Focus-Driven Contrastive Learning for Medical Question Summarization

Anonymous ACL submission

Abstract

Automatic medical question summarization can significantly help the system to understand consumer health questions and retrieve correct answers. The Seq2Seq model based on maximum likelihood estimation (MLE) has been applied in this task, which faces two general problems: the model can not capture well question focus and and the traditional MLE strategy lacks the ability to understand sentence-level semantics. To alleviate these problems, we propose a novel question focus-driven contrastive learning framework (QFCL). Specially, we propose an easy and effective approach to generate hard negative samples based on the question focus, and exploit contrastive learning at both encoder and decoder to obtain better sentence-level representations. On three medical benchmark datasets, our proposed model achieves new state-of-the-art results, and obtains a performance gain of 12.2%, 28.7% and 9.6% over the baseline BART model on three datasets respectively. Further human judgement and detailed analysis prove that our QFCL model learns better sentence representations with the ability to distinguish different sentence meanings, and generates high-quality summaries by capturing question focus.

1 Introduction

A growing number of health questions are raised by consumers on websites nowadays, which are usually written in natural language and including detailed and peripheral information not related to the answers. Summaries of such questions can greatly improve the performance in retrieving relevant answers (Ben Abacha and Demner-Fushman, 2019). Accordingly, the medical question summarization task is defined as summarizing the consumer health questions (CHQ) into frequently asked questions (FAQ), which are shorter but remain essential information of the original question to get correct answers. An example of medical question summarization is shown in Table 1.

Input question: consumer health question (CHQ) subject: gender dysphoria message: no health care on my son suffering from gender dysphoria what can we do to help him he worked out of high school no problems now not working and about shutting himself in his room 24/7 theres nothing this condition in our area we live in [location].no help in area what can we do he has had bad thoughts already please help us with some sort of info thank yuo [name] [location]
Golden summary: frequently asked question (FAQ): Where can I find information on treatment and resources for gender dysphoria?
Summary by BART (baseline): What are the treatments for weight loss?
Summary by our model: What are the treatments for gender dysphoria?

Table 1: An example of medical question summarization in MeqSum dataset, where the question focus is highlighted in green. Summaries generated by BART and our model are also listed.

The Seq2Seq neural models have been widely used in abstractive summarization (Nallapati et al., 2016; Lewis et al., 2020; Zhang et al., 2020) and show promising potentials, and they have also been applied in medical question summarization and achieve current state-of-the-art results. Ben Abacha and Demner-Fushman (2019) apply the pointer-generator model for this task. Yadav et al. (2021a) present a reinforcement learning framework with question-type identification reward and question-focus recognition reward. Mrini et al. (2021b) propose a multitask learning method by treating recognizing question entailment as an auxiliary task.

In the medical question summarization task, the input question CHQ is always lengthy and contains redundant information, where some salient medical entities and the semantic focus of question are vital to understand users' intention. But it still remains a challenging task for the existing methods to capture the question focus. As described in the example 1, the focus "*gender dysphoria*" is mis-replaced by "*weight loss*" in the summary generated by the

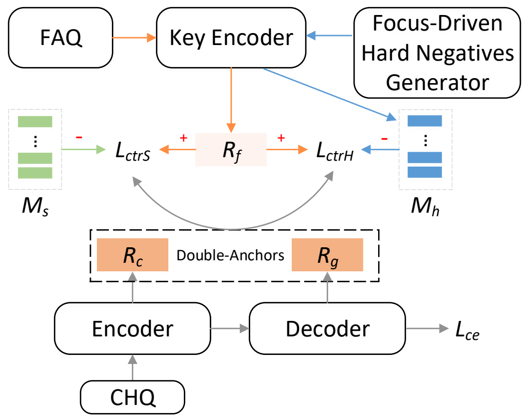

Figure 1: Sketch of our proposed contrastive learning framework. M_s,M_h represents the memory bank that contains simple negative samples and hard negative samples respectively. $\mathcal{R}_f$, $\mathcal{R}_c$, $\mathcal{R}_g$ denotes the sentence representation of FAQ, CHQ and generated summary. $\mathcal{L}_{ctrS}$ and $\mathcal{L}_{ctrH}$ are contrast learning loss on simple negative samples and hard negative samples respectively. $+$ indicates the positive sample, and $-$ indicates the negative sample.

fine-tuned BART, resulting in a complete different meaning from the original sentence.

