A Paraphrase Generation System for EHR Question Answering

Sarvesh Soni, Kirk Roberts
School of Biomedical Informatics
University of Texas Health Science Center at Houston
Houston TX, USA
{sarvesh.soni, kirk.roberts}@uth.tmc.edu

Abstract

This paper proposes a dataset and method for automatically generating paraphrases for clinical questions relating to patient-specific information in electronic health records (EHRs). Crowdsourcing is used to collect 10,578 unique questions across 946 semantically distinct paraphrase clusters. This corpus is then used with a deep learning-based question paraphrasing method utilizing variational autoencoder and LSTM encoder/decoder. The ultimate use of such a method is to improve the performance of automatic question answering methods for EHRs.

1 Introduction

The useful information present in electronic health records (EHRs) is hard to access due to many of its usability issues (Zhang and Walji, 2014). Question answering (QA) systems have the potential to reduce the time it takes for users to access information present in the EHRs. However, the effectiveness of such QA systems largely depends on the variety of questions they are capable of handling. Automated paraphrasing techniques are known to improve the performance of QA models in general domain by generating different variations of a question (Duboue and Chu-Carroll, 2006; Fader et al., 2013; Berant and Liang, 2014; Bordes et al., 2014a,b; Dong et al., 2015; Narayan et al., 2016; Chen et al., 2016; Dong et al., 2017; Abujabal et al., 2018b). Thus, automatic generation of high quality paraphrases for patient-specific EHR questions has the potential to improve performance of the clinical QA systems.

Paraphrasing is a technique of rewording a given phrase such that its lexical and syntactic structure is different but its semantic information is retained (Bhagat and Hovy, 2013). For instance, the following two questions can be considered as paraphrases of each other.

- What *medications* am I currently taking?
- What are my current *medications*?

Such EHR-related questions are usually targeted toward specific clinical information (Roberts and Demner-Fushman, 2016). For example, the aforementioned questions are intended to get information regarding *medications*. In such a scenario, paraphrases can be considered as different ways of accessing the same medical data. As such, automatic clinical paraphrase generation can help in increasing the breadth of questions for training a clinical QA system.

While automated paraphrase generation is well-studied in the general domain (Madnani and Dorr, 2010; Androutsopoulos and Malakasiotis, 2010), very few studies have focused on clinical paraphrasing (Hasan et al., 2016; Adduru et al., 2018; Neuraz et al., 2018). On the other hand, clinical text simplification, which aims at generating easier to read paraphrases, has received relatively more attention (Zeng-Treitler et al., 2007; Elhadad and Sutaria, 2007; Deléger and Zweigenbaum, 2008; Kandula et al., 2010; Pivovarov and Elhadad, 2015; Qenam et al., 2017; Adduru et al., 2018; Bercken et al., 2019). However, these works in the clinical domain are not representative of QA needs as the usefulness of paraphrases is largely application-specific (Bhagat and Hovy, 2013). Also, existing datasets for clinical paraphrasing consist of either short phrases (Hasan et al., 2016) or webpage title texts (Adduru et al., 2018), both of which are not suitable to build a paraphrase generator for QA. One can resort to using external tools such as Google Translate for generating question paraphrases (Neuraz et al., 2018), but these general-purpose tools are not tailored to the medical domain (Liu and Cai, 2015).

In this paper, we propose a clinical paraphrasing corpus CLINIQPARA with questions which

Proceedings of the BioNLP 2019 workshop, pages 20–29
Florence, Italy, August 1, 2019.

can be answered using EHR data[1]. We further propose a deep learning-based automated clinical paraphrasing system utilizing a variational autoencoder (VAE) and a long short-term memory recurrent neural network (LSTM) (Gupta et al., 2018). To our knowledge, this is the first work aimed at automatically generating paraphrases without using any external resource for questions specifically focused on retrieving patient-specific information from EHRs. Our main contributions are summarized as follows:

- Crowdsourcing a large paraphrasing corpus of questions which are answerable using the data from EHR.
- Application of VAE in context to clinical paraphrasing task.

The rest of the paper is structured as follows. Section 2 explores related work in the domain of clinical paraphrasing. Then, Sections 3 and 4 discuss our dataset generation and model implementation details respectively. Next, Section 5 evaluates the results of our clinical paraphrasing system. Finally, Section 6 discusses our findings, and Section 7 provides a concluding summary.

