How You Ask Matters: The Effect of Paraphrastic Questions to BERT Performance on a Clinical SQuAD Dataset

Sungrim Moon[1], **Jungwei Fan**[1,2]
[1]Division of Digital Health Sciences
[2]Center for the Science of Health Care Delivery
Mayo Clinic
200 1st Street SW, Rochester, MN 55905
moon.sungrim@mayo.edu
fan.jung-wei@mayo.edu

Abstract

Reading comprehension style question-answering (QA) based on patient-specific documents represents a growing area in clinical NLP with plentiful applications. Bidirectional Encoder Representations from Transformers (BERT) and its derivatives lead the state-of-the-art accuracy on the task, but most evaluation has treated the data as a pre-mixture without systematically looking into the potential effect of imperfect train/test questions. The current study seeks to address this gap by experimenting with full versus partial train/test data consisting of paraphrastic questions. Our key findings include 1) training with all pooled question variants yielded best accuracy, 2) the accuracy varied widely, from 0.74 to 0.80, when trained with each single question variant, and 3) questions of similar lexical/syntactic structure tended to induce identical answers. The results suggest that how you ask questions matters in BERT-based QA, especially at the training stage.

1 Introduction

In clinical NLP, there has been vital interest in developing question-answering (QA) systems, e.g., AskHERMES (Cao et al., 2011), MiPACQ (Cairns et al., 2011), and MEANS (Abacha and Zweigenbaum, 2015). One specific type of clinical QA targets on locating any suitable answer within a given document (a.k.a. reading comprehension), which is helpful for answering patient-specific questions based on information mentioned in clinical notes. Recently, BERT (Devlin et al., 2018) and its derivatives have struck impressive success in this task for general English, represented by SQuAD (Rajpurkar et al., 2018), and for clinical text with promising results (Wen et al., 2020; Soni and Roberts, 2020).

However, an under-explored area in performing BERT-assisted QA is: how the system would behave if the input question is asked in a different (paraphrastic) way? Most existing experiments have assumed that the train and test data belong to a closed space with pre-assembled syntactic and lexical diversity (i.e., paraphrastic questions) representing what the users could ever ask, and faithfully evaluate the both diverse train/test sets in a symmetric manner. In practice, there are at least two possible scenarios concerning the potential effect from a differently-asked test question: 1) the system is trained with, or has seen, the question construct, 2) the system has never seen the question construct during training. Here by "construct" we refer to paraphrases like: "why is the patient prescribed medication-X?" versus "why does the patient take medication-X?". More examples are in Table 1.

Ideally, a BERT QA model is supposed to provide a consistent answer as long as the user asked a semantically-equivalent (paraphrastic) question. This is important in a production system because a user should not be required to ask only questions conforming to "template" constructs. Therefore, in this study we set to understand how such paraphrastic perturbation in asking would affect a BERT-based QA model. We used a dataset that contained finite question constructs, but purposefully injected experiments with using

Proceedings of the 3rd Clinical Natural Language Processing Workshop, pages 111–116

limited constructs in the training and/or testing to simulate the asymmetric perturbations of interest. For example, training on only one question construct and testing on the other different constructs (i.e., unseen ways of asking).

Our major findings can be summarized as follows:

1. Models trained with all pooled constructs still gave the best accuracy.
2. When training was limited to each single construct, certain constructs gave overall higher accuracy across all test constructs. Accuracy also varied depending on the test construct, but the effect was not strong as the choice of the training construct.
3. Certain test question constructs tended to induce identifical answers, as revealed via a clustering analysis.

2 Related work

Pampari et al. (2018) created the emrQA corpus by template-based semantic extraction from the i2b2 NLP challenge datasets (i2b2 2019). The emrQA includes more than 400,000 QA pairs and has served as a valuable resource in clinical QA research (Wen et al., 2020; Soni and Roberts, 2020). Most of the previous studies reported strong performance by BERT, especially when it was pre-trained with domain-specific text, e.g., the Clinical BERT (Alsentzer et al., 2019) was trained with about 2 million clinical notes from the MIMIC-III database (Johnson et al., 2016). For compatibility with using BERT-based reading comprehension QA, the SQuAD format is commonly adopted. Task-wise, the SQuAD 2.0 (Rajpurkar et al., 2018) also introduced unanswerable questions that require QA systems to know when not to answer if no suitable evidence is present in the text.