For the medical question summarization task, the generated question summary is required to semantically close to the reference question. However, in most of current pre-trained models such as BART (Lewis et al., 2020) or Pegasus (Zhang et al., 2020), the model adopts maximum likelihood estimation (MLE) and mainly focuses on the accuracy of the prediction of masked tokens, but does not guarantee to the semantic similarity or dissimilarity of the whole sentences. To address this issue, some previous works adopt reinforcement learning (RL) in text summarization task (Li et al., 2019; Paulus et al., 2018), but RL suffers from the noise gradient estimation problem (Greensmith et al., 2004), which makes the training process unstable and sensitive to hyper-parameters.

To alleviate these problems, we propose a novel question focus-driven contrastive learning (QFCL) framework for medical question summarization, as illustrated in Figure 1. In our model, we introduce a "*double anchors*" strategy for contrast learning, by utilizing the sentence representation of CHQ as an anchor and the generated summary as another anchor, and regarding the golden reference FAQ as the positive sample. In addition, we present a "*focus-driven hard negatives generator*" to construct hard negative samples, by replacing the focus phrases with other phrases sharing the same attribute.

Through contrast learning, we minimize the distance between CHQ/generated summary and golden reference, and maximize the distance between CHQ/generated summary and other negative samples. By using the *double anchors*, our model is able to extract sentence-level semantic features to alleviate the problem of MLE. With the help of *hard negatives generator*, the model learns to pay more attention to question focus and thus produces high quality summary.

We conduct extensive experiments on three medical question summarization datasets: Meqsum (Ben Abacha and Demner-Fushman, 2019), HealthCareMagic and iCliniq (Zeng et al., 2020). Our proposed model outperforms previous best results by a wide margin, achieving new state-of-the-art results on all three datasets. Compared with the baseline BART, our model brings a relative performance gain of 12.2%, 28.7% and 9.6% on Meqsum, Cliniq and HealthcareMagic respectively. Through analysis, we prove that our model significantly gains the power of distinguishing the semantics between generated summaries and negative samples, and our model generates high-quality summaries capturing more question focuses.

2 Ralated Work

2.1 Medical Question Summarization

The medical question summarization task is defined by Ben Abacha and Demner-Fushman (2019). They construct a benchmark dataset Meqsum, and apply a pointer-generator model to generate question summary. At the question summarization campaign of MEDIQA-21 organized by Ben Abacha et al. (2021), almost all approaches rely on the fine-tuning of pre-trained transformer models. Transfer learning, knowledge-base, and ensemble methods are widely utilized by participanting teams to achieve better performance (He et al., 2021; Yadav et al., 2021b; Mrini et al., 2021c; Sänger et al., 2021). In this paper, we also base our method on the strong pre-trained BART model.

Recently, Yadav et al. (2021a) propose a RL framework with two question-aware semantic rewards: question-type identification reward (QTR) and question-focus recognition reward (QFR). QTR is to identify whether the question types

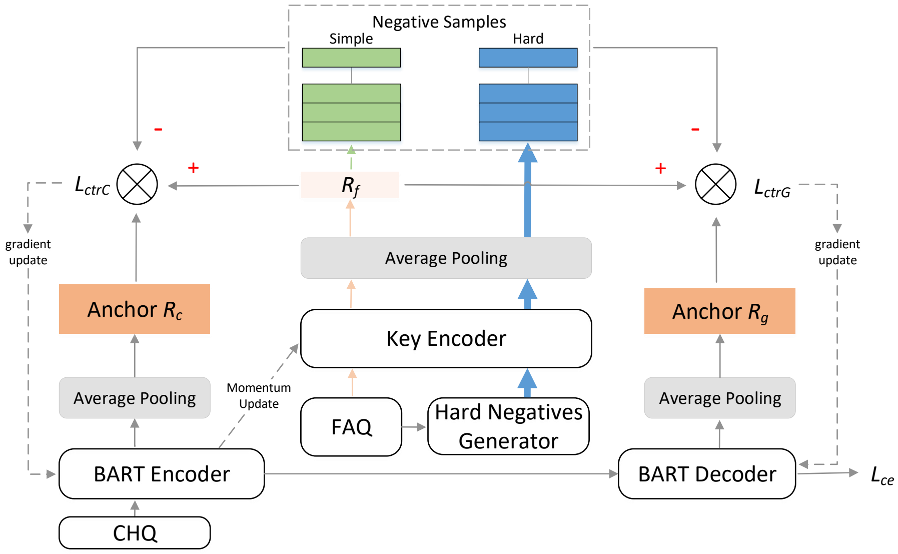

Figure 2: The overall framework of QFCL. $\mathcal{L}_{ctrC}$ and $\mathcal{L}_{ctrG}$ are contrast learning loss on the two anchors respectively.

