Towards Summarizing Healthcare Questions in Low-Resource Setting

Shweta Yadav and **Cornelia Caragea**
Department of Computer Science,
University of Illinois at Chicago, Illinois
{*shwetay, cornelia*}@uic.edu

Abstract

The current advancement in abstractive document summarization depends to a large extent on a considerable amount of human-annotated datasets. However, the creation of large-scale datasets is often not feasible in closed domains, such as medical and healthcare domains, where human annotation requires domain expertise. This paper presents a novel data selection strategy to generate diverse and semantic questions in a low-resource setting with the aim to summarize healthcare questions. Our method exploits the concept of guided semantic-overlap and diversity-based objective functions to optimally select the informative and diverse set of synthetic samples for data augmentation. Our extensive experiments on benchmark healthcare question summarization datasets demonstrate the effectiveness of our proposed data selection strategy by achieving new state-of-the-art results. Our human evaluation shows that our method generates diverse, fluent, and informative summarized questions.

1 Introduction

Online health information search is becoming conventional for more and more consumers every day. A recent survey showed that on average eight million people in the United States seek health-related information on the Internet every day[1]. One challenge towards assisting consumers in their healthcare information search is automatic question understanding. Generally consumers' questions are overly descriptive and include several peripheral information (as shown in Figure-1), which are not necessary to answer questions. Therefore, in this study we tackle the task of consumer health question understanding by summarizing the question.

Automatic text summarization is a non-trivial task in Natural Language Processing (NLP) that aims to generate human-readable, concise text containing salient information of the original document. The recent development in large-scale neural language models (Devlin et al., 2019; Raffel et al., 2020) have led to significant performance on several abstractive summarization task. However, their accuracy is partially due to the availability of large-scale human-annotated training data. Moreover, some domains such as biomedical and medical require domain experts to create high-quality training datasets, which is tedious to create at a large-scale level.

Consumer Health Question

Hello, my name is [NAME] , and I have Fibromyalgia terribly. My father has MS, and my mother's sister has been diagnosised with ALS. I was wondering if that would have any bearing on why my Fibromyalgia is hit me so hard. I have been experiencing s/s since as far back as 3rd grade that medical records can show and I'm 40 years old now. By the time I was 18 I was told I need psychiatric help because I was to young to complain like an 80 year old.

Summary

Is fibromyalgia hereditary and related to multiple sclerosis and amyotrophic lateral sclerosis?

Figure 1: The highlighted text shows important key aspects of the question which need to be considered while generating the summary.

A potential solution that has shown effectiveness in other generation and translation tasks is to augment the large-scale synthetically generated samples with a human-annotated training set. However, a limited study focused on data selection strategy in summarization, particularly for abstractive summarization. The majority of the traditional data selection methods are based on word replacement that mainly generates a synthetic sentence by changing one or multiple words with their synonyms (Zhang et al., 2015) or with a language model predicted words (Kobayashi, 2018). However, these methods make minor changes to the original sentence and therefore fall short of generating a diversified sentence.

To address this research gap, we present a novel data selection strategy for abstractive consumer

[1] https://pewrsr.ch/3l6m3mv

Proceedings of the 29th International Conference on Computational Linguistics, pages 2892–2905
October 12–17, 2022.

health question (CHQ) summarization task. Inspired by the success of the round-trip translation (RTT) (Hoang et al., 2018) – a process of translating the sentences to a pivot language and then back translating to the original language, we aim to explore the effect of RTT as a data augmentation method in CHQ summarization. However, not all the data samples obtained from RTT are diverse and can contain redundant information.

Towards this, we enhance the capability of RTT by devising multiple optimal data selection strategies to select diverse and informative questions, which leads to the significant performance improvement of the CHQ summarization system. Our first data selection strategy **Frechét Question Distance** (**FQD**) is based on Frechét distance (Dowson and Landau, 1982), which measures the distribution distance between the gold and round-trip translated question. The FQD ensures that questions having near similar or very different distributions should not be selected as additional data to train the summarization system. We propose **Precision Recall Question Distance** (**PRQD**) as our second data selection strategy, which disentangles the question distributions divergence into two components: *precision* and *recall*. These two components ensure that the selected additional data brings diversity to the whole training dataset. It is achieved by finding the trade-off between precision and recall of the distributions of the gold and round-trip translated questions. Our final data selection strategy **Question Semantic Volume** (**QSV**) is based on maximizing the semantic area formed by the points obtained from the semantic representation of the questions. The QSV aims to select the questions which maximize the semantic area leading to the selection of the additional questions which are non-redundant and diverse in nature.

We evaluated the effect of the additional data generated using the RTT and our proposed data selection objective measures on benchmark CHQ summarization dataset and two low-resource open domain datasets. We assess the role of each objective measure in RTT based data selection technique using five different pivot languages. Our results show that the RTT-based data selection method helps to improve the performance of the summarization system. We summarize the contribution of the work as follows:

1. We explored the role of the RTT-based data selection technique on CHQ summarization by experimenting with five different pivot languages.
2. We introduced the semantic-volume and diversity-based data selection objective measure in RTT to optimally select the diverse and informative synthetic questions.
3. Our unsupervised method achieves state-of-the-art performance on benchmark consumer healthcare question summarization datasets. Further, our human analysis confirms the effectiveness of our proposed approach in generating fluent and informative summary.

2 Related Work

Neural Abstractive Summarization: The recent advancement of neural networks models, particularly sequence-to-sequence (seq2seq) (Sutskever et al., 2014) models, attention mechanism (Bahdanau et al., 2015), copy mechanism (Gu et al., 2016), coverage mechanism (See et al., 2017) has propelled the development of efficient abstractive summarization approaches on numerous open-domain datasets. Several other methods have utilized the reinforcement learning (RL) (Paulus et al., 2018; Pasunuru and Bansal, 2018; Zhang and Bansal, 2019) to guide the models to generate faithful summaries. Recently, several studies have investigated the pre-trained language models in the abstractive summarization task (Qi et al., 2020; Liu and Lapata, 2019) and have achieved the state-of-the-art performance. Besides the supervised models, various other unsupervised approaches have utilized variational autoencoders for automatic summarization (Laban et al., 2020; Bražinskas et al., 2020; Baziotis et al., 2019).

