Paraphrasing to improve the performance of Electronic Health Records Question Answering

Sarvesh Soni, MS, Kirk Roberts, PhD
School of Biomedical Informatics
University of Texas Health Science Center at Houston
Houston TX, USA

Abstract

This paper describes a paraphrasing approach to improve the performance of question answering (QA) for electronic health records (EHRs). QA systems for structured EHR data usually rely on semantic parsing, which aims to generate machine-understandable logical forms from free-text questions. Training semantic parsers requires large datasets of question-logical form (QL) pairs, which are labor-intensive to create. Considering the scarcity of large QL datasets in the clinical domain, we propose a framework for expanding an existing dataset using paraphrasing. We experiment with different heuristics for multiple sample sizes and iterations to assess the effect of adding paraphrasing to the task of semantic parsing. We found that adding paraphrases to an existing dataset based on TERTHRESHOLD *scores results in an improved performance in the majority (*74%*) of the experimental runs. Hence, the proposed paraphrasing-based framework has the potential to improve the performance of QA systems using a limited set of existing QL annotations.*

Introduction

Electronic health records (EHRs) contain a wealth of useful patient information. However, navigating such information through EHR interfaces is often difficult due to many of their usability issues[1,2]. Question answering (QA) techniques aim to reduce this effort by providing the means to access the records using natural language queries. The performance of QA systems is dependent on the variety of questions they accommodate. Since a question can be asked in many different ways, increasing the range of a QA system for handling this variation is important. Consider the following questions asking about the time of administrating a *flu shot*.

- When did the patient last get a *flu shot*?
- When was her most recent *flu vaccine*?
- When was she last administered an *influenza vaccine*?

All the above questions seek the same information – the time when some person of interest was administered a given influenza vaccine. In other words, all these questions carry the same semantic information but have different lexical and syntactic structures. Such groups of sentences are called paraphrases[3] and the natural language processing (NLP) technique for automatically generating paraphrases for a given sentence is known as paraphrasing.

Several works in the general domain have exploited paraphrasing techniques to improve the performance of QA models[4–12]. The main idea behind using these techniques is to introduce more natural language variation to the QA systems. Use of such techniques can have even more impact in the scenarios where the availability of datasets for training QA models is limited.

The structured information present in EHRs can be effectively queried using semantic parsing[13]. This technique aims at mapping the natural language questions to their machine-comprehensible logical forms[14]. Further, these unambiguous logical forms can be converted into structured queries such as Fast Healthcare Interoperability Resources (FHIR)[15] for querying EHRs.

Semantic parsers are usually trained on annotated datasets containing questions and their logical forms (QLs)[16]. Annotating the QLs is a time-consuming process, and does not promise a wide coverage of question surface forms. Our previous work aimed at reducing the time for annotating such corpora by automating a time-consuming step of concept normalization[17]. However, the issue of limited question variety remains due to the restricted availability of such

datasets in the clinical domain (because of several privacy and EHR data complexity issues). This further signifies the need of paraphrasing techniques to build a diverse corpus of questions.

In this paper, we propose a framework for improving the performance of a semantic parser by automatically generating paraphrases for an existing question-logical form (QL) dataset. Specifically, we first generate multiple paraphrases for the questions present in our QL corpus[17] using an improved version of our previously published paraphrase generation system based on a variational autoencoder (VAE) and a long short-term memory recurrent neural network (LSTM)[18]. Then, we apply various heuristics on these generated paraphrases to construct a set of paraphrases for training a semantic parser. Finally, we experiment by drawing different samples from the dataset for various heuristic criteria to determine the impact of adding paraphrases to the original dataset on the performance of our semantic parsing system. To our knowledge, this is the first work aimed at improving the performance of clinical semantic parsing by supplementing an existing QL dataset with question paraphrases.

Related Work

We present a brief review of the existing work where paraphrasing techniques are leveraged for improving QA in the general domain. A subset of these papers focused on using paraphrases for semantic parsing.