are consistent with the gold question, and QFR is designed to capture question focus. But in their work, the question types and question focuses in the dataset should be manually labeled, which is both time-consuming and labor-intensive for large-scale datasets such as HealthcareMagic and iCliniq. Moreover, the RL training process is unstable. Mrini et al. (2021b) claim an equivalence between medical question summary and recognizing question entailment(RQE), and employ multi-task learning to train the model to not only perform next-word-prediction but also carry question entailment recognition. These two studies demonstrate that the pre-trained models achieve better performance after capturing the underlying sentence semantics of generated questions. Different from these works, we exploit contrastive learning to obtain focus-aware question representations.

2.2 Contrastive Learning

Different from the traditional methods which learn representations in pixel-level for computer vision tasks, contrastive learning encodes high-level features to distinguish different objects and has achieved great success (Henaff, 2020; Chen et al., 2020; Misra and van der Maaten, 2020; He et al., 2020), and it has also been applied in several NLP tasks such as machine translation (Pan et al., 2021), pre-training (Chi et al., 2021) and question answering (Yang et al., 2021).

In contrastive learning, the memory bank (Wu et al., 2018; Tian et al., 2020) is introduced to store large volumes of negative samples. Chen et al. (2020) prove that a large batch size will improve the performance of contrastive learning, but it will bring heavy burden on computation cost. To address this issue, He et al. (2020) propose MoCo, which maintains a queue as the memory bank to store negative samples. MoCo adopts two encoders with the same structure: a key encoder and a query encoder, where the key encoder is momentum updated from the query encoder. In our work, we apply MoCo for efficient contrastive learning.

3 Model

Given an input question CHQ, which is written by consumers and contains lengthy and complex information, the medical question summarization task aims to automatically generate a question summary that is a frequently asked question (FAQ), capturing the essential information to help efficiently retrieve correct answers. A more detailed structure of our proposed QFCL model is presented in Figure 2.

3.1 Contrastive Learning Architecture

We employ the pre-trained BART (Lewis et al., 2020) as our basic model to generate question summaries. For contrastive learning, we adopt the MoCo architecure (He et al., 2020), which contains a key encoder E_k with the same structure as the BART encoder E_q, and a queue to store simple negative samples with large volume. The simple negative samples in the queue are progressively replaced by current mini-batch of representations extracted from the key encoder. In addition, QFCL employs a *hard negatives generator* to generate hard negative samples.

In our model, the BART encoder E_q and the decoder are updated via back propagation by combining three types of loss functions, as described in the subsequent sections. The parameters of E_k are frozen and updated slowly towards that of E_q:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{1}$$

where m is a momentum coefficient.

At the inference, only the BART encoder and decoder are retained, but other parts such as the key encoder, the queue, and the *hard negatives generator* are all discarded, which are only used to optimize the sentence representations in the training stage.

3.2 Simple Negative Samples

In the medical question summarization task, the input question CHQ should be semantically close to its reference summary FAQ but different from other question summaries. Therefore, we regard the CHQ c_i in the i-th pair as the anchor, FAQ f_i in the same pair as the positive sample and randomly select f_j from other different pairs to serve as simple negative samples.

Let $\mathcal{R}_s$ denote the average decoded output of an arbitrary sentence s, the objective function of the simple contrastive learning is defined as:

$$\mathcal{L}_{ctrCS} = -log \frac{e^{sim(\mathcal{R}_{ci},\mathcal{R}_{fi})/\tau}}{\sum_{\mathcal{R}_{fj}\in M_s} e^{sim(\mathcal{R}_{ci},\mathcal{R}_{fj})/\tau}} \tag{2}$$

where $\mathcal{R}_{ci}$ indicates the sentence representation of the i-th CHQ extracted from E_q, and $\mathcal{R}_{fi}$ and $\mathcal{R}_{fj}$ are extracted from the key encoder E_k for the i-th and j-th FAQ respectively. The operation sim is to calculate the cosine similarity, τ is a temperature hyper-parameter. M_s is the memory bank which contains one positive sample and K simple negative samples in the queue with respect to an anchor.