2 Background

We begin this section by detailing work related to clinical text simplification and paraphrasing in Sections 2.1 and 2.2 respectively. Then, we highlight some of the current work in general-domain paraphrasing for QA as part of Section 2.3.

2.1 Clinical Text Simplification

As stated earlier, many studies have focused on clinical text simplification. Text simplification differs from paraphrasing as the former is a unidirectional task whereas the latter can be considered as bi-directional textual entailment (Androutsopoulos and Malakasiotis, 2010), but the methods nonetheless provide useful context for our work. Elhadad and Sutaria (2007) and Deléger and Zweigenbaum (2008) relied on parallel or comparable corpora to construct paraphrase pairs of specialized and lay medical texts. Zeng-Treitler et al. (2007) and Kandula et al. (2010) either replaced the difficult clinical phrases in text with simpler synonyms or included uncomplicated explanations for them. Qenam et al. (2017) concentrated on just substituting the difficult terms with more comprehensible ones. Much of the simplification work in the clinical domain has been targeted toward lexical methods to convert or append the complex phrases present in the original sentence with their simpler alternatives (Pivovarov and Elhadad, 2015). Such simplification approaches usually make use of external vocabularies to map the difficult clinical terms. While these techniques reduce the complexity of a sentence at the lexical level, they generally leave the syntactic structure of a sentence unchanged. For instance,

- Patient suffered from *myocardial infarction*.
- Patient suffered from *heart attack*.

These variations correspond to a specific category of paraphrases named synonym substitution (Bhagat and Hovy, 2013) and amount to a smaller subset of possible paraphrases.

Alternatively, Adduru et al. (2018) and Bercken et al. (2019) constructed clinical simplification datasets from various web-based sources such as WebMD, MedicineNet, Wikipedia, and SimpleWikipedia utilizing sentence alignment techniques. While this approach is capable of generating more variations of a given sentence, it is still a simplification task and hence not suitable to be incorporated in a QA system (Bhagat and Hovy, 2013).

2.2 Clinical Text Paraphrasing

Comparatively, less focus has been drawn toward clinical paraphrase generation. Hasan et al. (2016) built their dataset by combining an existing general domain paraphrasing corpus PPDB 2.0 (Pavlick et al., 2015) with the UMLS (Unified Medical Language System) metathesaurus. Specifically, they utilized fully specified names of medical concepts present in UMLS. Though their corpus contains medical terms, it comprises of comparatively shorter length phrases rather than complete sentences.

Adduru et al. (2018) also created a paraphrasing corpus utilizing the titles of web articles from Mayo Clinic along with Wikipedia. While this dataset consists of complete clinical sentences, they are atypical of the patient-specific EHR questions.

Neuraz et al. (2018) used the Google Translate API to generate paraphrases for question templates in French. They utilized these generated template paraphrases to augment the size of their

[1]The corpus is available upon request.

Scenario 18: You've been having some low back pain recently, and want to make an appointment with your doctor's office through the doctor's website, but the system isn't clear. Write a short (up to 15 word), grammatical, one-sentence question asking how you make an appointment. No need to state it is confusing, simply ask a question. **Question**: How do I make an appointment?
Scenario 41: Your elderly mother has been taking Metformin (a diabetes drug). She is forgetful and requires someone to organize her pills for each day. However, the person that normally organizes her pills hasn't done it for this week, and you need to know what the instructions are for your mother's Metformin prescription. Write a short (up to 15 word), grammatical, one-sentence question asking her doctor for this dosage information. You question must contain the word 'Metformin'. **Question**: What are my mother's Metformin dosage instructions?
Scenario 43: You recently had an automobile accident, and you've started taking physical therapy to help recover. Your first appointment went well, but you forgot to write down when your next appointment was scheduled for. Write a short (up to 15 word), grammatical, one-sentence question asking your doctor for this information. Your question must contain 'physical therapy'. **Question**: When is my next physical therapy appointment?

Table 1: Three scenarios used to build the CLINIQPARA corpus, along with a canonical question (not provided to annotators).

development dataset for natural language understanding task without evaluating the quality of the paraphrases. Such general-purpose machine translation systems lack the ability to capture the domain-specific nuances of biomedicine (Liu and Cai, 2015). This suggests the need for a question paraphrasing dataset targeted toward clinical domain.