Besides the relevant backgrounds above, there have been NLP studies that reported the effect of paraphrastic questions in QA system performance. Buck et al. (2017) and Dong et al. (2017) developed approaches to paraphrasing questions for optimal answer accuracy in retrieval-based QA, where candidate answers were searched and ranked from a large set of documents. The closest work for reading comprehension QA we identified was by Gan and Ng (2019), which investigated the effect of question variants on a general English SQuAD dataset. They demonstrated that unseen paraphrastic test questions hurt the accuracy of deep learning QA models, and proposed a countermeasure by pre-augmenting the training data with machine-generated paraphrastic questions.

3 Methods

3.1 Research questions

We designed our experiments around the following three research questions:

- How does the accuracy change by training the model with a pool of multiple question constructs versus training by only each construct?
- How does the accuracy vary across different training question constructs and across different test question constructs?
- Do some of the test question constructs tend to elicit similar answers out of a trained model?

3.2 Dataset

We used the emrQA as the base dataset and selected only those "why"-questions in this study due to our application research interest. Within the why-QA subset, we further considered three levels of QAs, from broad to specific:

All – all the why-QAs (see Appendix A).

Med – why-QAs about medication.

Q0~Q8 – the 9 individual question constructs in the *Med* set, as elaborated in Table 1.

Label	Question construct
Q0	Why was [medication] prescribed?
Q1	Why was [medication] originally prescribed?
Q2	Why was the patient prescribed [medication]?
Q3	Why is the patient prescribed [medication]?
Q4	Why has the patient been prescribed [medication]?
Q5	Why was the patient on [medication]?
Q6	Why is the patient on [medication]?
Q7	Why does the patient take [medication]?
Q8	Why is the patient taking [medication]?

Table 1: The 9 different question constructs in the medication why-QAs.

Corpus		*All*	*Med*	*Qi*
Train	HasAns	8,835	3,807	423
	NoAns	7,985	3,726	414
Dev	HasAns	3,024	1,260	140
	NoAns	2,725	1,242	138
Test	HasAns	9,232	4,572	508
	NoAns	8,204	4,428	492

Table 2: Number of the train/dev/test QAs in the three cascaded levels. *Qi* represents each of *Q0-Q8*.

All the QAs were prepared into the SQuAD 2.0 format. The train/dev/test splits are detailed in Table 2, which also breaks down with showing the answerable (HasAns) versus unanswerable (NoAns) QA counts. The dev partition was for setting the optimal threshold of "do not answer" before processing the final held out test partition. Note that the numbers in column *Qi* were made identical across *Q0~Q8* respectively, so there should not be any bias in inflating any of them.

3.3 Training and evaluation

All of the models started from the pre-trained Clinical BERT, followed by a modest fine-tuning with 1,833 general English why-QAs from the SQuAD 2.0 corpus. On top of that, the experiments involved three parts (Table 2 for denotations):

A. Fine-tune/calibrate on *All* train/dev (one model), test on each *Qi* test set.

B. Fine-tune/calibrate on *Med* train/dev (one model), test on each *Qi* test set.

C. Fine-tune/calibrate on *Qi* train/dev (nine models), test on each *Qi* test set. This is basically crossover between *Q0~Q8*.

Each fine-tuning (or simply referred as "training") was done with 10 epochs, batch_train_size=32, learning_rate=3e-5, and max_seq_length=128. The jobs were run on a Tesla V100 with compute capability 7.0 and 18 GB of memory.

The official SQuAD 2.0 evaluation script was used, and we reported primarily the accuracy as F1-weighted overlaps between the gold and the system answers. As a semi-qualitative assessment of question similarity (in terms of the triggered model behavior), we computed the number of agreed (case-insensitive and remove articles) answers between each pair of *Qi* test sets in experiment B above and performed hierarchical clustering to group the 9 question constructs.