3.3 Focus-Driven Hard Negative Samples

The above simple negative samples are randomly selected. As claimed by (Kalantidis et al., 2020), hard negative samples that are more similar to positive samples can facilitate the model to get better performance. Inspired by this, we build a bridge between hard sample generation and question focus prediction.

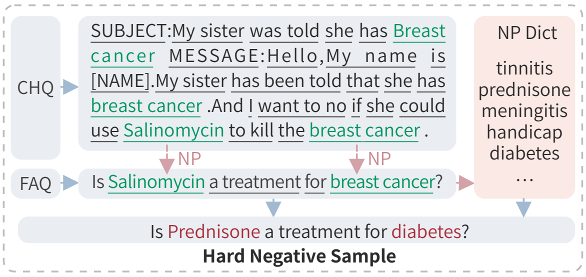

Figure 3: The method of hard negative samples generation.

3.3.1 Question Focus Identification

As mentioned before, the question focus is essential to understand a consumer health question. If some focus phrases are missing in the generated summary, the semantic will drift far away from the original user's intention. So we construct difficult negative samples based on the question focus to enhance contrastive learning. Specially, we replace the focus phrases with some other phrases of the same attribution, and keep other words of the sentence unchanged. An example of hard negative sample generation is shown in Figure 3.

One issue for our method is how to automatically annotate question focus. Yadav et al. (2021a) manually labeled the question focus in MeqSum dataset. However, this is quite time-consuming and labor-intensive, driving us to find a method which can automatically mark the question focus in larger datasets, such as HealthcareMagic and iCliniq. We analyzed the manually labeled MeqSum dataset, and found that in 340 of the total 500 records (up to 68%), the question focuses are the overlap phrases between CHQ and FAQ. Accordingly, we hypothesize that the same phrases appearing both in the source question and the golden summary have a high probability to be key-phrases. This idea is also proved to be effective in (Li et al., 2020).

Since the question focus is usually a phrase rather than a single word, we need to split one sentence into phrases. We apply the chunker (Akbik et al., 2018) to the CHQ and FAQ text, and record the chunk label of each phrase. Then the consistent phrases appearing both in CHQ and FAQ are labeled as the question focuses.

3.3.2 Hard Negative Sample Generation

We constructed a NP dictionary by concatenating all "NP" phrases of the FAQ sentences in the training set. To generate hard negative samples for contrastive learning, the question focuses are randomly replaced by other phrases of the same chunk label from the NP dictionary. As shown in Figure 2, "*breast cancer*" is replaced by "*diabetes*" since they share the same label "*NP*". We repeat this process N_h times to construct N_h different hard negative samples for each CHQ-FAQ pair.

3.3.3 Contrstive Learning on Hard Negative Samples

The sentence representation of hard sample $\mathcal{R}_h$ is extracted from the key encoder E_k. We define the hard loss function of contrastive learning as:

$$\mathcal{L}_{ctrCH} = -log \frac{e^{sim(\mathcal{R}_{ci}, \mathcal{R}_{fi})/\tau}}{\sum_{\mathcal{R}_h \in M_h} e^{sim(\mathcal{R}_{ci}, \mathcal{R}_h)/\tau}} \quad (3)$$

where M_h denotes the memory bank containing one positive sample and N_h hard negative samples.

This loss function forces the model to not only shorten the distance between CHQ and FAQ, but also expand the gap between the anchor CHQ and hard negative samples. In this way, we achieve the goal of making the model pay more attention to the question focus, and obtain a focus-aware representation.

3.4 Contrastive Learning at Decoder

An imbalance existing in the above method is that contrast learning is only utilized at the encoder. We fine-tuned BART on iCliniq dataset, and found that the decoder lacks the ability to distinguish the representations between the generated summary and the positive samples/unrelated negative samples, as $s^+_{g\_faq}$, $s^-_{g\_sim}$, $s^-_{g\_hard}$ shown in Figure 4. Therefore, we try to improve the similarity between the generated summary and its reference FAQ, and at the same time enlarge the dis-similarity between the generated summary and other unrelated questions.

Specially, we regard the generated summary as an extra anchor, and denote the representation of the generated summary as g_i. Since the output summary should be semantically consistent with the corresponding FAQ, we consider the representation of the FAQ f_i in the same pair as the positive sample, and select the simple negative samples randomly from the queue and generate hard negative samples using the *hard negatives generator*. The object functions of contrast loss $\mathcal{L}_{ctrGS}$ and $\mathcal{L}_{ctrGH}$ at the decoder end are defined in a similar style as Equation 2 and 3, except that the anchor c_i is replaced by another anchor g_i.