Many works have investigated the usefulness of direct incorporation of good quality paraphrases in a QA system to improve the overall performance of question-answering. Duboue et al.[4] employed rule-based machine translation technique to generate question paraphrases and emphasized the selection of good paraphrases to be used in a QA system. They evaluated the QA system by selecting the best paraphrase which resulted in an improvement of 35% in mean reciprocal rank (MRR) compared to the original question. Fader et al.[5] proposed a novel approach to learn structures having lexical equivalence for relations, entities, and question templates as well as rank them with the aim to map questions to queries for answering a broad variety of open-domain questions. Bordes et al.[7,8] used paraphrases in learning the question embeddings by incorporating the task of predicting a pair of questions as paraphrases in training the QA system. Similarly, a study by Dong et al.[9] leveraged the usage of question paraphrases to learn representations of words and question patterns in a multi-task learning approach. The incorporation of the paraphrases showed improved performance in question understanding and answer ranking. A recent work by Dong et al.[12] demonstrated the effectiveness of co-training both paraphrase model and QA model for predicting a distribution over answers provided a question. The paraphrasing component incorporates a scoring mechanism that predicts a paraphrase quality based on their likelihood to get to the correct answers. This end-to-end framework proved to improve the performance of QA over Freebase and answer sentence selection.

Paraphrasing has also been utilized to train semantic parsers. Berant et al.[6] designed a method to exploit a paraphrasing system to select the best canonical utterance (which is closer to the input utterance) corresponding to a logical form among all the canonical utterances of the candidate logical forms. This approach demonstrated improvement in semantic parser performance on standard question-answering datasets without relying much on knowledge bases. Narayan et al.[10] applied grammar-based method to generate question paraphrases whose structures were isomorphic to the gold graphical forms representing correct answers in Freebase. This approach had the potential to mitigate the problems related to cases where the graphical representations corresponding to the original question and the correct answer are not isomorphic. Chen et al.[11] proposed a method based on sentence rewriting for semantic parsing with a particular intention to alleviate the vocabulary mismatch problem between the natural language and its target logical form.

In the clinical domain, although a few works have focused on creating paraphrasing corpus[18], none of them so far has attempted to use the paraphrases in a QA model. This work presents an approach of directly using the generated paraphrases corresponding to clinical questions and utilize the expanded dataset to improve the performance of a semantic parser in clinical QA.

Materials and Methods

The proposed framework for improving semantic parsing using paraphrasing is shown in Figure 1. We begin by introducing the datasets used in this study under Section 1. We then provide a brief overview of our paraphrasing model in Section 2. Further, we detail the process of filtering and applying heuristics to the generated paraphrases

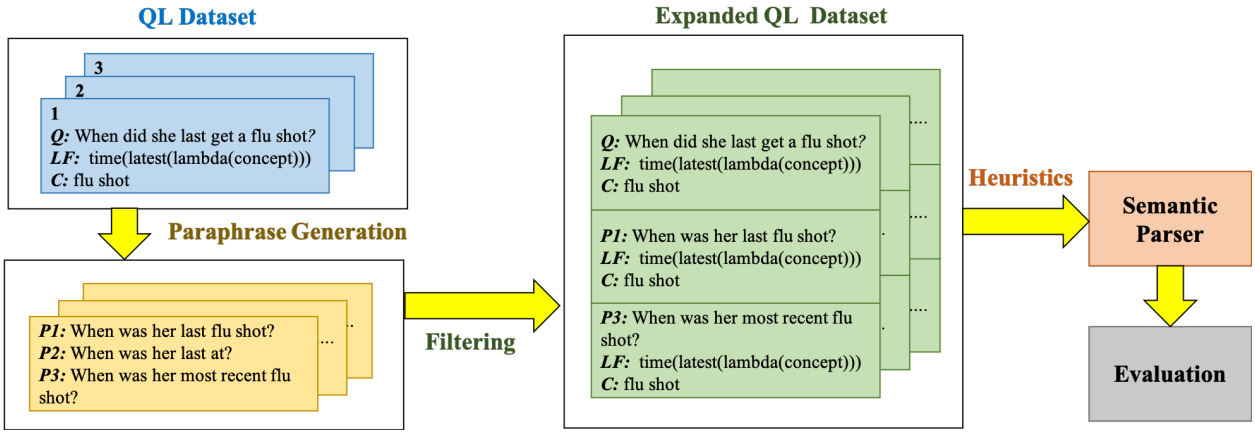

Figure 1: Framework for improving medical question answering using paraphrases. **Q** – Question, **LF** – Logical Form, **C** – Concept.

in Sections 3 and 4 respectively. After providing a brief description of our semantic parsing model in Section 5, we finally explain our evaluation methods in detail under Section 6.