4 Results

4.1 Pooled training made stronger model

The model accuracies are reported in Figure 1, where Figure 1b is specifically to show precision (positive predictive value, or PPV) on those HasAns QAs that were indeed answered by each model. It can be seen that the *All* model (blue line at the top) outperformed the *Med* model (orange line) and every individual *Qi* model, suggesting that training with additional non-medication why-QA entries still benefited the accuracy. The benefit is more apparent in PPV (Figure 1b) and fluctuates mildly across the test constructs *Q0~Q8* (X-axis). In comparison to the individual *Qi* models, the pooled *Med* model also exhibits clear advantage but with varying margins (elaborated in 4.2).

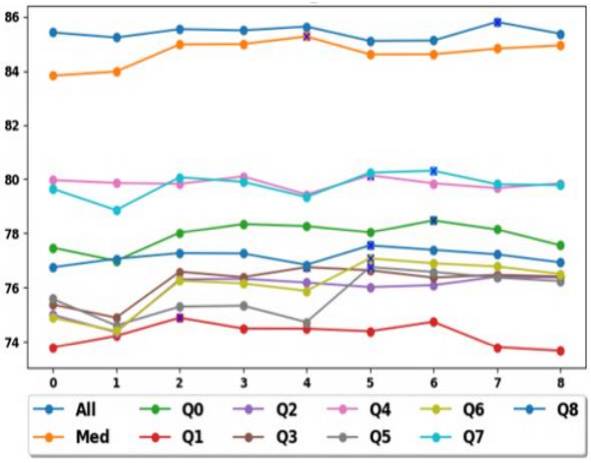

Figure 1a: Accuracies of the different models. X-axis: test question construct, Y-axis: accuracy.

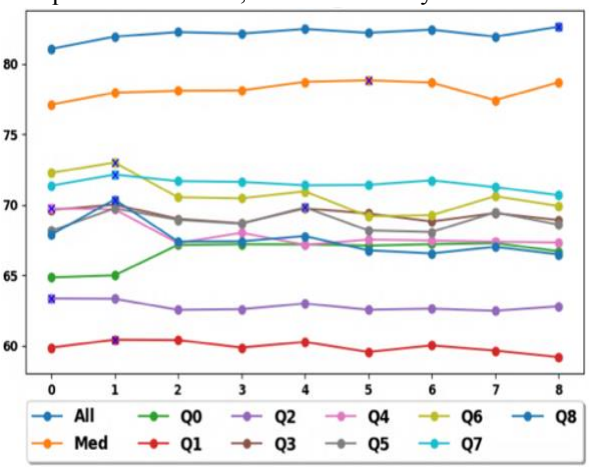

Figure 1b: Accuracies of the different models, on the answerable and indeed answered subset.

4.2 Accuracy varied depending on the question construct

The accuracy appears to be strongly affected by which specific question construct was used for training. For example, *Q7* (cyan line in Figure 1) exhibits about 4% drop compared to *Med* (orange), while *Q1* (red) has 10% or wider gap below *Med*. Manual inspection of 569 disagreements between *Q1* and *Q7* did show the *Q1* model frequently refrained from answering (141/569=25%) or gave irrelevant answers (286/569=50%). In addition, such question-dependent behavior changes again when we look at PPV specifically. For example, in Figure 1a the *Q4* model (pink) performs comparably well as the *Q7* model, but in Figure 1b its relative rank drops to the middle tier indicating that *Q4* gave many incorrect answers.