3.5 Overall Objective Function

For predicting next tokens in the generated summary, we use the cross entropy loss $\mathcal{L}_{ce}$:

$$\mathcal{L}_{ce} = -\frac{1}{|T|} \sum_{t \in T} log(p(y_t | x, y_{1:t-1}, \theta)) \quad (4)$$

In our model, the overall loss function consists of five parts: the cross entropy loss $\mathcal{L}_{ce}$ and four different loss functions of contrastive learning: $\mathcal{L}_{ctrCS}$, $\mathcal{L}_{ctrCH}$ for the anchor at the encoder end, $\mathcal{L}_{ctrGS}$, $\mathcal{L}_{ctrGH}$ for the anchor at the decoder end. We define the contrastive learning loss with respect to these two anchors as:

$$\begin{aligned} \mathcal{L}_{ctrC} &= \alpha \mathcal{L}_{ctrCS} + \beta \mathcal{L}_{ctrCH} \\ \mathcal{L}_{ctrG} &= \alpha \mathcal{L}_{ctrGS} + \beta \mathcal{L}_{ctrGH} \end{aligned} \quad (5)$$

where α, β are hyper-parameters to control the balance between simple negatives and hard ones. The weights of contrastive learning loss at the encoder and decoder are considered as equal, and the overall loss is defined as:

$$\mathcal{L} = \mathcal{L}_{ce} + \frac{1}{2}\mathcal{L}_{ctrC} + \frac{1}{2}\mathcal{L}_{ctrG} \quad (6)$$

4 Experiments

4.1 Datasets

We conduct experiments on three English benchmark medical question summarization datasets, including Meqsum, HealthcareMagic and iCliniq. Meqsum is a high-quality dataset from NIH [1], constructed by Ben Abacha and Demner-Fushman (2019). Mrini et al. (2021a) extracted HealthCareMagic and iCliniq datasets from MedDialog (Zeng et al., 2020) , which are collected automatically from the online healthcare service platforms [2] [3]. We list some statistics of these datasets in table 2. Following previous works, we adopt ROUGE (Lin, 2004)[4] as the evaluation metric.

[1] www.nlm.nih.gov/medlineplus
[2] www.healthcaremagic.com
[3] www.icliniq.com
[4] https://pypi.org/project/py-rouge

Dataset	Train	Dev	Test
MeqSum	400	100	500
HealthCareMagic	181,122	22,641	22,642
iCliniq	24,851	3,105	3,106

Table 2: Statistics of three medical question summarization datasets.

4.2 Training Details

We utilize BART-large (Lewis et al., 2020) in huggingface[5] as our pre-trained model. The learning rate of BART baseline is set to 3e-5 as the same with Mrini et al. (2021b). For contrastive learning in QFCL, the learning rate is optimized to 1e-5. Betas of Adam optimizer is set to 0.9 and 0.999. Batch size is set to 16. The number of hard negative samples n_h is set to 64. For Moco, the queue size K is set to 4096, temperature τ is 0.07, and the momentum coefficient m is 0.999. In Equation 5, α and β are set to 1 and 0.5 respectively through grid search on MeqSum development set. Experiments were all performed on a single NVIDIA RTX 3090 GPU. The average runtimes of each epoch for MeqSum, iCliniq and HealthcareMagic are 4.2h, 0.6h and 0.1h respectively.

4.3 Overall Performance

We report our experimental results in Table 3. Our model achieves new state-of-the-art results on all three datasets. Compared with the previous best results, we obtain an improvement of 0.99 ROUGE-L score on MeqSum, 8.44 on iCliniq, and 0.51 on HealthcareMagic, respectively.

MTL+Data augmentation (Mrini et al., 2021b) obtains the previous state-of-the-art results on iCliniq and HealthcareMagic, which utilizes the question entailment data to augment summarization data. In contrast, our method doesn't need other classification models or external data. The work of ProphetNet+QTR+QFR (Yadav et al., 2021a) gets the previous best result on MeqSum, which presents a reinforcement learning-based framework with question-aware rewards. Comparing with this competitive model, our method obtains consistent better performance on all metrics, with 2.28 improvement on R1, 4.66 improvement on R2 and 0.89 improvement on RL.