1 Data

For training our paraphrase generation system we utilized a clinical paraphrasing corpus CLINIQPARA. Detailed information about this dataset can be found in our previous work[18], we only provide a brief overview here. This dataset consists of questions which can be answered using EHR data, making it a good choice for improving EHR QA. It contains over $10,000$ unique questions which are spread across 946 distinct paraphrase clusters such that all the questions in a specific cluster are paraphrases of each other.

For semantic parsing, we used an annotated dataset of clinical questions and their corresponding logical forms. Again, we refer the reader to our previous work[17] for complete information about this dataset, only a high-level description is provided here. There are a total of 1000 questions in this dataset which are created using a FHIR server and, hence, are answerable using EHR data. Hereafter, we refer to this corpus as FHIRDATASET.

2 Paraphrase Generation

As mentioned earlier, our paraphrase generation model is based on VAE-LSTM[19]. The model is trained on a set of question-paraphrase (QP) pairs which are constructed using the semantically equivalent paraphrase clusters. We use the same model architecture and hyperparameters as our previous work[18]. However, we employ a different method for QP pair creation than our original work. In our previous work, we generated all combinations of QP pairs from the questions present in a particular paraphrase cluster. Here, we take a conservative approach like Gupta et al.[19] and restrict each paraphrase to appear in at most one pair.

The quality of generated paraphrases was determined using the standard similarity metrics[20] – BLEU (BiLingual Evaluation Understudy)[21], METEOR (Metric for Evaluation of Translation with Explicit ORdering)[22], and TER (Translation Error Rate)[23]. BLEU is a measure of phrase-based similarity between generated and reference or ground truth paraphrases. METEOR further considers synonyms and word stems while calculating the similarity. On the other hand, TER calculates an edit distance (number of additions or deletions required to convert one sentence to another) between the two paraphrases. All of these scores lie between the range of 0 and 100. Note that higher BLEU and METEOR scores are better, in which cases 100 is considered a perfect match. Differently, a lower TER score is preferable with 0 being the best.

We use our best paraphrasing model variant to generate paraphrases for the questions present in the FHIRDATASET. Precisely, we generate 300 paraphrases for each question by choosing the top 3 variations from each of the 100 beam searches (size $= 10$) used for sampling the paraphrases from our generative paraphrasing model.

3 Filtering

All the generated paraphrases are not equally apt and using them directly can adversely impact the performance of QA. Since our generative paraphrasing model can generate duplicate paraphrases, we first filter these as they do not add any variation to the existing dataset. Also, in the context of semantic parsing, the generated paraphrases must contain same number/type of concepts and other references such as person references, temporal references, and measurements. For instance, paraphrases for the question *"When did she last get a flu shot?"* must have a *concept – flu shot* and a *person reference – she*. Any generated paraphrase which does not meet this concept and reference matching criteria is filtered out from further processing. For example, the generated paraphrase *"When was her last at?"* is filtered out because of a missing *concept*. Hence, the filtering step refines the generated paraphrases by removing some of the less useful paraphrases.

4 Heuristics

Though some of the most obvious bad paraphrases are removed in the filtering step, not all the paraphrases which remained after this step are of good quality. We therefore define a set of heuristics for prioritizing the paraphrases. Our first class of heuristics is based on the rank of paraphrases on the basis of various similarity metrics. For this we calculate the similarity scores for the generated paraphrases against their original question (using which the paraphrases were generated). We use these scores to generate a ranked list of paraphrases such that the candidates with better scores are ranked higher. A paraphrase with the best score for a specific metric among the candidates is ranked 1^{st} for that metric. Hence, there are 3 different heuristics in this class, namely, BLEURANK, METEORRANK, and TERRANK.