As discussed earlier, existing clinical paraphrasing datasets are not suitable for building a paraphrase generation system for clinical questions. To the best of our knowledge, the proposed paraphrasing corpus is the first which aims at clinical questions.

2.3 Paraphrasing for Question Answering

There are several question paraphrasing corpora available for the general domain such as WikiAnswers (Fader et al., 2013), PPDB (Ganitkevitch et al., 2013), PPDB 2.0 (Pavlick et al., 2015), GraphQuestions (Su et al., 2016), and ComQA (Abujabal et al., 2018a). However, there is a scarcity of such datasets for clinical questions.

The proposed corpus consists of questions which can be answered using EHR data. Such a corpus would have utility beyond QA systems as well, like in question similarity (Luo et al., 2015; Nakov et al., 2017), and in particular could serve as a standard paraphrase corpus for the medical domain.

3 Dataset Construction

In order to quickly and efficiently collect hundreds of paraphrases, we utilized the crowdsourcing platform Amazon Mechanical Turk (AMT). Instead of prompting AMT workers with a question and directly asking for paraphrases–which could prime the workers and bias them toward very similar paraphrases–we presented them with a short, 3-6 sentence imaginary scenario that placed them in a situation where a specific piece of information was required (such as their current medications). The workers were then asked to provide questions directed to their doctor to answer that information need. After the crowd-sourced questions were collected, they were manually organized into distinct paraphrase clusters. This was necessary because some questions address the information need but are not logically equivalent paraphrases. These steps are discussed in more detail below.

3.1 Scenario Creation

To ensure a wide variety of EHR questions, we first came up with 11 top-level topic categories people might ask about: medications, other treatments, labs, immunizations, imaging, other exams, problem list, past medical history, family history, appointments, and documents. For each of these categories, 2-8 scenarios were created to capture relevant questions about the topic. In total, 50 scenarios were created. Table 1 shows three of these scenarios along with the canonical question expected by the scenario.

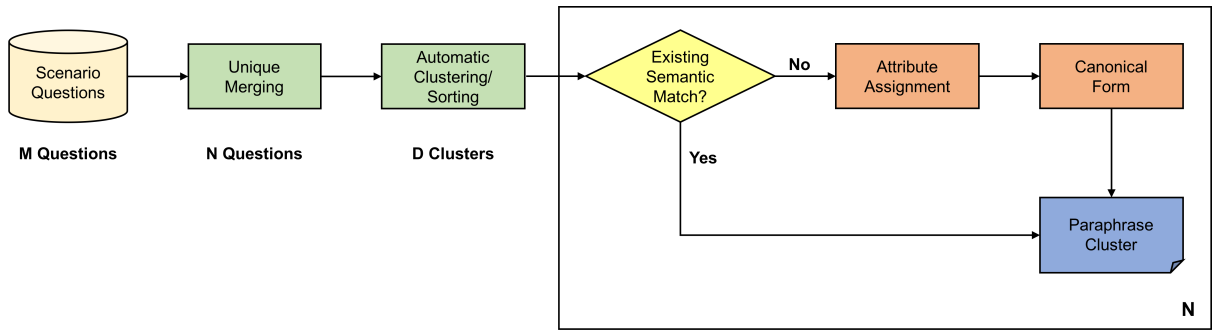

Figure 1: Paraphrase cluster creation process.

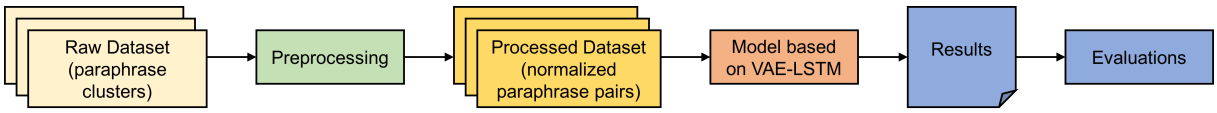

Figure 2: Framework of our paraphrasing system.