[5]huggingface.co/facebook/bart-large

4.4 Ablation Study

We perform ablation study to evaluate the impacts of different components employed in QFCL, and report the results in Table 3. In particular, for Meqsum dataset, due to the small size which may cause the training unstable, we conducted five separate experiments and computed the average ROUGE score of these five checkpoints as the final result. Compared with the base BART model, we obtain an absolute improvement of 5.33 points on average. T-test is implemented on such five ROUGE scores and the p-value is less than 1e-2, validating that this improvement is significant. On Cliniq the absolute improvement is 12.85 points and on HealthcareMagic 3.81 points. In comparison to BART, the relative improvements of our model are 12.2%, 28.7% and 9.6% on Meqsum, Cliniq and HealthcareMagic respectively.

The results demonstrate that each component of our model is helpful. On MeqSum, there is an increase of 3.15 points for BART+S compared to the baseline, indicating that the contrastive learning on simple negative samples largely improves model performance. It shows an continuous increase of 0.77 points for BART+S+H, and the highest ROUGE-L score is obtained when three parts are all implemented in our model. It suggests that each component in QFCL contributes positively, and metrics like ROUGE evaluating the similarity between whole sentences benefit from our contrastive learning strategy.

4.5 Human Evaluation

To quantitatively assess the results, we compare our method with the baseline BART through human judgement. We randomly selected 50 samples from each of three datasets, and hired 3 graduate students to categorize each generated summary into one of the following categories: 'Incorrect', 'Acceptable', and 'Perfect'. We compute the average number of each category, and report the result in Table 4. The average Spearman correlation coefficient between three annotators is 0.68, which guarantees a high quality of our annotation data. The evaluation results show that our model generates a higher proportion of perfect samples and a lower proportion of incorrect ones, by enhancing the model's ability of capturing sentence semantics and question focuses.

Model	MeqSum			iCliniq			HealthCareMagic		
	R1	R2	RL	R1	R2	RL	R1	R2	RL
Pointer-Generator Networks (PG)(See et al., 2017)	32.41	19.37	36.53	-	-	-	-	-	-
PG+Data augmentation(Ben Abacha and Demner-Fushman, 2019)	44.16	27.64	42.78	-	-	-	-	-	-
ProphetNet + QTR + QFR(Yadav et al., 2021a)	45.52	27.54	48.19	-	-	-	-	-	-
MTL+Data augmentation(Mrini et al., 2021b)	49.20	29.50	44.80	54.20	36.90	49.10	45.90	24.30	42.90
BART (Lewis et al., 2020)	46.17	28.05	43.75	48.79	25.47	44.69	42.33	23.07	39.60
BART + S	49.30	31.78	46.89	56.58	36.43	52.06	44.35	24.73	41.46
BART + S + H	49.96	32.72	47.66	58.26	40.08	55.34	45.52	25.71	42.51
BART + S + H + D (QFCL)	**51.48**	**34.16**	**49.08**	**60.09**	**43.22**	**57.54**	**46.42**	**26.47**	**43.41**

Table 3: Experimental results on three medical question summarization datasets. S denotes the contrastive learning on simple negative samples at the encoder end; H denotes the contrastive learning on hard negative samples at the encoder end; D denotes the decoder end's contrastive learning. The top group lists the existing state-of-the-art results on three datasets, and the bottom group shows our ablation study on different components.

Model	MeqSum			iCliniq			HealthCareMagic		
	I	A	P	I	A	P	I	A	P
BART	28.7	17.3	4.0	12.3	17.0	20.7	20.7	20.3	9.0
QFCL	12.0	18.0	20.0	6.3	17.7	26.0	5.7	16.3	28.0

Table 4: Human evaluation of the summaries generated by BART and QFCL respectively. The metric **I** means the number of incorrect samples, **A** means acceptable, **P** means perfect.

	Case1
CHQ	MESSAGE: Is it okay to drink **alcohol** in moderation when taking **Ampicillin**. I was told it negates any medical effect of the drug
FAQ	Can I drink **alcohol** while taking **Amoxicillin**?
BART	What are the side effects of **Ampicillin**?
QFCL	Is it okay to drink **alcohol** with **Ampicillin**?
	Case2
CHQ	Hi..... I have 3 month old baby girl...... I don t have **breast milk** from the beginning due to some reason. I can not give formula milk to baby...... So right now i m giving buffelo milk .... What else i should give her for better nourishment????? .... She has constipation problem may be due to milk but i cant give her breastmilk or formula .... How to overcome it?????......... Please help me
FAQ	Suggest ways to feed newborn other than **breast milk**
BART	Suggest treatment for constipation in a child
QFCL	Suggest better nourishment for baby other than **breast milk**

Table 5: Examples of generated question summaries by BART and our QFCL model. The question focuses are highlighted.