Consumer Health Question (CHQ) Summarization: While major progress has been made in open-domain abstractive summarization, CHQ summarization is a relatively new task. Ben Abacha and Demner-Fushman (2019) defined the task of summarizing CHQ and introduced a benchmark dataset containing 1000 consumer questions summaries. Recently, a first shared task was organized by Ben Abacha et al. (2021) with the task of summarizing consumer health questions, radiology reports, and multi-document answers. The majority of the works (Lee et al., 2021; He et al., 2021; Sarrouti et al., 2021; Sänger et al., 2021) used pre-trained language models, ensemble approaches, and knowledge-based methods for the CHQ summarization task. A few other new methods (Yadav

et al., 2021a) have enhanced the capability of transformer model by inducing the latent knowledge. In the literature, several works have explored the concept of RTT in machine translation (Nguyen-Son et al., 2021), sentence construction (Zhou et al., 2021), and style-transfer (Zhang et al., 2020b).

Our works advances the existing studies in the consumer health question summarization by proposing an unsupervised framework to optimally select the diverse and information RTT questions, which leads to significant improvement without the need of additional labelled data.

3 Methods

3.1 Background

Given a consumer health question $Q = \{q_1, q_2, \ldots, q_M\}$, the goal of this task is to generate a summarized question $S = \{s_1, s_2, \ldots, s_N\}$ that contains the key information of the original question. Towards this, we build our question summarization model over the Transformer-based seq2seq (Vaswani et al., 2017) architecture. It aims to learn the conditional likelihood $p(S|Q) = \prod_{t=1}^{t=N} p(s_t|s_{<t}, Q)$, where, $s_{<t}$ denotes all generated target tokens before s_t. We utilized the pre-trained ProphetNet (Qi et al., 2020), as the strong base model to summarize the questions. We choose ProphetNet as it is specifically designed for sequence-to-sequence training and it has shown near state-of-the-art results on language generation and CHQ summarization task (Yadav et al., 2021a).

3.2 CHQ Summarization with Round-trip Translation

To train an effective neural network model for language generation tasks, the requirement of sufficient training data is indispensable. Synthetic data augmentation is a way to mitigate the data scarcity issue. It helps the model to reduce the brute-force memorization and also introduce a regularization effect.

In the literature, existing works (Yu et al., 2018; Xie et al., 2020) have shown that the RTT-based data augmentation methods create diverse samples while preserving the semantics. Inspired by these studies, we perform RTT to generate the paraphrases of the source CHQ that could lead to a better summarization system. In order to avoid the noise and keeping the fact intact, we did not paraphrase the gold summarized questions.

Specifically, for a given original dataset $\mathcal{D}_{orig} = \{(Q_i, S_i) \mid i = 1, 2, \ldots, L\}$, we translate the source CHQ $Q_i \in \mathcal{D}_{orig}$ into a non-English pivot language (xx) to obtain $\mathcal{D}^{en \rightarrow xx} = \{(Q_i^{xx}, S_i) \mid i = 1, 2, \ldots, L\}$ using the Google translation. We then back-translate the $\mathcal{D}^{en \rightarrow xx}$ to English and obtained $\mathcal{D}^{xx \rightarrow en}$. The final dataset is obtained from forward ($\mathcal{D}^{en \rightarrow xx}$) followed by the backward ($\mathcal{D}^{xx \rightarrow en}$) translation as:

$$\mathcal{D}_{rtt}^{en \leftrightarrow xx} = \{(\hat{Q}_i, S_i) \mid i = 1, 2, \ldots, L\} \quad (1)$$

Further, to enhance the model's generalization ability, we enrich the original training dataset $\mathcal{D}_{orig}$ with the additional RTT-based generated data $\mathcal{D}_{rtt}^{en \leftrightarrow xx}$. We call this as the augmented dataset $\mathcal{D}_{aug}^{en \leftrightarrow xx}$:

$$\mathcal{D}_{aug}^{en \leftrightarrow xx} = \mathcal{D}_{orig} \cup \mathcal{D}_{rtt}^{en \leftrightarrow xx} \quad (2)$$

3.3 Data Selection Objective Measures

We define three different data selection objective measures: **(i)** Frechét Question Distance, **(ii)** Precision Recall Question Distance, and **(iii)** Question Semantic Volume, that assess both diversity and quality of the RTT question by assigning low scores to less informative questions (*i.e.*, questions having factual errors and lacking salient medical information as present in the original question) or have low-diversity.

In the literature, there are few metrics like BLEU, Self-BLEU, Negative Log-Likelihood (NLL) that individually account for quality and diversity in the generated text. Alihosseini et al. (2019) shows that these metrics neglect either the quality (in the case of Self-BLEU) or the coverage (in metrics like BLEU, NLL). Thus, it is necessary to have a measure that could jointly consider both quality-diversity in the generated text. We argue that the distribution distance between the semantic representations of the round-trip generated question and the original question can be used simultaneously to select the diverse and informative round-trip generated question.

Given a gold question Q and round-trip generated question $\hat{Q}$, we first extract the question semantic representations h_Q and $h_{\hat{Q}}$ from a Transformer-based (Vaswani et al., 2017) pre-trained language model, which encodes the contextual information of the questions. Unlike the other work (Xiang et al., 2021), where the BERT has been used to derive fixed-size sentence embedding, we follow the idea of sentence-BERT (Reimers

and Gurevych, 2019) which uses the siamese and triplet networks (Schroff et al., 2015) to update the weights such that the generated semantically similar question representations are close in vector space. Towards this, we utilized the pre-trained `MPNet` (Song et al., 2020) model, which is fine-tuned using the siamese and triplet networks as discussed in Reimers and Gurevych (2019). We obtain the semantic representation of the questions from fine-tuned `MPNet` as:

$$\begin{aligned} h_Q &= \texttt{MPNet}(q_1, q_2, \ldots, q_M) \\ h_{\hat{Q}} &= \texttt{MPNet}(\hat{q}_1, \hat{q}_2, \ldots, \hat{q}_{\hat{M}}) \end{aligned} \tag{3}$$

In the following sub-sections, we use the semantic representation of the questions to devise multiple objective measures to select the diverse and informative round-trip generated question.

