CS230

Retrieving Duplicates in the Extensive Set of Medical Questions

Alexander Sivura
SCPD
Stanford University
asivura@stanford.edu

1 Problem Description

HealthTap[1] allows its members to ask medical questions and get answers from US doctors for more than ten years. During that time, a knowledge base with 1,896,988 of medical patient questions and 2,661,762 doctor answers was collected. Many of these questions duplicate each other and require doctors to spend their time answering them again. Duplicated questions make it difficult for users to search content on the website when they get many pages with repeated questions for their search queries. Product feature for identifying duplicated questions can serve users with answers almost immediately by recycling answers to repeated questions for the new user question user wasting the time of doctors on it. Other websites and medical institutions that built databases with medical questions and answers (like ADAM[2], WebMD[3]) could face similar problems.

2 Datasets

The same problem has been tackled for general questions in the context of online user forums [1], community QA [2], and question-answer archives [3]. Models trained on the datasets with general questions do not perform well for medical question similarity [4]. Another challenge for solving this is that medical question can imbibe a lot of information what is critical for the answer. So changing even a single word in the question can make an answer completely irrelevant. For instance, a pair of questions, "Do symptoms like atopic dermatitis go away quickly?" and "Do symptoms like urticaria go away quickly?" are dissimilar despite being different in only a single word.

2.1 Medical Question Pair (MQP) dataset

Medical Question Pair (MQP) dataset[4] is the only publicly available dataset for the medical question similarity task. This dataset contains 3048 similar and dissimilar medical question pairs hand-generated and labeled by Curai's doctors. Authors of this dataset randomly sampled a list of 1524 patient-asked questions of HealthTap. They asked doctors to rewrite each question in two different ways to generate a positive question pair (similar) and a negative question pair (different) for each question from the list.

In this work, we trained models on the MQP dataset to find duplicates in the HealthTap user questions. Unfortunately, these models do not perform well on medical question pairs where both questions in the pair are medical questions from HealthTap users. With the support of HealthTap doctors, we

[1] www.healthtap.com
[2] www.adam.com
[3] www.webmd.com
[4] github.com/curai/medical-question-pair-dataset

created a new dataset by annotating candidates for duplicated questions from the complete set of HealthTap questions.

2.2 HealthTap Medical Question Pair (HT-MQP) dataset

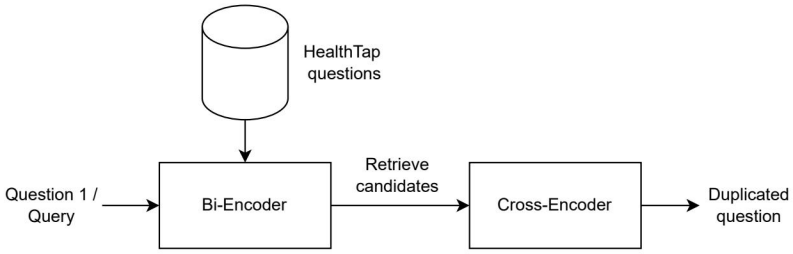
We built a new dataset where both questions in each pair are real user questions to. We randomly sampled a list of 10,000 questions from the full set of HealthTap questions. We selected a few candidates with top scores based on the model trained on the MQP dataset for each question from this list. Doctors were asked to put a positive label for a question pair only if questions are identical and have the same answer. Detailed annotation guidelines with examples are presented in the appendix A.

HealthTap doctors annotated 19,167 question pairs. To make HT-MQP similar to the MQP dataset, we included in the MQP-HT only a single positive pair and only a single negative pair for each question we sampled initially and excluded all questions where we have only positive or negative examples. It makes HT-MQP balanced dataset and evaluation metrics more objective. So HT-MQP contains 997 positive examples and 997 negative examples. We put the rest of the annotated question pairs (6,429 positive and 10,744 negative) into the HT-MQP-EXT dataset. HT-MQP-EXT imbalanced dataset can have many positive and negative pairs for each question. HT-MQP-EXT, together with the Curai MQP dataset, can be used for the initial fine-tuning model to improve the model performance on the HT-MQP.

3 Methods and Experiments

We created a pipeline (fig. 1) of two models, retrieval Bi-Encoder and re-ranker Cross-Encoder, to get duplicated questions for a given question from the extensive set of HealthTap medical questions.

