringes and needles fell all over me. She sent me home saying all is okay. But I am worried about an infection. Should I schedule another appointment for any blood tests			same
wrong	wrong	wrong	wrong	wrong	

B APPENDIX: FULL EXPERIMENTAL RESULTS

Set Up	Intermediate Task	Split 1	Split 2	Split 3	Split 4	Split 5
Train Set 1.9k bert-base-cased	None	78.0	78.9	79.9	79.3	76.5
	QQP	78.2	76.9	78.0	78.6	79.2
	QA	**80.8**	**81.2**	**82.8**	**81.4**	**81.6**
	AA	76.6	74.1	77.4	78.2	76.5
	QC	72.3	72.9	73.2	73.5	72.1
	QA (full corpus)	**84.8**	**84.6**	**85.5**	**83.7**	**84.0**
Train Set 1.2k bert-base-cased	None	77.8	80.4	79.8	78.3	78.1
	QQP	79.0	76.9	77.0	76.2	78.1
	QA	**81.7**	**82.5**	**81.7**	**81.3**	**82.4**
	AA	77.1	77.5	76.6	76.8	76.5
	QC	71.5	71.1	71.5	72.5	71.1
Train Set 500 bert-base-cased	None	73.9	75.8	75.0	76.3	76.3
	QQP	78.2	78.9	77.6	78.9	80.1
	QA	**81.4**	**81.1**	**81.6**	**81.0**	**80.8**
	AA	74.0	74.9	75.7	74.6	76.5
	QC	66.0	68.7	67.5	67.5	68.7
Train Set 250 bert-base-cased	None	74.0	73.8	74.1	73.3	75.1
	QQP	76.8	76.2	76.9	77.8	78.2
	QA	**80.2**	**80.8**	**79.2**	**79.9**	**80.0**
	AA	70.4	68.9	71.4	71.1	73.9
	QC	65.9	67.1	65.9	65.3	64.3
Train Set 1.9k xlnet-base-cased	None	78.9	79.3	74.1	78.3	78.0
	QQP	78.0	78.2	78.3	78.0	78.4
	QA	**81.6**	**82.4**	**82.6**	**83.7**	**82.8**
	AA	78.3	77.2	77.8	75.7	73.7
	QC	67.3	66.1	66.9	70.4	69.0
Train Set 1.9k bert-base-cased	QA WebMD	**78.2**	78.1	**80.6**	78.9	**80.4**
	QA HealthTap (small)	78.2	**78.7**	77.7	**79.3**	77.8
	QQP (small)	77.4	77.5	78.4	78.2	78.1
Train Set 1.9k bert-base-cased	BioBERT	79.2	77.6	78.0	78.3	79.3
	SciBERT	76.4	76.3	74.0	77.1	75.4
	ClinicalBERT	75.4	73.5	76.3	73.8	71.9