3.3.1 Frechét Question Distance

Heusel et al. (2017) introduced the metric Fréchet Inception Distance (FID) to evaluate the performance of the Generative Adversarial Networks (Goodfellow et al., 2014) based image generation models. FID is based on the Fréchet distance (Dowson and Landau, 1982) and is used to measure the similarity of generated images to real ones. Inspired by FID, we introduce FQD, which measures the distributional distance between the semantic representation of the gold question and the round-trip generated question. We assume that question semantic representations follow the multi-dimensional Gaussian distribution with first two moments: *mean* and *covariance*. The distance between these two Gaussian distributions is measured by the Fréchet distance.

Let the semantic representation h_Q of the gold question follow the Gaussian: $h_Q \sim \mathcal{N}(\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q)$ with mean $\boldsymbol{\mu}_q$ and co-variance matrix $\boldsymbol{\Sigma}_q$. Similarly, let the semantic representation of the round-trip question follow: $h_{\hat{Q}} \sim \mathcal{N}(\boldsymbol{\mu}_{\hat{q}}, \boldsymbol{\Sigma}_{\hat{q}})$. The Frechét Question Distance between Q and $\hat{Q}$ is computed as follows:

$$d_{\text{FQD}}(Q, \hat{Q}) = \|\boldsymbol{\mu}_q - \boldsymbol{\mu}_{\hat{q}}\|_2^2 + \text{Tr}(\boldsymbol{\Sigma}_q + \boldsymbol{\Sigma}_{\hat{q}} - 2(\boldsymbol{\Sigma}_q \boldsymbol{\Sigma}_{\hat{q}})^{1/2}) \tag{4}$$

where $Tr(X)$ is the trace of matrix X. To produce a uniform FQD score, we linearly scale the $d_{\text{FQD}}(Q, \hat{Q})$ in the range $[0, 1]$ using the following min-max normalization:

$$FQD(Q, \hat{Q}) = \frac{d_{\text{FQD}}(Q, \hat{Q}) - min(d_{\text{FQD}})}{max(d_{\text{FQD}}) - min(d_{\text{FQD}})} \tag{5}$$

where $min(d_{\text{FQD}})$ and $max(d_{\text{FQD}})$ represent the minimum and maximum FQD in the dataset. When the distribution of gold question is close to the distribution of the round-trip generated question, the FQD score is close to zero. In order to have the diverse, informative, and non-redundant samples in the training set, one does not need to include the round-trip generated questions whose FQD scores with gold questions are either low (near same question) or high (entirely different questions). Toward this, we aim to select the round-trip generated questions such that FQD score with gold questions is found to be in an optimal range. Given the round-trip generated questions $\mathcal{D}_{rtt}^{en \leftrightarrow xx}$ with pivot language (xx), we select a subset of the questions as follows:

$$\mathcal{D}_{\texttt{rtt+fqd}}^{en \leftrightarrow xx} = \{(\hat{Q}_i, S_i) \mid \mu_1 < FQD(Q_i, \hat{Q}_i) < \mu_2\} \tag{6}$$

where μ_1 and μ_2 are hyper-parameters (i.e., the optimal threshold) chosen based on the performance of CHQ summarization system on the validation dataset.

3.3.2 Precision Recall Question Distance

Inspired by the work of Sajjadi et al. (2018), which uses the notion of precision and recall to compare the reference and hypothesis distribution, we propose our second objective measure Precision Recall Question Distance. Similar to the FQD, it measures the distributional distance between semantic representations of the gold and round-trip generated questions; however, it does not require estimating the moments of the probability distributions. Intuitively *precision* measures how much of $h_{\hat{Q}}$ can be generated by a portion of h_Q. In contrast, *recall* measures how much of h_Q can be generated by a portion of $h_{\hat{Q}}$. Hence, the precision and recall should be high for the approximately same question distributions, whereas, if the question distributions are disjoint in nature, the precision and recall will be zero. Therefore, we aim to select the RTT questions whose precision and recall lies between the optimal range to ensure diversity. To compute PRQD, we follow the algorithm proposed by Sajjadi et al. (2018), which is based on the precision-recall distance (PRD) curve. Toward this, we compute pairs of precision $prec(\alpha)$ and recall $rec(\alpha)$ for an equiangular grid of values of α.

$$\begin{aligned} prec(\alpha) &= \sum_{\boldsymbol{v} \in \mathcal{V}} \min(\alpha h_Q(\boldsymbol{v}), h_{\hat{Q}}(\boldsymbol{v})) \\ rec(\alpha) &= \sum_{\boldsymbol{v} \in \mathcal{V}} \min\left(h_Q(\boldsymbol{v}), \frac{h_{\hat{Q}}(\boldsymbol{v})}{\alpha}\right) \end{aligned} \tag{7}$$

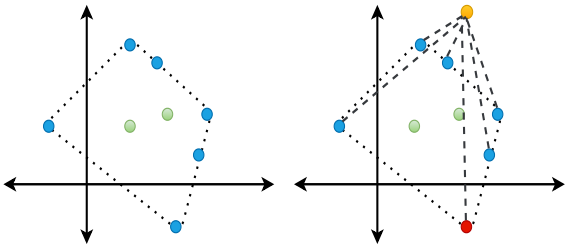

Figure 2: Question semantic volume maximization using convex hull. The ● and ● are the selected and non-selected candidates RTT questions using convex hull. The **left** side figure shows the toy-example of the convex hull. The **right** side figure shows the selected RTT question ● with respect to the gold question ●.

where h_Q and $h_{\hat{Q}}$ probability distributions are defined on a finite state space $\mathcal{V}$. In order to compute a single-value metric, we compute the F1-score corresponding to each α and select the maximum F1-score as the PRQD distance $d_{\text{PRQD}}(\hat{Q}, Q)$ as follows:

$$d_{\text{PRQD}}(\hat{Q}, Q) = \max \left\{ \frac{2 * prec(\alpha) * rec(\alpha)}{prec(\alpha) + rec(\alpha)} \;\middle|\; \alpha \in \Lambda \right\} \tag{8}$$

where $\Lambda = \{\tan(\frac{i}{p+1}\frac{\pi}{2}) | i = 1, \cdots, p\}$ and $p \in \mathbb{N}$ refers to the angular resolution, which is a hyper-parameter. Similar to FQD, we linearly scale the $d_{\text{PRQD}}(\hat{Q}, Q)$ in the range of $[0, 1]$ following Eq. 5 and obtained the normalized score $PRQD(\hat{Q}, Q)$.