4.6 Case Study

To clearly show the output question summary, we list two samples to compare our model with BART in Table 5. In Case 1, BART captures the question focus "*Ampicillin*" but misses "*drink alcohol*", and in Case 2 it misses the question focus "*breast milk*". In contrast, our model successfully extract multiple question focuses from the lengthy CHQ, and generate summaries which are more conform to the meaning of original questions.

5 Model Analysis

5.1 Correlation of Sentence Representations

Since the auxiliary structures are discarded at the inference stage, we make further analysis to check that whether the retained model has the ability to distinguish different sentence-level semantics when facing unknown data. We train QFCL and BART on the training set for 20 epochs and save each checkpoint, and evaluate these checkpoints on the development set.

Four types of sentence representations are extracted from these checkpoints: CHQ's representation $\mathcal{R}_c$, FAQ's representation $\mathcal{R}_f$, hard negatives' representation $\mathcal{R}_h$, and the generated summary's representation at decoder end $\mathcal{R}_g$. Then we calculate the cosine similarity between them, and draw the relationship between these similarity scores and the epoch numbers, as shown in Figure 4.

Regarding the anchor CHQ in the curve of iCliniq, $s^{+}_{c\_faq}$, $s^{-}_{c\_sim}$ and $s^{-}_{c\_hard}$ are very close to each other at epoch 0, suggesting that the initial encoder lacks the ability to capture different semantics. With the increase of training steps, $s^{+}_{c\_faq}$ changes smoothly, while $s^{-}_{c\_sim}$ decreases sharply to near zero and $s^{-}_{c\_hard}$ decreases gradually and converges at a middle level between $s^{+}_{c\_faq}$ and $s^{-}_{c\_sim}$. This suggests that, powered by contrastive learning, our model has learned to distinguish sentences of different meanings at the encoder end.

With the generated summary as another anchor, we find out that $s^{+}_{g\_faq}$, $s^{-}_{g\_sim}$, $s^{-}_{g\_hard}$ are all near to 0

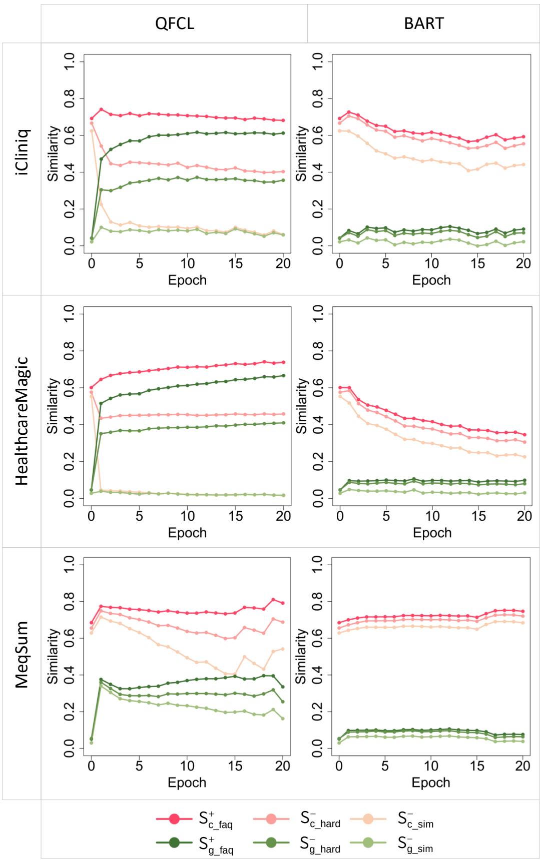

Figure 4: Correlation between sentence representation similarities and epoch numbers on dev set. The red lines are about the anchor CHQ. $s^+_{c\_faq}$ is the average cosine similarity between CHQ and related FAQ, $s^-_{c\_sim}$ is between CHQ and simple negative samples (other FAQs), $s^-_{c\_hard}$ is between CHQ and hard negative samples. The green lines are about the anchor of generated summary. $s^+_{g\_faq}$ is the average cosine similarity between the generated summary and FAQ, $s^-_{g\_sim}$ is between generated summary and simple negatives, $s^-_{g\_hard}$ is between generated summary and hard negatives. The epoch number equaling 0 denotes the initial pre-trained model.

initially, which depict that the decoder is also weak in representing sentence-level semantics. After training, $s^+_{g\_faq}$ increases significantly, $s^-_{g\_hard}$ converges between $s^+_{g\_faq}$ and $s^-_{g\_sim}$, and $s^-_{g\_sim}$ keeps very low all the time. It suggests that the decoder has strengthened its power to distinguish different semantics as the same to the encoder end.