3.2 Crowdsourcing

The 50 scenarios were uploaded to AMT in three batches, one scenario per Human Intelligence Task (HIT). Workers were required to provide three questions per HIT, since first question might be obvious and not result in a particularly diverse set of paraphrases. Each HIT was assigned to 100 workers and the annotators were paid $0.08/HIT. Workers were required to be proficient in English, but otherwise no requirements were imposed and no demographic data was collected.

The initial validation process was minimal. HITs were rejected if the workers did not provide 3 questions, or if none of the questions were valid. 93% of submitted HITs were approved. Of the rejections, 73% were due to not providing 3 questions. Many of the rejections due to invalid questions were for questions that were completely unrelated to the scenario.

3.3 Paraphrase Cluster Creation

After collecting a set of questions for each scenario using crowdsourcing, the next step was to manually organize the questions into paraphrase clusters (Figure 1). We consider a paraphrase cluster to be composed only of exact paraphrases. That is, questions are paraphrases only if they should have the same semantic representation.

The first two steps in paraphrase construction were designed to ease the manual burden of paraphrase cluster assignment. First, questions were merged into case-independent unique sets. Second, questions were clustered using Dirichlet Process Mixture Model clustering with unigram and bigram features. This allowed us to sort the questions so that very similar questions, which are likely to be paraphrases, are annotated in succession. The remainder of this process required manual annotation for each question (with some computer assistance).

Each paraphrase cluster is represented by a canonical form. For each unique question, given the correct list of paraphrase clusters, the annotator selected a cluster that is the semantic match, or created a new cluster if none existed. Each new paraphrase cluster was assigned several values, notably including whether it was grammatical. Invalid questions (non-responsive, spurious responses that are common with crowdsourcing) were placed in either the INVALID-related cluster (invalid questions which were related to the scenario), or the INVALID-unrelated cluster. Finally, a canonical form was assigned to valid clusters.

The entire process in Figure 1 was repeated for each scenario. Since there were 100 workers per HIT, and 3 questions per worker, up to 300 questions needed to be clustered per scenario (with 50 scenarios, there were 15,000 questions). There were much fewer than 300 unique questions per scenario, and the process took between 30-40 minutes for most scenarios.

After ignoring casing and whitespace, there were an average of 240 unique questions per scenario. Three annotators manually clustered the questions (three scenarios were clustered as a group, with the remaining scenarios being clustering individually). Ignoring invalid questions (9%), and ungrammatical questions (6%), there were a

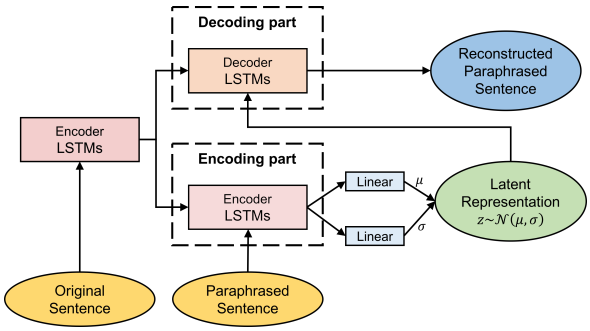

Figure 3: Architecture of the paraphrasing model based on VAE-LSTM.

median of 2.8 and mean of 5.6 paraphrase clusters, with a minimum of 5 questions, per scenario. Table 2 shows the paraphrase clusters for one of the scenarios.

4 Paraphrase Generator

An overall framework of our paraphrasing system is presented in Figure 2.

4.1 Preprocessing

First, we normalize the medical concepts and mask the person references and digits present in the question. This is carried out to make sure the questions from different scenarios are consistent. Consider the following questions and their masked versions:

- What types of *cancer* does my *father* have?
  → What types of *concept* does my *human* have?
- Was it in *2003* that I had my *appendectomy*?
  → Was it in *digits* that I had my *concept*?

After this step, we further deduplicate the questions and remove clusters with only 1 question (as a minimum of two questions are required for evaluating paraphrasing).

We then construct paraphrase pairs using the created clusters of paraphrases. Specifically, we generate all combinations of questions which are present in the same cluster. This results in over 258,000 unique question-paraphrase pairs for 10,578 questions distributed across 946 semantically distinct paraphrase clusters.