Given the round-trip generated questions $\mathcal{D}_{rtt}^{en \leftrightarrow xx}$ with pivot language (xx), we select a subset of questions as follows:

$$\mathcal{D}_{\texttt{rtt+prqd}}^{en \leftrightarrow xx} = \{(\hat{Q}_i, S_i) \mid \beta_1 < PRQD(\hat{Q}_i, Q_i) < \beta_2\} \tag{9}$$

where β_1 and β_2 are the optimal thresholds which are chosen similar to μ_1 and μ_2.

3.3.3 Question Semantic Volume

Existing work in the literature (Yogatama et al., 2015) shows that the sentences which maximize the semantic volume in a distributed semantic space are the most diverse and have least redundant sentences. Motivated by this, first, we aim to find the most diverse and least redundant round-trip generated questions from the pool of RTT questions generated by considering different pivot languages. Later, we devise a simple yet effective measure to quantify the candidate RTT questions with respect to the gold questions in terms of their semantic distance. Specifically, for the given gold question Q and a set of K RTT generated questions $\{\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_K\}$, first, we extract (*cf.* Eq. 3) the semantic representation h_Q for gold question and each RTT questions $\{h_{\hat{Q}_1}, h_{\hat{Q}_2}, \ldots, h_{\hat{Q}_K}\}$ and form a data matrix $H \in \mathcal{R}^{(K+1)\times d}$. Later, we perform the linear dimensionality reduction using Principal Component Analysis to project the data matrix H to a lower dimensional space and obtain the transformed data matrix $\overline{H} \in \mathcal{R}^{(K+1)\times 2}$. In order to find and compare the most diverse round-trip candidate questions, we exclude the point corresponding to the gold question from $\overline{H}$. To find a convex maximum volume, we find the convex hull using the Quickhull algorithm (Barber et al., 1996) as follows:

$$\{p_1, p_2, \ldots, p_C\} = \texttt{ConvexHull}(\overline{h}_1, \overline{h}_2, \ldots, \overline{h}_K) \tag{10}$$

The convex hull are the smallest convex set that includes all points $\overline{h}_1, \overline{h}_2, \ldots, \overline{h}_K$. The points $\{p_1, p_2, \ldots, p_C\}$ are the vertices of the convex hull. It also guarantees to obtain the maximum semantic area with the selected points. Intuitively, it selects the RTT questions which are diverse in nature.

However, the vertices of the convex hull do not reduce the redundant points over the convex hull, and it lacks the notion of semantic distance from the point representing the gold question. Due to this, it usually selects the redundant round-trip generated questions (*cf.* Figure 2). To tackle this, first, we compute the euclidean distance $d(p_g, p_i)$ between the point p_g representing the gold question and each point p_i from the vertices of convex hull. Then, we only select the farthest apart round-trip question $\hat{Q}_j$ to include in the dataset if their semantic point in vector space represented by p_j is greater than an optimal threshold.

$$\begin{aligned} D &= \{d(p_g, p_i) \mid i = 1, 2, \ldots, C\} \\ p_j &= \underset{p_1, p_2, \ldots p_C}{\arg\max}(D) \end{aligned} \tag{11}$$

Finally, we select the optimal subset of the questions as follows:

$$\mathcal{D}_{\texttt{rtt+qsv}} = \{(\hat{Q}_j, S_j) \mid d(p_g, p_j) > \lambda\} \tag{12}$$

where λ is an optimal threshold and chosen based on the performance of CHQ summarization on the validation dataset.

4 Experiments

4.1 Datasets

We experimented with a benchmark CHQ summarization dataset (MEQSUM) (Ben Abacha and

Demner-Fushman, 2019). The MEQSUM dataset consists of domain-expert labeled 1000 question-summary pairs. The dataset is derived from de-identified consumer health questions (CHQs) received by the U.S. National Library of Medicine, National Institute of Health. Similar to the Ben Abacha and Demner-Fushman (2019), we augmented additional $4,655$ pairs of medical questions and shorter questions obtained from (Ely et al., 2000) to the original MEQSUM dataset. We use $5,055$ question-summary pairs as a training dataset, 100 sample pairs for validation, and 500 sample pairs for testing. We also experimented on an additional test collection containing 100 question-summary pairs released in BioNLP 2021 MEDIQA-QS shared task challenge (Ben Abacha et al., 2021) that has the same training set as the MEQSUM dataset.

We evaluated the performance of the proposed models using ROUGE (Lin, 2004). Following the existing works (Fabbri et al., 2021; Yadav et al., 2022b; Gliwa et al., 2019; Yadav et al., 2022a, 2021b), we reported the Rouge-1, Rouge-2, and Rouge-L. Additional implementation details are in the **Appendix**.

4.2 Experimental Setups

We design the following experiments to assess and compare the role of round-trip translation and the proposed data selection objective measures.

1. **Original Data**: We trained the question summarization system with the gold-standard training dataset ($\mathcal{D}_{orig}$) which consist of source question and target summary and evaluated the performance on the test dataset.
2. **RTT**: We augmented the RTT questions with the original data and obtained $\mathcal{D}_{aug}^{en \leftrightarrow xx}$ (*cf.* Eq. 2). We performed this experiment with five different languages (*xx*): Spanish (*es*), German (*de*), Japanese (*ja*), Chinese Simplified (*zh-CN*), and Chinese Traditional (*zh-TW*).
3. **RTT + FQD**: We utilize the FQD based objective measure to select the optimal subset of RTT synthetic questions. The selected synthetic questions ($\mathcal{D}_{\texttt{rtt+fqd}}^{en \leftrightarrow xx}$) with the original questions ($\mathcal{D}_{orig}$) are used to train the question summarization system.
4. **RTT + PRQD**: We use the PRQD based objective measure to select the optimal subset of round-trip translated synthetic questions. Similar to the **RTT + FQD**, we use selected synthetic questions ($\mathcal{D}_{\texttt{rtt+prqd}}^{en \leftrightarrow xx}$) along with the original questions ($\mathcal{D}_{orig}$) to train the question summarization system.
5. **RTT + QSV**: With this experimental setup, we select the optimal subset from round-trip translated synthetic questions based on question semantic volume obtained from the five different languages. We train the system with $\mathcal{D}_{\texttt{rtt+qsv}}$ dataset (*cf.* Eq. 12) along with the original questions ($\mathcal{D}_{orig}$).