Within each line (a trained model), the variance of accuracy across different test questions does not appear as drastic (up to ~2%) compared to that observed across models. However, one puzzling observation is that the peak accuracy within each line of *Q0~Q8* is usually not at where the train and test question align. (e.g., train on *Q0*, test on *Q0*)

4.3 Some questions were more likely to obtain same answers

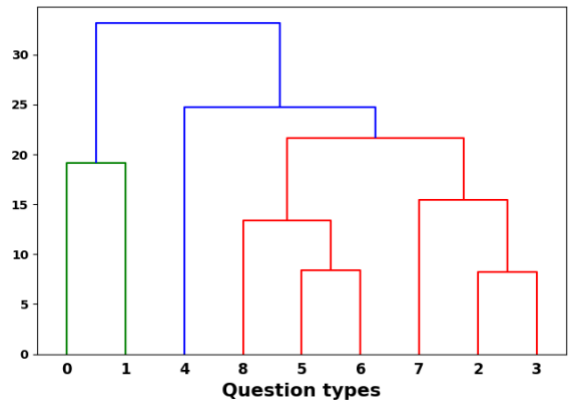

Figure 2: Question similarity clustering based on answer agreement. X-axis: test questions *Q0~Q8*. Y-axis: # of agreed answers by the model *Med*.

The hierarchical dendrogram for clustering the question constructs is shown in Figure 2. When trained on the pooled of *Med* constructs (i.e., the orange line in Figure 1a), some of the test question constructs turned out to yield closer answers than others. Specifically, *Q0* and *Q1* form a cluster (green in Figure 2), *Q4* sort of stands alone, and the others form another subtree (red) enclosing further sub-clusters. Some intuitive explanations could be derived by inspecting the lexical/syntactic contents of the questions: For example, between *Q0* and *Q1* the only difference is an additional "originally" in *Q1*. Likewise, a single switch of tense between "was" and "is" appears to account for the two tight clusters (*Q2* and *Q3)* as well as (*Q5* and *Q6)*.

5 Discussion

5.1 Recap the rationale

Many would think it a trivial fact that different questions surely contribute to varying answers. However, we believe it is worth a break-down analysis beyond the monolithic thinking of just "the more the better". No matter how well planned, one can always legitimately ask "what if" the training questions were not exhaustive and some real user threw in unexpected questions. Therefore, this study was meant to expose such behavior out of a BERT-based QA model by putting it under the stress of partial train/test questions.

5.2 Featured findings, raised questions

Our findings did validate some trivial knowledge such as the more robust models gained by pooling diverse training data and that similar questions tended to elicit similar answers. On the other hand, we have findings to highlight: 1) a model's accuracy is determined strongly by what questions it is trained on and not as much by what test questions it is asked to answer, 2) there appear to be "better ways to ask" especially in training that would yield generally higher accuracy. That said, our findings somewhat circle back to corroborating the common strategy that focuses on enriching the training diversity to achieve robust performance – plus, the "good" questions need adequate presence.

The micro-level, quality-oriented observations could not be revealed without diving into those question-specific comparisons. For fundamental computational linguistics research, our findings pointed an interesting direction to explore: why certain question constructs (e.g., *Q7* "Why does the patient take [medication]?") appear to be more transferrable (at least accuracy-wise) after being learned by BERT? Methods for inspecting the attention mechanism under the hood might help, but we suspect that new approaches of even better interpretability likely need to be designed. Specifically, one phenomenon that puzzled us was the peak accuracy did not always happen at the point when a single-question model was tested on questions of the same training construct.

5.3 Limitations

Our experiments on why-QAs and the focus on medication-related questions could limit the generalizability. The diversity of those question variants was bound to what emrQA had offered, and we did not know how that compared to the natural distribution of variants asked by humans. Besides, the emrQA corpus might have embedded noise and quality issues (Yue et al., 2020) that affected the results. Lastly, we still do not have explanation to many findings, and it is unclear if BERT can represent other QA models especially in terms of the question-specific behaviors.

5.4 Future work

The current study looked mainly into syntactic variants of the questions, and we will further research the independent or interactive effect of lexico-syntactic variants of concepts (e.g., medication) mentioned in both the question and answer document. Based on our findings, we plan to experiment optimizing the QA accuracy through ensemble approaches such as voting - the hypothesis is the chance of achieving a convergent (correct) answer should be increased by asking the same question in different ways.