Figure 1: Pipeline for duplicated questions retrieval



3.1 Bi-Encoder

Bi-Encoder is a model for obtaining embeddings for medical questions. We feed the query question into the BioLinkBERT [5] transformer network for producing contextualized word embeddings for all input tokens in the question. We added a mean-pooling layer to average all contextualized word embeddings to get a fixed-sized output representation. Pooling layer gives us a fixed 768 dimensional output vector independent of how long our input question was. To find duplicated questions, we use cosine similarity for question embeddings. We train the model using a Siamese Network Architecture (fig. 2) [6]. For each question pair in the dataset, we pass question 1 and question 2 through to obtain embeddings u and v. The cosine similarity of these embeddings gives us a prediction score if pair of questions are duplicates. We fine-tune the model with $ContrastiveLoss$ on the training question pairs:

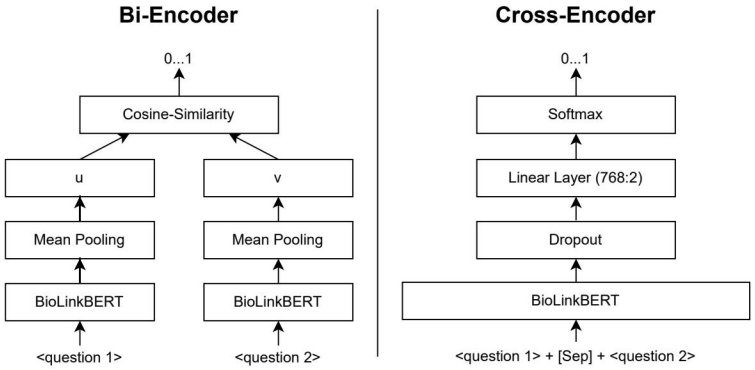
$$ContrastiveLoss = y(1-d)^2 + (1-y)max(margin - d, 0)^2, \quad (1)$$

where distance function d is $1 - cosine\_similarity(u, v)$. The $margin$ is a model hyper-parameter which represents a minimal distance for negative examples.

We precalculate embeddings for all HealthTap questions with a bi-encoder model to retrieve all duplicates for a given question efficiently, as we only need to calculate embedding for a query question and calculate $cosine\_similarity$ between that embedding and embeddings for all HealthTap

questions. It takes just a few seconds to retrieve all duplicated questions out of 1.8M HealthTap questions for a given question with precalculated embeddings.

Figure 2: Bi-Encoder and Cross-Encoder architecture



3.2 Cross-Encoder

In contrast to Bi-Encoder, for a Cross-Encoder, we pass both questions simultaneously to the Transformer network by concatenating them with a [SEP] token. We send the output of the classification head of the transformer output to the linear layer with input size 768 and output size 2. We apply the softmax function to the outputs of the linear layer to get predictions. To train this model, we use $BinaryCrossEntropy$ loss.

Cross-Encoder models outperform Bi-Encoder models, but they work very slowly if we need to select duplicated questions from the set of 1.8M of HealthTap questions. It can take more than one hour on the GPU to get 1M of predictions, making these models alone unsuitable for our task. We use them to re-rank candidates from Bi-Encoders.

3.3 Experimental setup

We use 20% of records from the dataset to test the models. 80% of records we use to train the models using a 4-fold cross-validation technique. In other words, we train each model 4 times and use one fold as the development dataset and three folds as the training dataset. We report the average value for predictions from the four models trained during the cross-validation for all metrics.

We use F1 as a target metric to compare models between each other and choose the best model during training. Additionally, we report Accuracy, Precision, and Recall. [7]. During training, we find the thresholds for predictions that give the best F1 and Accuracy for the validation dataset. When we evaluate models against the test dataset, we use thresholds obtained on the development dataset.

For cross-encoder models we use a batch size 16, AdamW optimizer with learning rate 2e-5 and a linear learning rate warm-up over 500 training steps. For bi-encoder models we use a batch size 16, AdamW optimizer with learning rate 2e-5 and a linear learning rate over 1000 training steps. For the $ContrastiveLoss$ margin we use 0.3

All bi-encoder(BE) models bases on the BioLinkBERT-base transformer (110M params). Cross-encoder-base (CE-base) models bases on the BioLinkBERT-base transformer. Cross-encoder large (CE-large) models bases on the BioLinkBERT-large transformer (340M params)

3.4 Train models on the MQP dataset

To retrieve question pairs for the HT-MQP dataset annotation from the complete set of HealthTap questions, we trained models on the MQP dataset. After we finalized the HT-MQP dataset, we evaluated these models against a new dataset where both questions are real user questions (table 1). Precision on the HT-MQP drops significantly compared to the MQP dataset (for the CE-large model

from 0.88 to 0.585), which gives us much lower f1 scores. We assume it happens because examples in the MQP dataset are not as hard as those from real user questions. The models trained on the MQP dataset could learn patterns of how doctors artificially modify questions to make them different or duplicated.