4.3 Results

We report the results on MEQSUM datasets in Table 1. The results shows that our proposed method outperforms all the baselines in terms of Rouge-1, Rouge-2 and Rouge-L metrics on MEQSUM. Additionally, we also compared our proposed methods with the state-of-the-art techniques on MEQSUM. As evident from Table 1, our method outperforms the array of existing approaches on both datasets (in term of Rouge-L) without the need for any additional human-annotated training dataset. On MEQSUM, Mrini et al. (2021) obtained the best performance in terms of Rouge-1 and Rouge-2. It is to be noted that (Mrini et al., 2021) performed experiments on large-scale datasets from various healthcare forums which are restricted for data sharing and crawling. Therefore, to not breach the privacy concern of users, we did not considered those datasets for our experiments.

To understand the role of different data selection method, we carried out a deep analysis of the results (*cf.* Table- 2 and Table 7 in Appendix) both in terms of the performance (Rouge-1, Rouge-2, and Rouge-L) and the number of training samples selected. The results show that augmenting data via RTT significantly improves the performance of the model on all the three metrics. Especially with Frechét Question Distance, we achieve the highest Rouge-1, Rouge-2, and Rouge-L scores 46.59, 29.33, and 49.68 respectively. We also observe a similar gain on all the other language pairs with FQD. The FQD proved to be better amongst all the measures as it consider the semantic distance between the gold question and RTT generated question in the distributional space compared to the PRQD which computes a more abstractive distance. The PRQD based objective measure also achieve significant performance improvement over RTT in all five languages. Our final semantic-volume-based objective measure obtained the improvement of $2.35/3.01/2.61$ on Rouge-1/Rouge-2/Rouge-L

Methods	Rouge-1	Rouge-2	Rouge-L
Baseline Methods			
Seq2Seq (Sutskever et al., 2014)	25.28	14.39	24.64
Pointer Generator (PG) (See et al., 2017)	32.41	19.37	36.53
BertSumm (Liu and Lapata, 2019)	26.24	16.20	30.59
T5 (Raffel et al., 2020)	38.92	21.29	40.56
PEGASUS (Zhang et al., 2020a)	39.06	20.18	42.05
$\text{BART}_{\text{LARGE}}$ (Lewis et al., 2020)	42.30	24.83	43.74
ProphetNet (Qi et al., 2020)	43.87	25.99	46.52
State-of-the-art on CHQ Summarization			
PG + Data Augmentation (Ben Abacha and Demner-Fushman, 2019)	44.16	27.64	42.78
BART + Data-Augmented Joint Learning (Mrini et al., 2021)	48.50	29.70	44.90
ProphetNet + RL rewards (Yadav et al., 2021a)	45.52	27.54	48.19
Proposed Method (RTT+FQD)	46.59	29.33	49.68

Table 1: Comparison of our proposed method with the SOTA and other existing methods on the MEQSUM.

Method	Rouge-1	Rouge-2	Rouge-L	% of Additional Samples
Original Data	43.87	25.99	46.52	–
RTT	44.67	27.68	47.34	100
RTT+FQD	46.59	29.33	49.68	13.16
RTT+PRQD	45.48	27.74	48.61	75.31
RTT+QSV	46.22	29	49.13	2.00

Table 2: Performance of proposed methods (best on *es* language) on MEQSUM. The results for remaining languages can be found in **Appendix** (Table 7).

points over the original data based experiment.

Our second set of experiments analyzed the number of training samples selected by different objectives measures. It can be visualized from Table 2 that QSV outperforms the other selection measures by selecting only 2% of the RTT samples and obtaining 46.22, 29 and 49.13 values of Rouge-1, Rouge-2, and Rouge-L respectively. With FQD, we observed a little higher improvement on *de* language and reported 46.5, 29.53, and 49.45 values for Rouge-1, Rouge-2, and Rouge-L by selecting 6.85% of RTT samples.

We also evaluated the performance of our proposed objective measures on the MEDIQA-QS test dataset. Since, the official training data for MEDIQA-QS was MEQSUM annotated questions, we used the best-performing system (across each language) developed on MEQSUM to evaluate the performance (*cf.* Table-3) of each objective measures on MEDIQA-QS test set. The results shows that our proposed approach outperforms the existing methods in terms of Rouge-1 and Rouge-L. This confirms our data selection measures ensure the training samples are diverse in nature which leads to enhanced learning capability of the summarization model.

Methods	Rouge-1	Rouge-2	Rouge-L
Baseline Methods			
T5 (Raffel et al., 2020)	29.6	10.7	26.7
PEGASUS (Zhang et al., 2020a)	31.2	11.8	28.1
BART (Lewis et al., 2020)	28.6	9.8	25.8
ProphetNet (Qi et al., 2020)	30.3	11.1	26.5
Existing Methods			
Adversarial Training (Xu et al., 2021)	34.03	13.98	29.62
Transfer Learning (Lee et al., 2021)	33.52	15.97	30.90
Generative Transformers (Sänger et al., 2021)	33.40	15.99	31.49
Knowledge-based Method (He et al., 2021)	35.14	16.08	31.31
Proposed Methods			
RTT	35.40	15.00	30.80
RTT+FQD	36.80	15.30	32.10
RTT+PRQD	36.20	15.10	31.60
RTT+QSV	36.50	15.40	32.00

Table 3: Comparison of our proposed methods with the best performing models on the MEDIQA-QS test set.