Another chart is drawn to show this relationship for BART baseline in Figure 4. The similarities between the anchor and the positive samples, negative samples are very close, and never improve significantly with the progress of training. This situation suggesting that the BART baseline has a relatively weaker performance to distinguish the sentences of different meanings at both encoder and decode, since it only focuses on the prediction of next tokens.

Model	C1	C2	C3	C4	C5	Mean
BART	33.37	40.76	39.78	35.34	36.21	37.09
QFCL	47.41	42.24	45.20	45.20	47.17	**45.44**

Table 6: Accuracy of question focuses in generated summaries. C1-C5 means 5 different checkpoints trained by each model.

We also draw this correlation curve on MeqSum and HealthcareMagic. The curve of HealthcareMagic is similar to iCliniq. On MeqSum, our model can still distinguish sentences with different semantics better than the baseline, but the signal is not as significant as iCliniq or HealthcareMagic due to the limited size of training set.

5.2 Capturing Question Focus

To study whether our model pays more attention to the question focus, we evaluate the accuracy of question focuses in generated summaries. We use the sequence labeling model trained by Yadav et al. (2021a) to predict question focuses on the MeqSum dataset, and regard the 812 predicted question focuses in test set as the gold-standard. For QFCL and BART, we train five checkpoints and generate summaries on these checkpoints, and compute the accuracy of question focuses on test set. As shown in Table 6, the average accuracy is 37.09% for BART and 45.44% for QFCL. Our model exceeds the baseline by 8.35 points for question focus generation. P-value of t-test on these two sets of results is 1.04e-3, indicating that this improvement is statistically significant.

6 Conclusion

In this paper, we introduce a novel question focus-based contrastive learning framework QFCL for medical question summarization. In the proposed model, we adopt a "double anchor" strategy, by considering both the input question CHQ and the generated summary as comparing anchors. And we exploit a "hard negatives generator" to generate hard negative samples based on the question focus. Our model significantly improves the performance on three medical question summarization datasets, and achieves new state-of-the-art results. In the future, we would like to find a more effective way to do question focus recognition.

References

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online. Association for Computational Linguistics.

Evan Greensmith, Peter L Bartlett, and Jonathan Baxter. 2004. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5(9).

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738.

Yifan He, Mosha Chen, and Songfang Huang. 2021. damo_nlp at MEDIQA 2021: Knowledge-based preprocessing and coverage-oriented reranking for medical question summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 112–118, Online. Association for Computational Linguistics.

Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 21798–21809. Curran Associates, Inc.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8196–8203.

Siyao Li, Deren Lei, Pengda Qin, and William Yang Wang. 2019. Deep reinforcement learning with distributional semantic rewards for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6038–6044, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6706–6716.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021a. Joint summarization-entailment optimization for consumer health question understanding. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 58–65, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021b. A gradually soft multi-task and data-augmented approach to medical question understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1505–1515, Online. Association for Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilia Farcas, and Ndapa Nakashole. 2021c. UCSD-adobe at MEDIQA 2021: Transfer learning and answer sentence selection for medical summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 257–262, Online. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258, Online. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

Mario Sänger, Leon Weber, and Ulf Leser. 2021. WBI at MEDIQA 2021: Summarizing consumer health questions with generative transformers. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 86–95, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision – ECCV 2020*, pages 776–794, Cham. Springer International Publishing.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3733–3742.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021a. Reinforcement learning for abstractive question summarization with question-aware semantic rewards. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 249–255, Online. Association for Computational Linguistics.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021b. NLM at MEDIQA 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 291–301, Online. Association for Computational Linguistics.

Nan Yang, Furu Wei, Binxing Jiao, Daxing Jiang, and Linjun Yang. 2021. xMoCo: Cross momentum contrastive learning for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6120–6129, Online. Association for Computational Linguistics.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.