Table 1: Performance of models trained on the MQP dataset and evaluated on the MQP test and HT-MQP test datasets

Model	F1		Precision		Recall		Accuracy	
	mqp	ht-mqp	mqp	ht-mqp	mqp	ht-mqp	mqp	ht-mqp
BE	0.819	0.707	0.77	0.549	0.896	0.995	0.803	0.597
CE-base	0.86	0.727	0.856	0.574	0.865	0.991	0.857	0.638
CE-large	0.888	0.736	0.884	0.585	0.892	0.991	0.883	0.651

3.5 Train models on the HT-MQP and HT-MQP-Ext datasets

We trained models on the HT-MQP dataset with and without pretraining on the extended dataset. We used HT-MQP-EXT and MQP datasets (20,021 examples) for training and HT-MQP validation for the pretraining stage. Results are present in the table 2. Bi-encoder and cross-encoder significantly outperform models after training on the HT-MQP dataset compared to training on the MQP dataset, as we use data from the same distribution for training. The pretraining stage with an imbalanced HT-MQP-EXT dataset significantly increases F1 score for BE and CE-large models, and the improvement for the CE-base model is relatively small. BE model has a higher recall than CE-large models and lower precision. Hence, combining the BE with the CE-large in the pipeline, where we use BE for selecting candidates and CE-large for the final predictions, will work well for retrieving duplicated questions from the big set of questions.

Table 2: Performance of models trained on the HT-MQP dataset with pretraining on the HT-MQP-EXT and without and evaluated on the HT-MQP test dataset.

Model	F1	Precision	Recall	Accuracy
BE	0.763	0.661	0.911	0.742
BE (pretrained)	0.807	0.729	0.905	0.793
CE-base	0.797	0.753	0.849	0.798
CE-base (pretrained)	0.813	0.833	0.795	0.819
CE-large	0.809	0.758	0.867	0.796
CE-large (pretrained HT-MQP-EXT)	0.842	0.832	0.853	0.838

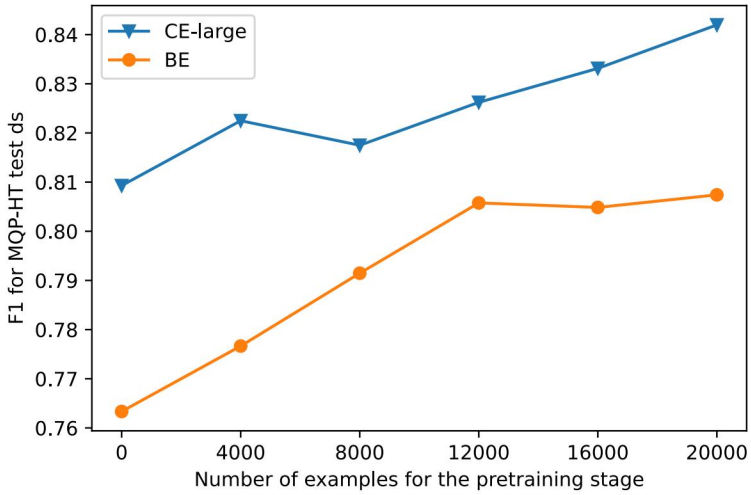
3.6 Pretraining dataset size

To study the impact of the dataset size we use for pretraining to F1 score, we pretrained BE and CE-large models with reduced HT-MQP-Ext datasets (figure 3). We can observe an increasing trend for F1 scores with an increasing number of dataset examples for pretraining, especially for the CE-large model, where the dependency is almost linear. Increasing the dataset size can improve the F1 score significantly.