4.2 Model

We use a deep learning model based on VAE-LSTM (Gupta et al., 2018), the architecture of which is presented in Figure 3. One of the main characteristics of VAE that makes it a good choice for paraphrasing task is that its latent representation is continuous. In other words, the encoder outputs a distribution rather than discrete values. This enables the decoder to produce naturalistic outputs even in the cases where latent code does not correspond to any of the already viewed inputs.

The model consists of two parts, namely, encoding and decoding. On the encoding side, the original sentence is first passed to an encoder LSTM which constructs a vector representation x for the sentence. Then, another encoder LSTM takes as input x along with the paraphrased sentence whose vector representation y is generated as the output. Finally, a feedforward neural network generates the VAE encoder's mean (μ) and standard deviation (σ) parameters using y.

Both original and paraphrased sentences are fed into their respective encoder LSTMs using word embeddings. We train the word embeddings on our paraphrasing corpus using word2vec (Mikolov et al., 2013) and keep them fixed while training the paraphrasing system.

In the decoding phase, we first generate a vector representation x by passing the original sentence to an encoder LSTM. Ultimately, a decoder LSTM reconstructs the paraphrased sentence using x and a latent code z which is sampled from $\mathcal{N}(\mu, \sigma)$. While x is fed to the decoder LSTM only at an initial stage, z is taken as input at each of its stages.

During training, we aim to maximize the objective function shown below in Equation 1, thereby learning the VAE parameters.

$$\begin{aligned}\mathcal{O}(\theta, \phi; x, y) &= \mathbb{E}_{q_\phi(z|x,y)}[\log(p_\theta(y|z,x))] \\ &\quad - \mathrm{KL}(q_\phi(z|x,y) \,||\, p(z)) \quad (1)\end{aligned}$$

where $q_\phi(z|x,y)$ is a posterior distribution (encoder model) on z that the VAE aims at keeping closer to its prior $p(z)$ (commonly a standard normal distribution). KL represents the Kullback-Leibler divergence which intuitively gives a similarity measure between the two distributions. At the decoder side, $p_\theta(y|z,x)$ is a distribution on y, given the latent code z and vector x, whose expectation $\mathbb{E}$ is taken with respect to $q_\phi(z|x,y)$. The objective function gives a lower bound on the true likelihood of the data. We follow the training mechanism proposed by Bowman et al. (2016).

During testing, the encoder part is ignored and paraphrases are generated for a given question using z sampled from a standard normal distribution.

Scenario: You just realized you should have a doctor's appointment coming up soon, but cannot find it on your calendar. Write a short (up to 15 word), grammatical, one-sentence question asking your doctor about your next appointment.
Cluster 1 (229 questions, 164 unique): When is my next appointment? When is my next appointment? (frequency = 32) What time is my next appointment? (6) When is my next scheduled appointment? (5) Can you tell me when my next appointment is? (4) When is my next appointment scheduled? (4) When is my next appointment scheduled for? (4) What is the date and time of my next appointment? (3) *(... 157 more ...)*
Cluster 2 (38 questions, 33 unique): Do I have an appointment soon? Do I have an appointment coming up? (3) Do I have a doctor's appointment coming up soon? (2) Do I have an appointment soon? (2) Do I have an upcoming appointment scheduled? (2) *(... 29 more ...)*
Cluster 3 (3 questions): Do I have an appointment this week? Am I scheduled to come in to your office this week for an appointment? Do I have an appointment this week? Is my appointment scheduled for this week?
Cluster 4 (2 questions): Can I make an appointment? Can I make an appointment? Will you be able to make an appointment any soon?
Cluster 5 (1 question): How long until my next appointment? How long until my next doctor's appointment?
Cluster 6 (1 question): Is my appointment this week or next? Is my appointment scheduled for this week or next week?
Cluster 7 (1 question): Is my appointment next week? Was my appointment scheduled for next week?
Cluster 8 (1 question): Is my appointment on Tuesday? Is my scheduled appointment for Tuesday?
Cluster 9 (1 question): Is my appointment this month? Do you have a record of my having made an appointment for later this month?
Cluster INVALID-related (34 questions) Can you give me an appointment card? How long will this appointment last? What happens if I miss the appointment? What will you be discussing in regards to my next check up? Will I be meeting with you or with your assistant? *(... 29 more ...)*
Cluster INVALID-unrelated (17 questions) According to my lab results, what vitamins or supplements should I be taking? Do you have the results of my mri? How is my BMI? What does this medicine do? What symptoms should I watch for? *(... 12 more ...)*

Table 2: Paraphrase Clusters for Scenario 3. Only a sample of questions are shown.

