Annotating and Characterizing Clinical Sentences with Explicit Why-QA Cues

Jungwei Fan
Division of Digital Health Sciences, Mayo Clinic
200 1st Street SW, RO-HA-2-CSHCD, Rochester, MN 55905
fan.jung-wei@mayo.edu

Abstract

Many clinical information needs can be stated as why-questions. The answers to them represent important clinical reasoning and justification. Clinical notes are a rich source for such why-question answering (why-QA). However, there are few dedicated corpora, and little is known about the characteristics of clinical why-QA narratives. To address this gap, the study performed manual annotation of 277 sentences containing explicit why-QA cues and summarized their quantitative and qualitative properties. The contributions are: 1) sharing a seed corpus that can be used for various QA-related training purposes, 2) adding to our knowledge about the diversity and distribution of clinical why-QA contents.

1 Introduction

The thought process involved in clinical reasoning and decision-making can be naturally framed into a series of questions and answers. In addition to the tangible value as handy assistance, making computers handle question-answering (QA) is considered a remarkable achievement in artificial intelligence. Accordingly, there has been vital interest in developing clinical QA systems, e.g., AskHERMES (Cao et al., 2011), MiPACQ (Cairns et al., 2011), and MEANS (Abacha & Zweigenbaum, 2015). Among the targets, why-QA represents a special category that deals with cause, motivation, circumstance, and purpose (Verberne, 2006). Within the top ten question types asked by family doctors (Ely et al., 1999), 20% of them can actually be paraphrased into a why-question. Besides the sizable presence, clinical why-QA is both semantically and pragmatically rich because: 1) toward the deep explanatory end the task almost resembles expert-level synthesis and inference, 2) toward the shallower end it usually involves identifying the documented reason that a decision was made.

It is worth clarifying here two different scenarios that QA tasks are defined. The first aligns more along consulting knowledge sources to answer a question that is not patient-specific, e.g., *Why do phenobarbital and Dilantin counteract each other?* This is also the scenario that most of the existing clinical QA systems handle. The second scenario (focus of this study) is to find the answer within a given document (a.k.a. reading comprehension), which can especially benefit patient-specific QA based on information mentioned in clinical notes. In the general domain such reading comprehension QA has more than a decade of research, with widely used corpora such as the SQuAD (Rajpurkar, Zhang, Lopyrev, & Liang, 2016) and that by Verberne, Boves, Oostdijk, & Coppen (2006). There have not been comparable resources in the clinical domain until a couple of works in 2018 (see Related work section).

The recently developed corpora in clinical reading comprehension QA are extremely valuable, but also limited with regard to why-QA research because 1) their coverage and analysis did not emphasize on why-questions, 2) the annotation methods could have missed many representative why-QA targets. Therefore, the current study aims to compensate for these oversights through systematic inspection into clinical sentences that contain the intuitive cues "because" and "due to". The rationale is: we might never know what can be missed by diving right into complex cases, unless the low-hanging offers are well understood first. In fact, the results revealed many informative clinical topics and patterns involved in why-QA. Along with the diverse topics, the well-formed linguistic constructs based on the two unambiguous cues make this small corpus an ideal seed training set to stabilize models or to bootstrap other solutions.

Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 101–106
Minneapolis, Minnesota, June 7, 2019.

2 Related work

There has been considerable annotation research for why-QA in non-medical domains. As part of developing a why-QA system, Higashinaka & Isozaki (2008) used information retrieval to search documents possibly relevant to each why-question, followed by manual validation of qualified QA pairs. Mrozinski, Whittaker, & Furui (2008) used Mechanical Turk to recruit annotators for reading Wikipedia articles and generating why-questions based on the contents. Dulceanu et al. (2018) applied web scraping over community forums to collect why-QAs about Adobe Photoshop usage. The answer quality was backed either by questioner feedback or by community votes. Prasad & Joshi (2008) proposed leveraging causal relations in the richly annotated Penn Discourse Treebank to derive why-QAs.

In the clinical domain there were two corpora developed for reading comprehension QA based on electronic medical records (EMR), and both had broad coverage not limited to only why-QAs. In Raghavan, Patwardhan, Liang, & Devarakonda (2018), medical students were presented with structured and unstructured EMR information of each patient and were instructed to come up with realistic questions for a hypothetical office encounter. The patient's notes were then loaded into an annotation tool for them to mark answer text spans. Pampari, Raghavan, Liang, & Peng (2018) developed emrQA, a large clinical QA corpus generated through template-based semantic extraction from the i2b2 NLP challenge datasets.[*] The emrQA contains 7.5% of why-QAs, but they mainly ask about why the patient received a test or treatment, due to the partial interest of the original challenge annotations.

3 Methods

The study notes were from the 2010 i2b2/VA NLP challenge (Uzuner, South, Shen, & DuVall, 2011), obtained through an academic data use agreement.[†] The corpus consists of 426 discharge summaries from Partners Healthcare and Beth Israel Deaconess Medical Center. The two considerations in choosing this dataset were: 1) the sentences were pre-chunked that made the downstream analysis easier, 2) it overlapped with the emrQA corpus and thus allowed comparison of coverage, etc.