3.7 Qualitative Analysis

In the table 3 we present some examples for bi-encoder and cross-encoder models from the validation dataset. There is a typo in the word "Whar" in question pair 9. Likely it confuses the models to give correct predictions. When we fix the type CE-large model gives the correct prediction. The BE model sometimes confuses different medications for the same treatment, as in example 3. Questions in example 1 are labeled as different, but they might involve the judgment call of the annotator; others can treat this example as an exact duplicate. Examples 2 and 5 are hard for all models as questions

Figure 3: Impact of the dataset size for the pretraining stage to F1 score for BE and CE-large models



look very similar, but answers will be different. For example, there is a difference from the medical point of view between "every meal" and "breakfast," which could confuse all models. It might be hard to say that these pair of questions are different for people without medical expertise. Some medical terms can subsume others, like "herpes" and "genital herpes" in example 10, and it confuses BE model, but CE-large gives the correct prediction. Due to the limited time for the project, we did only the initial review of the model predictions. We plan to do it more systematically later by categorizing wrong predictions into categories.

4 Conclusions and Future Work

We released HT-MQP and HT-MQP-Ext datasets of medical question pairs sampled from the HealthTap user questions in this work. We demonstrated that models trained on artificially created datasets like MQP, where doctors are asked to modify questions to create different and duplicated question pairs, lead to a significant drop in the model performance when we evaluate them on the real question pairs.

We built a pipeline to retrieve duplicated questions from the large set of questions by combining a fast bi-encoder model to retrieve candidate questions for duplicates and another slow cross-encoder model to re-evaluate retrieved candidate questions for the final output.

We demonstrated that a bigger dataset for model pretraining leads to a higher F1 score. Increasing the size of the dataset for pretraining could be the easiest way to improve the model performance.

We did only initial analyzes of predictions. Splitting wrong predictions by categories and adding more training examples from those categories could improve performance. The performance of bi-encoder models is lower than cross-encoder models. We can create a dataset with weak labels from the cross-encoder models for the additional step of fine-tuning bi-encoder models to improve their performance.

References

[1] 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*, pages 123–131, Beijing, China, July 2015. Association for Computational Linguistics.

[2] Kai Zhang, Wei Wu, Haocheng Wu, Zhoujun Li, Ming Zhou, Wei Wu, and Ming Zhou. Question retrieval with high quality answers in community question answering. In *Proceedings of the*

Table 3: Examples of model predictions

	Question 1		Question 2	Label
	Bi-Encoder	**Cross-Encoder-base**	**Cross-Encoder-large**	
1	Is the heimlich maneuver always safe for everyone?		Can having the heimlich maneuver done be dangerous?	diff
	wrong	correct	wrong	
2	I got some deodorant in eye what can I do?!?		What will deodorant do if it got into the eye?	diff
	wrong	wrong	wrong	
3	Is triamcinolone acetonide a good steroid for poison ivy?		Is permethrin cream a good steroid for poison ivy?	diff
	wrong	correct	correct	
4	What causes loss of bladder control?		Why causes sudden loss of bladder control?	diff
	wrong	correct	correct	
5	Would drinking too much water affect pulmonary edema?		Could increased water consumption cause pulmonary edema?	diff
	wrong	correct	wrong	
6	Please explain why does my stomach hurt after every meal eat?		For what reason does my stomach hurt after i eat breakfast?	diff
	wrong	wrong	wrong	
7	Can psoriasis be hereditary?		How can psoriasis be transmitted?	diff
	wrong	correct	correct	
8	Is it possible that bladder cancer wouldn't show up on an ultrasound?		Is there a possibility bladder cancer would not show up on a sonogram of the bladder?	dup
	correct	correct	correct	
9	Whar are some natural foods that will increase my metabolics?		I am trying to eat healthy what are some good foods that would increase metabolism?	dup
	wrong	wrong	wrong	
10	What will help relieve genital herpes symptoms?		What will help to treat herpes?	diff
	wrong	wrong	correct	
11	I burned my lip (not by sunburn) and there's a blister. What can I do to treat this?		What's to be done about burn blister on lip?	dup
	wrong	correct	correct	

23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM'14), pages 371–380. ACM, November 2014.

[3] Zongcheng Ji, Fei Xu, Bin Wang, and Ben He. Question-answer topic model for question retrieval in community question answering. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, page 2471–2474, New York, NY, USA, 2012. Association for Computing Machinery.

[4] Clara H. McCreery, Namit Katariya, Anitha Kannan, Manish Chablani, and Xavier Amatriain. Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery amp; Data Mining*, KDD '20, page 3458–3465, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Michihiro Yasunaga, Jure Leskovec, and Percy Liang. Linkbert: Pretraining language models with document links. 2022.