The presence of input question at the decoder side enables the model to generate its paraphrases.

We utilize the same model parameters as Gupta et al. (2018). Namely, the dimension of the word embedding is 300; the dimension of the encoder and decoder is 600; the latent space dimension is 1100; the encoder has 1 layer; the decoder has 2 layers; the learning rate is 5 x 10^{-5}; the dropout rate is 30%; the batch size is 32. We use PyTorch for implementing the model and run all our experiments on an NVidia Tesla V100 GPU (32G).

4.3 Evaluation

The paraphrased questions generated by the model are re-incorporated with the concept, person names, and digits which were extracted during the preprocessing step. The paraphrases are evaluated using standard paraphrase evaluation metrics such as BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and TER (Snover et al., 2006), which are shown to work well for the paraphrase identification task (Madnani et al., 2012). BLEU score assesses the lexical similarity of generated paraphrases with the reference ones using exact matching while METEOR additionally takes into account the word stems and synonyms. TER score measures the edit distance (number of edits required to convert one sentence into another) between generated and reference paraphrases. So, higher BLEU and METEOR scores are better whereas a lower TER score is preferable. Since we have multiple paraphrases for each question in our corpus, we calculate these metrics for the generated paraphrases against all the available ground truth paraphrases.

To evaluate the performance measures on all the parts of CLINIQPARA dataset, we perform 10-fold cross validation. Specifically, we split our dataset by scenarios (into 10 groups each containing 5 scenarios) and sequentially test the performance of model on each group of 5 scenarios after training it on the other 45. We report the individual and average scores from all these runs in our results.

We further evaluate the performance of our model on the Quora dataset[2], which contains over 400k pairs of questions of which around 150k pairs are paraphrases. We train on 90% of these paraphrase pairs and test on the remaining 10%.

We also perform human evaluation of the generated paraphrases for quantifying the aspects not covered solely by the automated evaluation metrics. For the CLINIQPARA dataset, we randomly select a set of 300 questions from all the scenarios. For each of these questions, we further choose a ground truth paraphrase as well as a system generated paraphrase in a random fashion. This result in a total of 600 pairs of question paraphrases, 300 from the gold dataset and 300 generated by the paraphrasing system. The constructed set is separately evaluated by two annotators who are asked to rate the paraphrases based on two parameters: *fluency* of the questions as natural language and their *relevance* to the original question. Both of these scores range from 1 (worse) to 5 (best). For each paraphrase, the final score is calculated by averaging the scores provided by the two annotators. The fact that a paraphrase is ground truth or generated by the model is hidden from the annotators to avoid bias. For the Quora dataset, we directly report the human evaluation results from Gupta et al. (2018).

Dataset	Metric		
	BLEU	**METEOR**	**TER**
Quora	16.70	20.60	77.4
CLINIQPARA	13.25	21.47	91.93

Table 3: Performance of our paraphrasing system using automated evaluation metrics.

Fold	Metric		
(Scenarios)	**BLEU**	**METEOR**	**TER**
1-5	**19.25**	23.56	92.58
6-10	12.27	19.25	94.01
11-15	18.79	21.93	**78.17**
16-20	9.72	19.30	91.46
21-25	9.20	20.97	103.25
26-30	16.45	**23.66**	84.98
31-35	6.07	19.84	111.62
36-40	11.24	20.40	95.05
41-45	14.08	22.33	85.18
46-50	15.48	23.44	82.97
Average	**13.25**	**21.47**	**91.93**

Table 4: Results on CLINIQPARA using automated evaluation metrics for 10-fold cross validation. Each fold contains 5 scenarios over which the model is tested after being trained on the other 45 scenarios.

[2]https://data.quora.com/First-Quora-Dataset-Release-Question-Pairs

Dataset	Type	Relevance	Fluency
Quora	Ground Truth	4.82	4.94
	VAE-LSTM	3.57	4.08
CLINIQPARA	Ground Truth	4.69	4.70
	VAE-LSTM	1.88	3.65

Table 5: Results of human evaluation. Range of scores is between 1 (worst) and 5 (best).