The second class of heuristics is based on selecting the paraphrases on the basis of absolute values of similarity scores. Precisely, we fix a set of thresholds for each of the similarity metrics and select paraphrases which are better than the threshold for that particular metric. For instance, a threhsold of 40 for BLEU would mean selecting all the paraphrases which scored better than 40. Note that the comparison for better or worse is dependent on the type of metric. For BLEU and METEOR, we come up with a set of threshold values as $\{40, 30, 20, 10, 0\}$ while for TER we choose a set $\{60, 70, 80, 90, 100\}$. The choice of these sets are motivated by the range and count of scores that the generated paraphrases achieved for each of the similarity metrics. For BLEU and METEOR, majority of the scores were centered around 20 while only a few paraphrases scored better than 40. The choices for TER is made similarly while keeping in mind the nature of this score (i.e., lower is better). This class also contains 3 types of heuristics, namely, BLEUTHRESHOLD, METEORTHRESHOLD, and TERTHRESHOLD.

5 Semantic Parsing

The semantic parser used in this study is based on support vector machine (SVM), the full details of which are available in our previous work[13]. We do not update the lexicon, candidate generation rules, and machine learning features of the semantic parser. This is mainly because of our aim to assess the effect of paraphrasing on semantic parsing with minimal human intervention.

6 Evaluation

We consider BASELINE as the case when the original dataset (FHIRDATASET) or its sample was used for training and testing the semantic parser without adding any paraphrases. We experiment with different variations of the heuristics defined in Section 4. To incorporate the randomness of real world scenarios and assess the true effect of adding paraphrases, we choose to run each of our heuristic variations for multiple sample sizes and multiple iterations. Particularly, we run each of the 6 variations (3 each for rank and threshold) with sample sizes as $\{100, 200, 300, 400, 500, 600, 700, 800, 900\}$ and for each sample size the model is run for 50 iterations.

Note that all the variants with paraphrases based on some heuristics are also evaluated using leave-one-out validation. However, to keep the evaluation fair we refrain from using any paraphrase of the test question in training the semantic parser.

Table 1: Performance of our paraphrase generator system using CLINIQPARA. Higher BLEU and METEOR are better whereas a lower TER is better. **CV** – Cross Validation.

Dataset Variant	Metric		
	BLEU	**METEOR**	**TER**
All pairs (CV)	13.25	21.47	91.93
Restricted pairs (CV)	33.29	27.87	65.35
Restricted pairs (best fold)	**45.05**	**31.72**	**57.76**

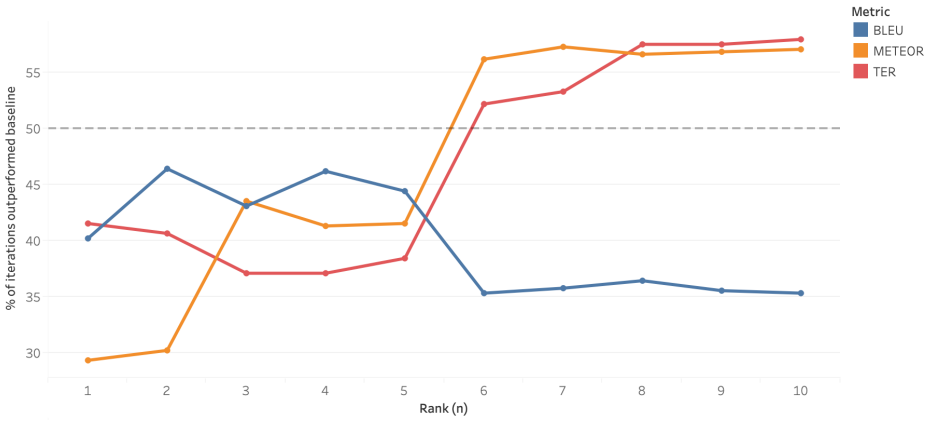

Figure 2: Percentage of iterations (out of 50) performed better than original sample (BASELINE) after adding top *n* paraphrases.