4.4 Discussion

The results thus satisfy our two major claims: **(i)** The data generated using the RTT helps to improve the performance of the CHQ summarization model by a significant margin, and **(ii)** our proposed diversity and semantic-volume-based objective measures are highly effective in filtering out redundant and undesirable RTT questions, which makes the augmented data more informative and helpful in further improving the performance of the system. Amongst all the objective measure QSV select least amount of RTT samples, it is because QSV follow the two-steps (hull formation, maximizing the distance from gold summary) process to evaluate the informative and diverse samples. We analyze the 82.3% samples was excluded at the first step as they do not form the hull.

From the obtained results, FQD can be chosen among the proposed three objective measures. Although the use of FQD does not lead to selection of least training RTT samples, the results obtained by FQD are consistent and are very near to the optimal solution across all the languages. The complexity of FQD lies in estimating the mean and co-variance of the Gaussian. For the PRQD computation, we need to compute multiple precision and recall to form the PRD curve. The computation of precision and recall is computationally intense as the samples should be compared based on statistical regularities, which requires to obtain the histogram over the k-means clustering of the union of two semantic representations as discussed in Sajjadi et al. (2018). For the QSV, we need to obtained multiple (K) round-trip translated questions followed by their 2-d projection using PCA which requires $\mathcal{O}(2 * d^2)$.

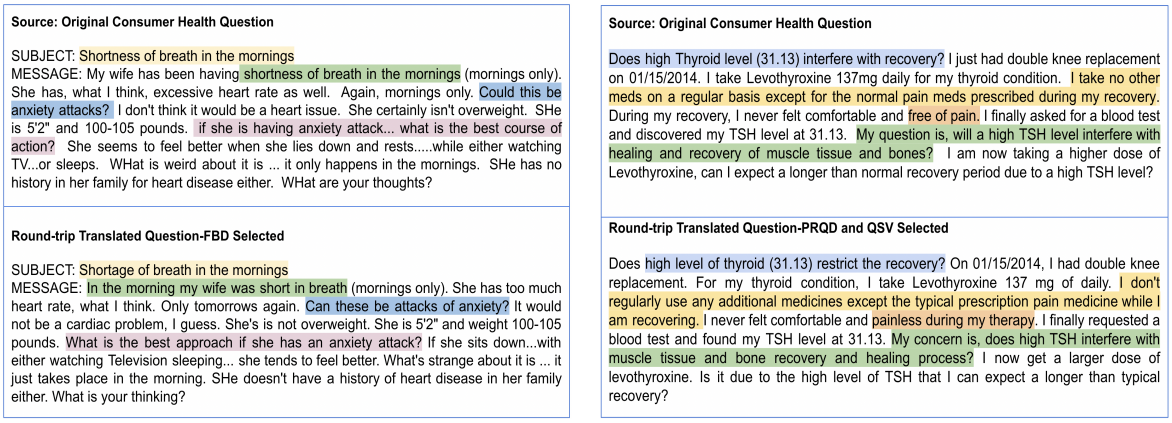

Figure 3: Selected RTT questions with the FQD, PRQD and QSV objective measures.

Thereafter, we need to obtain the convex hull of the projection which requires $\mathcal{O}(KlogK)$. Thus, the PRQD and QSV are computationally intense objectives compare to the FQD.

4.5 Evaluation on Healthcare Answer Retrieval Task

To determine whether the summarized questions can help in improving the answer retrieval performance, we performed experiments on the LiveQA 2017 test set (Abacha et al., 2017), consisting of 104 medical questions from the National Library of Medicine (NLM). The task aims to retrieve a correct answer to each medical question. Towards this, we used our best-performing method (FQD) on the CHQ summarization task to generate a summary for the LiveQA questions. We utilized the answer retrieval model developed in Yadav et al. (2022a) to retrieve the answer from the MedQuad collection[2]. We used the judgment scores[3] established by the LiveQA shared task to judge the quality of retrieved answers: "Correct and Complete Answer" (4), "Correct but Incomplete" (3), "Incorrect but Related" (2) and "Incorrect" (1). We excluded those questions for which the top answer's judgment score was unavailable. In this process, we evaluated common 48 questions for which human judgment scores were available across original questions, model-generated summarized questions and human-generated summarized questions.

Results We used the official evaluation metrics proposed by the LiveQA shared task to compare the performance of answer retrieval using the original versus summarized questions. Please note these metrics evaluate the first retrieved answer for each test question:

- avgScore(0-3): the average score for test questions by transferring 1-4 level grades to 0-3 scores. This is the main score to rank the LiveQA systems.
- succ@k: the number of questions with a score k or above ($k = \{2, 3, 4\}$) divided by the total number of questions in test set.
- prec@k: the number of questions with a score k or above ($k = \{2, 3, 4\}$) divided by the number of questions answered by the system.

Table 4 shows the results obtained by the QA system using: **(i)** the original questions, **(ii)** the summarized questions by FQD, and **(iii)** expert-created reference summaries as reported in (Ben Abacha and Demner-Fushman, 2019).

Measures	Original Questions	Human Generated Reference Summaries	FQD Generated Summarized Questions
avgScore(0-3)	0.384	0.557	0.48
succ@2+	0.23	0.336	0.288
succ@3+	0.115	0.144	0.144
succ@4+	0.038	0.076	0.048
prec@2+	0.5	0.72	0.62
prec@3+	0.25	0.312	0.312
prec@4+	0.083	0.016	0.104

Table 4: Evaluation of the answers retrieved using the original, human-generated, model-generated summaries based on the LiveQA metrics.

The results show that summarizing the CHQ can significantly improve the performance of the IR/QA system in retrieving relevant answers from the collection of curated answers. We also observe that the performance of the IR/QA model using the automatically summarized questions by our proposed approach is close to the performance achieved using the manually created reference summaries.

[2] https://github.com/abachaa/MedQuAD

[3] https://github.com/abachaa/MedQuAD(QA-TestSet)

Methods	RTT Questions			Generated Summary-MEQSUM				Generated Summary-MEDIQA-QS			
	Diversity (DI)	Informative (INF)	Factually Correct (FC)	Incorrect	Acceptable	Perfect	Fluent	Incorrect	Acceptable	Perfect	Fluent
ProphetNet	NA	NA	NA	23	18.5	8.5	23	25	17.5	7.5	26
FQD	38.5	41.5	41	9	12.5	28.5	37	11.5	10.5	28	35
PRQD	35.5	39.5	38.5	11.5	15.5	23	35	12	12	26	30
QSV	37	41.5	40.5	10	13	27	37	9.5	11	29.5	33

Table 5: Human evaluation on selected (50×5 languages) RTT questions. The metrics (DI, INF, FC) shows the average numbers of questions qualified for a given metric across all the 5 languages. The evaluation metrics (*Incorrect*, *Acceptable*, *Perfect*, *Fluent*) shows the average numbers of questions qualified for a given metric.