Input Question
Do you know when my next appointment is going to be?
Generated Paraphrases
1. Can you please confirm the date and time of my appointment? 2. On what day and what time do I have my appointment? 3. Do you have the date and time for my appointment? 4. Can you tell me when I am scheduled for my appointment.

Table 6: Example paraphrases generated by the model for an input question from Scenario 3 (Good).

Input Question
Is my latest CAT scan impression complete?
Generated Paraphrases
1. Was my CAT scan impression successful or not? 2. Was my CAT scan impression a success? 3. Was my diagnosis CAT scan impression? 4. does my father 's file show how many times he has CAT scan impression?

Table 7: Example paraphrases generated by the model for an input question from Scenario 32 (Moderate).

5 Results

The results on CLINIQPARA (our dataset) and Quora dataset using automated evaluation metrics are shown in Table 3. More granular cross validation results on CLINIQPARA are presented in Table 4. Moreover, the results of our human evaluation process are shown in Table 5. Some of the system-generated paraphrases are included in Tables 6 and 7. Table 6 shows the examples from a fold which performed well during the cross validation step whereas Table 7 includes examples from a low-performing fold.

6 Discussion

The quality of generated paraphrases is promising, but further investigation is required to determine if performance is sufficient for use in training a downstream QA system. We note that the METEOR score on CLINIQPARA was comparable to that of the results on the Quora dataset. This shows the potential of our paraphrasing system in generating paraphrases similar to the ground truth paraphrases. Our system performed well on the Quora dataset in terms of BLEU score, which can be attributed to the larger size of the Quora dataset in terms of unique questions (150k in Quora vs. 10.5k in CLINIQPARA).

On analyzing the results of the qualitative evaluation, we observe that the majority of the errors are related to change in the person reference or asking about frequency-related information. For instance, the original question *"When shall I come for my next physical therapy?"* asking about the user's next appointment for a therapy is modified to a question *"May I have the number of times my father has physical therapy?"* asking about the number of times the user's father has undergone the therapy. A similar trend can be seen in the second example where the original question *"Can you please give me the dosage details on the metformin mom takes?"* is concerned about getting the dosage information for the user's mother whereas the system generated question *"Could you tell me the amount of time my father has metformin?"* is related to the frequency of metformin intake of the user's father. Further qualitative evaluation can help pointing out more specific problems with the model.

Our future work includes experimenting with more advanced embedding techniques (Peters et al., 2018; Devlin et al., 2018). We also plan to handle some of the aforementioned errors by incorporating additional constraints such as restricting the question paraphrase pairs in our corpus to contain only semantically similar masked references.

7 Conclusion

Automatic paraphrase generation of clinical questions can improve the performance of the QA systems. Little work has been focused on clinical paraphrasing, let alone concentrating on clinical questions. We have proposed a new clinical paraphrasing corpus CLINIQPARA, containing questions which can be answered using EHRs. Our model based on VAE-LSTM has the potential to generate quality clinical paraphrases.

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

References

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2018a. ComQA: A Community-sourced Dataset for Complex Factoid Question Answering with Paraphrase Clusters. *arXiv preprint arXiv:1809.09528*. Version 2.

Abdalghani Abujabal, Rishiraj Saha Roy, Mohamed Yahya, and Gerhard Weikum. 2018b. Never-Ending Learning for Open-Domain Question Answering over Knowledge Bases. In *Proceedings of the 2018 World Wide Web Conference*, pages 1053–1062.

Viraj Adduru, Sadid A. Hasan, Joey Liu, Yuan Ling, Vivek Datla, Kathy Lee, Ashequl Qadir, and Oladimeji Farri. 2018. Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In *CEUR Workshop Proceedings*, pages 45–52.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Jonathan Berant and Percy Liang. 2014. Semantic Parsing via Paraphrasing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1415–1425.

Laurens van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating Neural Text Simplification in the Medical Domain. In *The World Wide Web Conference*, pages 3286–3292.

Rahul Bhagat and Eduard Hovy. 2013. What Is a Paraphrase? *Computational Linguistics*, 39(3):463–472.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014a. Question Answering with Subgraph Embeddings. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 615–620.