Results

The results of our clinical paraphrasing system are shown in Table 1. Though the variant using the restricted dataset was trained on far fewer paraphrase pairs, it performed better than the variant using all possible pairs. This is likely due to the excessive paraphrase variation in the all-pairs version. For instance, some paraphrase clusters in the CLINIQPARA corpus contain as many as 181 unique question paraphrases. In such cases, the all-pairs variant included each paraphrase in a paraphrase cluster to at most 180 pairs, effectively creating a highly imbalanced corpus biased toward these larger paraphrase clusters.

Figure 2 illustrates the proportion of iterations for which adding the ranked paraphrases resulted in an improvement in the semantic parsing accuracy. We found that, in the case of both METEOR and TER scores, the semantic parser performance improved as we added more ranked paraphrases to the original dataset. It can be seen that, based on the TER metric, adding the top 10 paraphrases to the FHIRDATASET resulted in improving the semantic parser accuracy in 58% of the iterations (indicated by red line in Figure 2). Also note that adding a minimum of 6 top paraphrases (for METEOR and TER metrics) was required to improve the semantic parser performance for more than 50% of the iterations. However, the addition of ranked paraphrases according to the BLEU metric (BLEURANK) did not result in improving the semantic parser's performance for a majority of the iterations (shown by blue line in Figure 2).

Since adding ranked paraphrases based on the TER metric (TERRANK) performed better among all the three metrics in improving the semantic parsing performance (as demonstrated in Figure 2), we specifically present sample size-wise variations in the proportion of iterations where improvement was observed for this metric in Figure 3. It was found that as the sample size was increased there was a large jump in the performance between adding top 5 versus top 6 paraphrases to the samples. More specifically, for the sample sizes 700, 800, and 900, the percentage of outperforming iterations went from 20% to 96%, 32% to 92%, and 44% to 96%, respectively, when the number of paraphrases added was increased from top 5 to top 6. Also, for sample size 600, the percentage increased from 26% to 72% by increasing the number of added paraphrases from 6 to 9. Thus, Figure 3 demonstrates that including larger sample sizes and more ranked paraphrases (usually 6-10) favors the accuracy improvement of the semantic parsing system overall.

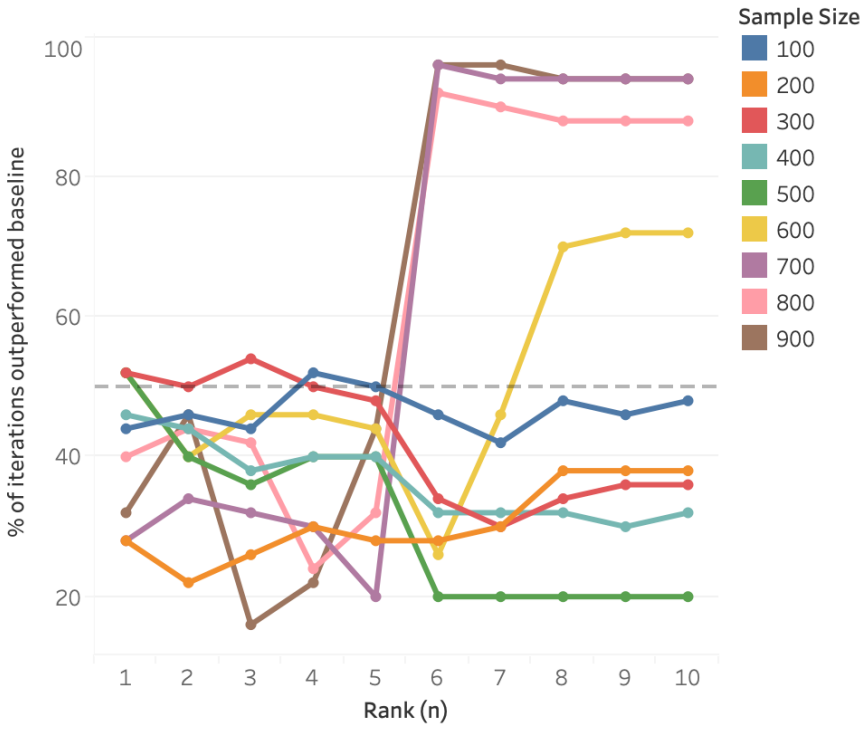

Figure 3: Percentage of iterations (out of 50) TERRANK performed better than original sample (BASELINE) after adding the top n paraphrases for each sample size.