Original Question-I: SUBJECT: health MESSAGE: I have been bleeding since 2010 and I have been having sharp pain on my left stomach since 2014 and my stomach is so big and I feel weak I have don a lot of test and nothing was seen. What could be wrong with me? And how can I conquer?

Reference: What are the causes of abdominal pain and swelling?

Proposed Approach: What are the causes and treatment for abdominal pain?

Original Question-II: SUBJECT: EPI 743 MESSAGE: My son, His age 4 month discovered it leigh disease infected from the mother side. and we have full family history with the leigh disease. my Daughter she lived for 7 years with the same disease, we have her Hospital reports, it is confirmed leigh Disease. Kindly, if there is any hope for my son with EPI743 treatment, and we are appreciate to accept him in the treatment study. i have a full reports for my son and the MRI... Hope get your help ASAP because he is in the first stage. and we have a we are all hope he will be better.

Reference: Is EPI743 an effective treatment for leigh syndrome?

Proposed Approach: Can leigh disease be treated with EPI743?

Table 6: Generated summaries on the MEQSUM. Example-I shows an acceptable summary and model capability of generating novel words (*"abdominal pain"*) without being present in the original question. The second example shows a semantically correct summary.

4.6 Human Analysis

To understand the role of each data selection measures, we conducted human analysis on randomly selected 50×5 languages samples from RTT datasets. A set of 2 annotators experts in medical informatics evaluated the selected questions on the basis of *diversity*, *informativeness*, and *factual consistency* to measure **(1)** whether the RTT questions have novel n-grams, **(2)** whether the semantics of the original question was retained in the RTT questions and **(3)** whether the salient medical information were present in the selected RTT questions. We also instructed annotators to annotate the generated summaries into one of the following categories: '*Incorrect*', '*Acceptable*', and '*Perfect*' and also report the whether the summary was '*fluent*' or not. We reported the detailed quantitative analysis in Table 5. The results shows that FQD outperforms the other objective measures in terms of selecting more diverse and factually correct questions. Figure 3 shows the de-identified CHQ selected by the different objective measures. In our second analysis on the generated summary (*cf.* Table 5), we again observed the superiority of defined objective measures over the ProphetNet model (trained without the augmented data). This confirms the effectiveness of data selection objective measures that enhance the model learning ability by introducing diverse and informative questions (*cf.* Table 6), leading to the higher proportions of perfect summaries. We also conducted error analysis on generated summaries and identified two main source of errors: **(i)** the original questions consists of multiple sub-questions, and, **(ii)** if the question focus (medical entities) are not transformed into correct medical terminologies.

5 Conclusion

This work propose novel data selection strategy based on the concept of round-trip translation for consumer health question summarization. We devised three major data selection objective measures: FQD, PRQD and QSV based on the distributional distance to optimally select the diverse and informative samples from the pool of round-trip translated data. Extensive experiments show that proposed methods can effectively improve the performance without any additional labelled data. We also achieves new state-of-the-art results on benchmark consumer healthcare question summarization datasets. In future, we plan to explore these objective measures on other resource-scarce tasks.

References

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC*.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and*

Evaluating Neural Language Generation, pages 90–98.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

C Bradford Barber, David P Dobkin, and Hannu Huhdanpaa. 1996. The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. SEQˆ3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 673–681, Minneapolis, Minnesota. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the role of question summarization and information source restriction in consumer health question answering. *AMIA Summits on Translational Science Proceedings*, 2019:117.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2228–2234. 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.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

DC Dowson and BV Landau. 1982. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455.

John W Ely, Jerome A Osheroff, Paul N Gorman, Mark H Ebell, M Lee Chambliss, Eric A Pifer, and P Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *Bmj*, 321(7258):429–432.

Alexander Richard Fabbri, Faiaz Rahman, Imad Rizvi, Borui Wang, Haoran Li, Yashar Mehdad, and Dragomir Radev. 2021. Convosumm: Conversation summarization benchmark and improved abstractive summarization with argument mining. 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 6866–6880.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237*.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. volume 27.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

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.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 452–457, New Orleans, Louisiana. Association for Computational Linguistics.

Philippe Laban, Andrew Hsi, John Canny, and Marti A. Hearst. 2020. The summary loop: Learning to write abstractive summaries without examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5135–5150, Online. Association for Computational Linguistics.

Lung-Hao Lee, Po-Han Chen, Yu-Xiang Zeng, Po-Lei Lee, and Kuo-Kai Shyu. 2021. NCUEE-NLP at MEDIQA 2021: Health question summarization using PEGASUS transformers. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 268–272, Online. Association for Computational Linguistics.

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.

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.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapandula Nakashole. 2021. 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.

Hoang-Quoc Nguyen-Son, Tran Thao, Seira Hidano, Ishita Gupta, and Shinsaku Kiyomoto. 2021. Machine translated text detection through text similarity with round-trip translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5792–5797.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

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. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. ProphetNet: Predicting future n-gram for sequence-to-SequencePre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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.

Mourad Sarrouti, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2021. Nlm at bioasq synergy 2021: Deep learning-based methods for

biomedical semantic question answering about covid-19.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

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. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS 2020*. ACM.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27*, pages 3104–3112. Curran Associates, Inc.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing dialogue systems with distribution distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2192–2198, Online. Association for Computational Linguistics.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33.

Liwen Xu, Yan Zhang, Lei Hong, Yi Cai, and Szui Sung. 2021. Chichealth@ mediqa 2021: Exploring the limits of pre-trained seq2seq models for medical summarization. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 263–267.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022a. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128:104040.

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*, pages 249–255.