Antoine Bordes, Jason Weston, and Nicolas Usunier. 2014b. Open Question Answering with Weakly Supervised Embedding Models. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 165–180.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21.

Bo Chen, Le Sun, Xianpei Han, and Bo An. 2016. Sentence Rewriting for Semantic Parsing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 766–777.

Louise Deléger and Pierre Zweigenbaum. 2008. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. In *AMIA Annual Symposium Proceedings*, page 146.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. Volume 2.

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. 2017. Learning to Paraphrase for Question Answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 875–886.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 260–269.

Pablo Duboue and Jennifer Chu-Carroll. 2006. Answering the question you wish they had asked: The impact of paraphrasing for question answering. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL*.

Noemie Elhadad and Komal Sutaria. 2007. Mining a Lexicon of Technical Terms and Lay Equivalents. In *Proceedings of BioNLP*, pages 49–56.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2013. Paraphrase-Driven Learning for Open Question Answering. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1608–1618.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase Database. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 758–764.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2018. A Deep Generative Framework for Paraphrase Generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5149–5156.

Sadid A Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. 2016. Neural Clinical Paraphrase Generation with Attention. In *Proceedings of the Clinical NLP Workshop*, pages 42–53.

Sasikiran Kandula, Dorothy Curtis, and Qing Zeng-Treitler. 2010. A semantic and syntactic text simplification tool for health content. In *Proceedings of the AMIA Annual Symposium*, pages 366–370.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An automatic metric for MT evaluation with high levels

of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231.

Weisong Liu and Shu Cai. 2015. Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study. In *Proceedings of BioNLP*, pages 134–140.

Jake Luo, Guo-Qiang Zhang, Susan Wentz, Licong Cui, and Rong Xu. 2015. SimQ: Real-Time Retrieval of Similar Consumer Health Questions. *J Med Internet Res*, 17(2):e43.

Nitin Madnani and Bonnie J. Dorr. 2010. Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods. *Computational Linguistics*, 36(3):341–387.

Nitin Madnani, Joel Tetreault, and Martin Chodorow. 2012. Re-examining Machine Translation Metrics for Paraphrase Identification. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 182–190.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26*, pages 3111–3119.

Preslav Nakov, Doris Hoogeveen, Lluís Màrquez, Alessandro Moschitti, Hamdy Mubarak, Timothy Baldwin, and Karin Verspoor. 2017. SemEval-2017 Task 3: Community Question Answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 27–48.

Shashi Narayan, Siva Reddy, and Shay B Cohen. 2016. Paraphrase Generation from Latent-Variable PCFGs for Semantic Parsing. In *Proceedings of the 9th International Natural Language Generation Conference*, pages 153–162.

Antoine Neuraz, Leonardo Campillos Llanos, Anita Burgun, and Sophie Rosset. 2018. Natural language understanding for task oriented dialog in the biomedical domain in a low resources context. *arXiv preprint arXiv:1811.09417*. Version 2.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 425–430.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237.

Rimma Pivovarov and Nomie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. 2017. Text Simplification Using Consumer Health Vocabulary to Generate Patient-Centered Radiology Reporting: Translation and Evaluation. *J Med Internet Res*, 19(12):e417.

Kirk Roberts and Dina Demner-Fushman. 2016. Annotating logical forms for EHR questions. In *Proceedings of the Language Resources & Evaluation Conference*, pages 3772–3778.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of 7th Conference of the Association for Machine Translation in the Americas*, pages 223–231.

Yu Su, Huan Sun, Brian Sadler, Mudhakar Srivatsa, Izzeddin Gur, Zenghui Yan, and Xifeng Yan. 2016. On Generating Characteristic-rich Question Sets for QA Evaluation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 562–572.

Qing Zeng-Treitler, Sergey Goryachev, Hyeoneui Kim, Alla Keselman, and Douglas Rosendale. 2007. Making texts in electronic health records comprehensible to consumers: a prototype translator. In *Proceedings of the AMIA Annual Symposium*, pages 846–850.

Jiajie Zhang and Muhammad Walji. 2014. *Better EHR, Usability, Workflow and Cognitive Support in Electronic Health Records*. University of Texas School of Biomedical Informatics at Houston.