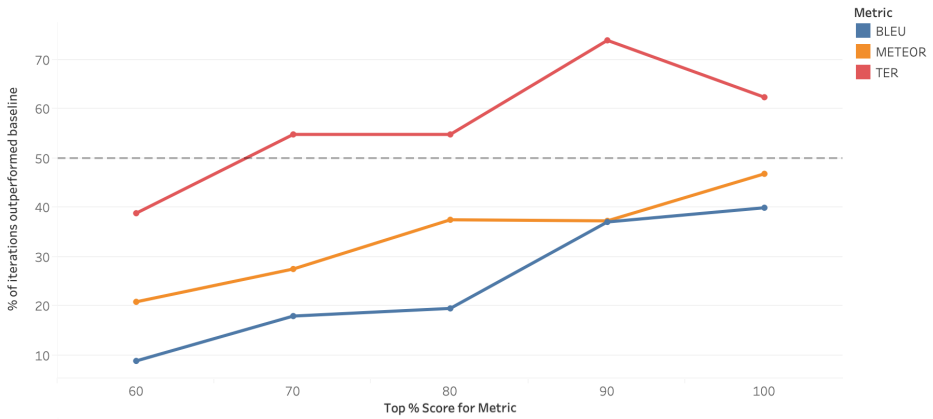

Figure 4: Percentage of iterations (out of 50 iterations) performed better than original sample (BASELINE) after adding the paraphrases of top $x\%$ score of each of the three metrics. For example, 90 along the x-axis represents adding paraphrases corresponding to top 90% metric scores (Top 90% corresponds to scores 0 through 90 for TER and 10 through 100 for both BLEU and METEOR).

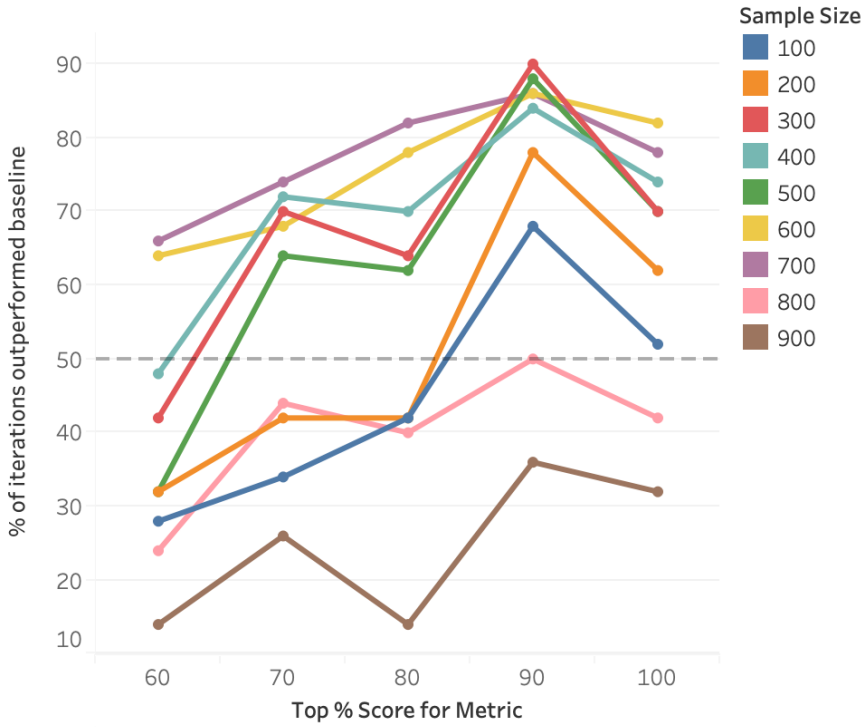

Figure 5: Percentage of iterations (out of 50) TERTHRESHOLD performed better than original sample (BASELINE) after adding the paraphrases of top $x\%$ score of TER metric for each sample size. For example, 90 along the x-axis represents adding paraphrases corresponding to top 90% TER scores (0 through 90).