6 Conclusion

We found that how you ask matters in a BERT-based clinical QA task, especially at the training stage. By controlling the train and test questions to individual lexical/syntactic constructs, our crossover evaluation showed that certain question constructs consistently yielded higher accuracy. Accordingly, it suggests that the most effective way to secure robust performance is still by training with diverse, sizable questions. Our results also brought up a somewhat nuanced inquiry: how come some question constructs seem to act "linguistically superior" to others, and whether it is a universal or BERT-dependent phenomenon?

Acknowledgments

Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY. We thank the anonymous reviewers for their constructive feedback.

References

Asma B. Abacha and Pierre Zweigenbaum. 2015. MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information processing & management* 51, no. 5:570-594. https://www.sciencedirect.com/science/article/pii/S0306457315000515

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B.A. McDermott. 2019. Publicly available clinical bert embeddings. *ClinicalNLP workshop at NAACL*. https://arxiv.org/pdf/1904.03323.pdf

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Wojciech Gajewski, Andrea Gesmundo, Neil Houlsby, and Wei Wang. "Ask the right questions: Active question reformulation with reinforcement learning." arXiv preprint arXiv:1705.07830 (2017). https://arxiv.org/abs/1705.07830

Brian L. Cairns, Rodney D. Nielsen, James J. Masanz, James H. Martin, Martha S. Palmer, Wayne H. Ward, and Guergana K. Savova. 2011. The MiPACQ clinical question answering system. *In AMIA annual symposium proceedings*, vol. 2011, p. 171. American Medical Informatics Association. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243235

YongGang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J. Cimino, John Ely, and Hong Yu. 2011, AskHERMES: An online question answering system for complex clinical questions. *Journal of biomedical informatics* 44, no. 2:277-288. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3433744/pdf/nihms400508.pdf

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Computing Research Repository*, https://arxiv.org/abs/1810.04805

Li Dong, Jonathan Mallinson, Siva Reddy, and Mirella Lapata. "Learning to paraphrase for question answering." arXiv preprint arXiv:1708.06022 (2017). https://arxiv.org/abs/1708.06022

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the Robustness of Question Answering Systems to Question Paraphrasing. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy. Association for Computational Linguistics. https://www.aclweb.org/anthology/P19-1610.pdf

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, https://www.nature.com/articles/sdata201635.pdf

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. *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 4356–4365, Hong Kong, China. Association for Computational Linguistics. https://www.aclweb.org/anthology/D19-1445/

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. arXiv preprint arXiv:1809.00732

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *In Proceedings of the Association for Computational Linguistics*. https://arxiv.org/abs/1806.03822

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250.

Sarvesh Soni and Kirk Roberts. 2020. Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering. *In Proceedings of the LREC*, p5534–5540. https://www.aclweb.org/anthology/2020.lrec-1.679.pdf

The i2b2 2019 NLP Research Data Sets. Secondary The i2b2 NLP Research Data Sets. https://www.i2b2.org/NLP/DataSets/Main.php (accessed August, 2020)

Andrew Wen, Mohamed Y. Elwazir, Sungrim Moon, and Jungwei Fan. 2020. Adapting and evaluating a deep learning language model for clinical why-question answering, *JAMIA Open*, 3(1):16-20. https://academic.oup.com/jamiaopen/article/3/1/16/5722318

Xiang Yue, Bernal J. Gutierrez, and Huan Sun. "Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset." arXiv preprint arXiv:2005.00574 (2020). https://arxiv.org/abs/2005.00574

Appendix A. All the why-question types

Why does the patient take |medication|
Why has the patient been prescribed |medication|
Why is the patient on |medication|
Why is the patient prescribed |medication|
Why is the patient taking |medication|
Why was |medication| originally prescribed
Why was |medication| prescribed
Why was the patient on |medication|
Why was the patient prescribed |medication|
Why did the patient have |test|
Why did they patient get |test|
Why was |test| done on this patient
Why was |test| performed
Why did the patient have |treatment|
Why did the patient need |treatment|
Why is the patient on |treatment|
Why was the patient on |treatment|