[6] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Appendices

A HT-MQP Annotation Guidelines

A.1 Duplicate "dup"

- Questions that are identical in meaning.
- Questions that have exactly the same answer because they mean the same thing – ie, the answer to one can be given as-is as the answer to the other.
- Operational test: all medical topics in one question are found in the other and the answer to one will be the same as the answer to the other.

Examples of near duplicate question pairs are present in the table 4

Table 4: Examples of duplicates

Question 1	Question 2	Label	Explanation
How to get rid of migraine?	What can I do to get rid of my migraine headaches ?	dup	Same topic in both Q (migraine) Users are asking about the same thing Answer will be the same for both Q
Why is it bad to consume alcohol while on antibiotics?	Why can't you drink alcohol when taking antibiotics?	dup	Same topic in both Q (Alcohol consumption while taking antibiotic) Users are asking about the same thing using different words Answer will be the same for both Q
Hello, what is the recommended daily intake of salt. Thank you?	What's the daily recommendation for sodium intake in a healthy person?	dup	Same topic in both Q ('Daily salt intake') Users are asking about the same thing using different words. Even the clarification in the second Q about 'healthy person' doesn't make it different from the first one. Answer will be the same for both Q
Fastest way to get rid of toe nail fungus?	How can I get rid of toenail fungus?	dup	Same topic in both Q ('toenail fungus') Users are asking about the same thing Answer will be the same for both Q

A.2 Near duplicate "ndup"

- Questions that look almost the same, but have subtle or small differences that make them nearly the same, but enough difference that the answer to the one would not be used verbatim as the answer to the other.
- Operational test: All significant medical topics in one question are found in the other but minor variation makes the answer possibly different.

Examples of near duplicate question pairs are present in the table 5

A.3 Related "rel"

- Questions that are on similar topics or asking about a variation of the question. It's a different question, but directly related to the interests of the person who asked the other question.
- Operational test: The single most important or most specific medical topic is the same OR are related as siblings or parent-child. IE, similar topics in one question are found in the other.

Examples of related question pairs are present in the table 6

Table 5: Examples of near duplicates

Question 1	Question 2	Label	Explanation
Can hydrocele cause infertility and what is the percentage.?	Will having a hydrocele make me infertile?	ndup	Same topic in both Q (infertility due to hydrocele) Users are asking about the same thing but the first question more detailed and includes clarifying question Answers will be possibly different
How can i get rid of acne scars?	I have scars and dark spots in some areas. How do I get rid of acne scars and dark spots?	ndup	Same main topic in both Q (acne scars) Users are asking about the same thing but the second Q includes additional question about dark spots Answers will be possibly different

Table 6: Examples of related question pairs

Question 1	Question 2	Label	Explanation
What are some cause's for cll?	How concerning is cll?	rel	Same topic in both Q (CLL) Users are asking about different things inside one topic
How to prevent motion sickness ?	Any tips on getting rid of motion sickness?	rel	Same topic in both Q (motion sickness) Users are asking about different things inside one topic Answers will be different
What are some causes of dry mouth?	What to do about dry mouth?	rel	Same topic in both Q (dry mouth) Users are asking about different things inside one topic Answers will be different
Is there an effective treatment for premature ejaculation?	What medications ar there for premature ejaculation?	rel	Same topic in both Q (premature ejaculation) Users are asking about the same thing but the first question more detailed and includes clarifying question Answers will be possibly different

A.4 Different "diff"

- Questions for which the core meaning or topic is different, so they are simply different questions, even though they may share many words or structure. The author of one is not particularly likely to be interested in the other question.
- The single most important or most specific medical topic is different.

Examples of different question pairs are present in the table 7

Table 7: Examples of Different question pairs

Question 1	Question 2	Label	Explanation
Do lots of people get disseminated intravascular coagulopathy, or is it unusual?	Do lots of people get pancoast syndrome, or is it unusual?	diff	different conditions
My mom has a major problem with her wrist. What kind of doctor should she see?	I've been having pain under my left armpit. What doctor should I see?	diff	different conditions same question what type of doctor to see however the answer will be different
I have been diagnosed with copd, why do I want to sleep all the time?	Why do I sleep all the time?	diff	Same topic in both Q (dry mouth) copd related sleepiness from fatigue vs. all causes of sleepiness