Instead of adding ranked paraphrases, in Figure 4, we present the proportion of iterations for which adding the top $x\%$ of the paraphrases based on the standard metric scores resulted in an improvement of the semantic parser's performance. Similar to the results in Figure 2, the TER metric (TERTHRESHOLD) performed the best compared to both BLEU (BLEUTHRESHOLD) and METEOR (METEORTHRESHOLD). It was found that the percentage of iterations outperforming the BASELINE increased and reached the highest when all the generated paraphrases were covered for both BLEU and METEOR. In the case of TER, however, adding paraphrases corresponding to the top 90% of the TER scores (TER score 0 – 90) resulted in 74% of the iterations to perform better than the BASELINE when no paraphrases were added.

The percentage of iterations outperforming the BASELINE for each of the sample sizes considered is shown for TER threshold scores in Figure 5. In this case, it was observed that for all the sample sizes in [100, 200, 300, 400, 500, 600, 700], the percentage of iterations where adding paraphrases corresponding to the top 90% TER scores outperformed the BASELINE were 68%, 78%, 90%, 84%, 88%, 86%, and 86%, respectively. Contrary to Figure 3 where ranked paraphrases were added based on TER, in this case including larger sample sizes such as 800 and 900 did not help much in improving the semantic parser's accuracy.

The actual improvements in the performance of semantic parser based on the highest performing metric (i.e., TER) are presented in Figure 6. We note that the absolute improvements in the performance are consistent with the trend displayed by Figures 2 and 4.

Discussion

The results of the paraphrase generator show satisfactory performance over all the evaluation metrics. Though the first variant in Table 1 used many more pairs than the following two variants, it achieved lower performance in all the evaluation metrics. This behavior is bolstered by Gupta et al.[19], who also utilized distinct sentences in the generation of their training pairs.

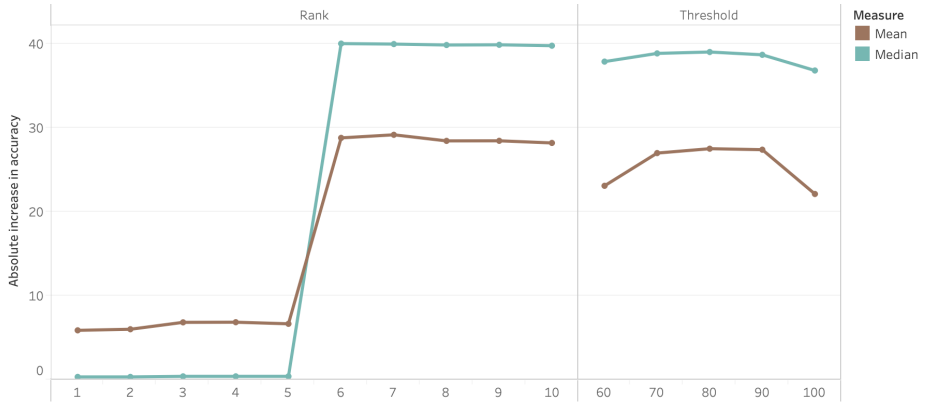

Figure 6: Descriptive statistics on performance improvements for the iterations in which incorporating paraphrases using TER metric outperformed the BASELINE. The results shown here combine the results of all the runs across different sample sizes.

Our experiments show that adding paraphrases is useful in improving the accuracy of a semantic parser compared to training the parser on the original dataset when no paraphrases are added. Among the three metrics, TER is the most effective metric for improving the semantic parser's performance both from the perspective of ranking and threshold score. Specifically, we note that adding more number of ranked paraphrases helps to improve the parser's accuracy. Also, our findings suggest that supplementing the original dataset with paraphrases falling in the range of top 90% TER scores is more useful, with 74% of the iterations outperforming the BASELINE considering all the sample sizes together.

Semantic parsing, or converting a natural language sentence to a machine-understandable logical form, is a crucial component of a QA system for answering questions over structured data. Performance of a semantic parser largely depends on the size of a QL dataset over which it is trained. We focus on increasing the size of a QL dataset using our paraphrase generation system and train our semantic parser over both original and expanded datasets.