Shweta Yadav, Deepak Gupta, and Dina Demner-Fushman. 2022b. Chq-summ: A dataset for consumer healthcare question summarization. *arXiv preprint arXiv:2206.06581*.

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.

Dani Yogatama, Fei Liu, and Noah A. Smith. 2015. Extractive summarization by maximizing semantic volume. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1961–1966, Lisbon, Portugal. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *International Conference on Learning Representations*.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Shiyue Zhang and Mohit Bansal. 2019. Addressing semantic drift in question generation for semi-supervised question answering. 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 2495–2509, Hong Kong, China. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28:649–657.

Yi Zhang, Tao Ge, and Xu Sun. 2020b. Parallel data augmentation for formality style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3221–3228, Online. Association for Computational Linguistics.

Wangchunshu Zhou, Tao Ge, Canwen Xu, Ke Xu, and Furu Wei. 2021. Improving sequence-to-sequence pre-training via sequence span rewriting. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 571–582, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experiments

A.1 Implementation Details

The pre-trained large uncased version[4] of ProphetNet is used as the base encoder-decoder model. We use the fine-tuned verison[5] of MPNet from Huggingface (Wolf et al., 2020) to extract the semantic representation of the questios. We use the Google translation[6] to translate the question into pivot language and then back-translate them into English. To decode the summary, we use beam search algorithm with beam size 4. We fine-tuned the summarization models on the respective training dataset for 15 epochs. The length of maximum original questions and summarized questions are set to 120 and 20, respectively. We choose the optimal value of pairs of hyper-parameters (μ_1, μ_2), (β_1, β_2) and λ using the grid search. We found the optimal pairs (μ_1, μ_2)=$\{(0.17, 0.4),\ (0.25, 0.35),\ (0.05, 0.23),\ (0.19, 0.3),\ (0.04, 0.17)\}$, (β_1, β_2)=$\{(0.3, 0.85),\ (0.3, 0.6),\ (0.3, 0.6),\ (0.55, 0.85),\ (0.4, 0.85)\}$ on languages *es*, *de*, *ja*, *zh-CN* and *zh-TW*, respectively. We obtain the 0.8 as optimal value of λ. To compute PRQD, we follow the official implementation[7] with the hyper-parameter value $p = 1001$. We obtained the first two principal components using the scikit-learn[8] library (Pedregosa et al., 2011). The convex hull is computed using the Sci-py Qhull[9] library (Virtanen et al., 2020). To compute Rouge, we use the py-rouge implementation[10].

To update the model parameters, we used Adam (Kingma and Ba, 2015) optimization algorithm with the learning rate of $7e-5$ in all the experiments. We also used the cosine annealing (Loshchilov and Hutter, 2017) based learning rate decay scheduler, where the learning rate decreases linearly from the initial learning rate in the optimizer to 0.

We have checked for the software usage agreements. The licence details of the used software are as follows: ProphetNet and MPNet Huggingface (Apache-2.0 License), Google Translate (Apache-2.0 License), scikit-learn (BSD-3-Clause License), scipy (BSD-3-Clause License).

Computing Infrastructure: We performed all the experiments on a single NVIDIA Tesla V100 GPU having GPU memory of 32GB.

Average Run Time: The average runtime (for each epoch) to fine-tuned the ProphetNet model on original and RTT augmented datasets are recorded as 10.4 and 20.5 minutes respectively. For the FQD, PRQD and QSV objective based methods the average run time range between 11.5 and 17.2 minutes. It depends upon the number of samples selected for a particular pivot language.

Number of Parameters: The ProphetNet model has 391.32 million parameters. Since, we used the same model for all our experiments there fore we have the same 391.32 million parameters in all variants of the proposed methods.

A.2 Experimental Setups

A.3 Limitation

In this study, we evaluated the model generated summary using Rouge-1, Rouge-2 and Rouge-L metrics. However, these automatic evaluation metrics do not fully capture the nuances of what should or should not be included in a consumer question summary. Although we have performed human evaluation on a subset of summary, it has to be val-

[4] https://huggingface.co/microsoft/prophetnet-large-uncased

[5] https://huggingface.co/sentence-transformers/paraphrase-mpnet-base-v2

[6] We also performed the initial experiment with `mbart-large-50-many-to-many-mmt` (Tang et al., 2021) and found that Google's translation quality was much better than mBART.

[7] https://github.com/msmsajjadi/precision-recall-distributions

[8] https://bit.ly/3DPdjeR

[9] http://www.qhull.org/

[10] https://pypi.org/project/py-rouge/

Method		Rouge-1	Rouge-2	Rouge-L	% of Additional Samples
Original Data		43.87	25.99	46.52	–
es	RTT	44.67	27.68	47.34	100
	RTT+FQD	46.59	29.33	49.68	13.16
	RTT+PRQD	45.48	27.74	48.61	75.31
de	RTT	45.43	29	48.41	100
	RTT+FQD	46.5	29.53	49.45	6.85
	RTT+PRQD	46.38	29.47	49.4	71.27
ja	RTT	45.86	27.8	48.32	100
	RTT+FQD	46.17	29.39	49.48	59.06
	RTT+PRQD	46.02	28.2	49.26	81.34
zh-CN	RTT	44.81	27.71	47.75	100
	RTT+FQD	45.75	28.04	48.71	17.55
	RTT+PRQD	45.66	28.54	48.6	11.36
zh-TW	RTT	45.23	27.76	48.12	100
	RTT+FQD	46.13	28.45	49.16	51.46
	RTT+PRQD	45.88	27.69	48.66	67.02
RTT+QSV		46.22	29	49.13	2.00

Table 7: Performance comparison across all the languages on the proposed methods.

idated by clinical expert on a larger representative collection.

A.4 Potential Risk

The ProphetNet pre-trained language model used in this study are not checked for social bias and diversity. It may not be the representative of the whole world population and may contains region, community, race or gender specific biases.

A.5 Ethics / Impact Statement

Our project involves publicly available datasets of consumer health questions. It does not involve any direct interaction with any individuals or their personally identifiable data and does not meet the Federal definition for human subjects research, specifically: “a systematic investigation designed to contribute to generalizable knowledge" and “research involving interaction with the individual or obtains personally identifiable private information about an individual."