In this study, we use the paraphrase generating system for increasing the size of the overall QL dataset. In the future, we plan to incorporate the paraphrasing system in a semantic parser itself[12]. We also aim to experiment utilizing multiple paraphrasing systems.

Conclusion

We demonstrated that an automated paraphrase generation system can be used to improve the performance of clinical QA system. We carried an array of experiments to assess the effect of adding paraphrases to an existing QL dataset on the performance of semantic parsing. We found that adding parapharses based on TER threshold 90 helped in improving the performance of the semantic parser.

Acknowledgments

This work was supported by the U.S. National Library of Medicine, National Institutes of Health (NIH), under award R00LM012104, as well as the Bridges Family Doctoral Fellowship Award.

References

[1] Zhang J, Walji M, Editors. Better EHR: Usability, workflow and cognitive support in electronic health records. National Center for Cognitive Informatics & Decision Making in Healthcare; 2014.

[2] Roman LC, Ancker JS, Johnson SB, Senathirajah Y. Navigation in the electronic health record: A review of the safety and usability literature. Journal of Biomedical Informatics. 2017;67:69–79.

[3] Bhagat R, Hovy E. What Is a Paraphrase? Computational Linguistics. 2013;39(3):463–472.

[4] Duboue P, Chu-Carroll J. 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; 2006. p. 33–36.

[5] Fader A, Zettlemoyer L, Etzioni O. Paraphrase-Driven Learning for Open Question Answering. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2013. p. 1608–1618.

[6] Berant J, Liang P. Semantic Parsing via Paraphrasing. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics; 2014. p. 1415–1425.

[7] Bordes A, Chopra S, Weston J. Question Answering with Subgraph Embeddings. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014. p. 615–620.

[8] Bordes A, Weston J, Usunier N. Open Question Answering with Weakly Supervised Embedding Models. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases; 2014. p. 165–180.

[9] Dong L, Wei F, Zhou M, Xu K. 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. vol. 1; 2015. p. 260–269.

[10] Narayan S, Reddy S, Cohen SB. Paraphrase Generation from Latent-Variable PCFGs for Semantic Parsing. In: Proceedings of the 9th International Natural Language Generation conference; 2016. p. 153–162.

[11] Chen B, Sun L, Han X, An B. Sentence Rewriting for Semantic Parsing. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics; 2016. p. 766–777.

[12] Dong L, Mallinson J, Reddy S, Lapata M. Learning to Paraphrase for Question Answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics; 2017. p. 875–886.

[13] Roberts K, Patra BG. A Semantic Parsing Method for Mapping Clinical Questions to Logical Forms. In: AMIA Annual Symposium Proceedings. vol. 2017. American Medical Informatics Association; 2017. p. 1478–1487.

[14] Kamath A, Das R. A Survey on Semantic Parsing. In: Automated Knowledge Base Construction; 2019.

[15] Health Level Seven International. Welcome to FHIR;. Available from: `https://www.hl7.org/fhir/`.

[16] Roberts K, Demner-Fushman D. Annotating logical forms for EHR questions. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, LREC 2016; 2016. p. 3772–3778.

[17] Soni S, Gudala M, Wang DZ, Roberts K. Using FHIR to Construct a Corpus of Clinical Questions Annotated with Logical Forms and Answers. In: AMIA Annual Symposium Proceedings. vol. 2019. American Medical Informatics Association; 2019. p. 1207–1215.

[18] Soni S, Roberts K. A Paraphrase Generation System for EHR Question Answering. In: Proceedings of the 18th BioNLP Workshop. Association for Computational Linguistics; 2019. p. 20–29.

[19] Gupta A, Agarwal A, Singh P, Rai P. A Deep Generative Framework for Paraphrase Generation. In: Thirty-Second AAAI Conference on Artificial Intelligence; 2018. p. 5149–5156.

[20] Madnani N, Tetreault J, Chodorow M. 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; 2012. p. 182–190.

[21] Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics; 2002. p. 311–318.

[22] Lavie A, Agarwal A. 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. Association for Computational Linguistics; 2007. p. 228–231.

[23] Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J. A study of translation edit rate with targeted human annotation. In: Proceedings of association for machine translation in the Americas. vol. 200; 2006. p. 223–231.