Case-insensitive word search was performed using "because" and "due to" into the 426 notes. To avoid massive false positives, highly ambiguous cues such as "for" were avoided in this pilot study. The author then manually reviewed the 280 hit sentences, of which 79 were from "because" and 201 from "due to". The review involved two tasks: 1) generate a QA pair from the sentence, and 2) categorize the question anchor and the answer. Using the following sentence as an example:

> *The patient had urinary tract infection and received Bactrim, which was stopped later because of diarrhea.*

The generated QA pair was:

Q: Why was the Bactrim for urinary tract infection stopped?
A: diarrhea

It was required that each answer must come from a substring of the source sentence. For each annotation, the line number and character offset of the answer were preserved so as to facilitate computable reuses. The types of question anchors and their answers were induced and consolidated throughout the entire review process. For example, the categorization for the specific QA pair above was:

Question anchor: medication avoidance
Answer reason: adverse effect

Upon completing the annotation, descriptive statistics were derived to show notable properties:

- Sentence coverage of the annotated why-QAs as compared to that of emrQA (Figure 1).
- Distribution of clinical notes with respect to the number of sentences that contain either of the why-cues (Table 1).
- Distribution of the categorized why-question anchors and answer types (Tables 2, 3, and 4).

4 Results

As a simple comparison of the question sources, sentence coverage of the annotated why-QAs versus the emrQA why-associated entries is

[*] https://www.i2b2.org/NLP/DataSets/
[†] Complying with the i2b2 NLP data use agreement, examples in this paper have been modified and differ from the original text.

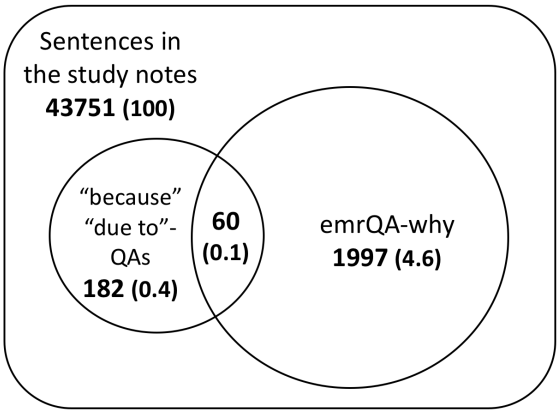

Figure 1: Venn diagram comparing the # (%) of the annotated sentences to that of emrQA

# of cue-containing sentences in the note	# (%) of notes
0	280 (65.7)
1	74 (17.4)
2	39 (9.2)
3	16 (3.8)
4	12 (2.8)
5	2 (0.5)
6	2 (0.5)
7	1 (0.2)

Table 1: Distribution of notes containing the "because" or "due to" cue

Why-question anchor	Answer reason type	# of QAs (%)
abnormal manifestation	disease-caused	51 (20.8)
abnormal manifestation	adverse effect	19 (7.8)
abnormal manifestation	manifestation elaborated	6 (2.4)
abnormal manifestation	disease interaction	3 (1.2)
abnormal manifestation	environment factor	2 (0.8)
procedure disposition	clinical indication	39 (15.9)
procedure disposition	patient preference	1 (0.4)
consultation, admission, discharge, or transfer event	clinical indication	34 (13.9)
consultation, admission, discharge, or transfer event	patient preference	2 (0.8)
consultation, admission, discharge, or transfer event	environment factor	1 (0.4)
medication avoidance	adverse effect	14 (5.7)
medication avoidance	disease interaction	8 (3.3)
medication avoidance	patient preference	2 (0.8)
medication avoidance	disease attribute	1 (0.4)
medication avoidance	procedure interaction	1 (0.4)
procedure avoidance	disease interaction	12 (4.9)
procedure avoidance	patient preference	3 (1.2)
procedure avoidance	procedure interaction	3 (1.2)
procedure avoidance	adverse effect	2 (0.8)
procedure avoidance	disease attribute	2 (0.8)
procedure avoidance	patient attribute	2 (0.8)
procedure unsuccessful	patient attribute	9 (3.7)
procedure unsuccessful	disease interaction	6 (2.4)
procedure unsuccessful	environment factor	2 (0.8)
procedure unsuccessful	disease attribute	1 (0.4)
procedure unsuccessful	disease-caused	1 (0.4)
procedure unsuccessful	procedure interaction	1 (0.4)
medication administered	clinical indication	12 (4.9)
medication administered	patient attribute	1 (0.4)
patient interpretation	patient assessment	1 (0.4)
procedure effective	patient attribute	1 (0.4)
social background	family factor	1 (0.4)
nonmedical treat	patient preference	1 (0.4)

Table 2: Detailed distribution of QA pairs by type

illustrated in Figure 1. There were a total of 43,751 sentences (including those short section headers) in the study corpus of 426 clinical notes. The emrQA used 2,057 sentences in generating its QA pairs, which were basically all about reasons for ordering a test or treatment. The cue-based annotation used 242 sentences, yet the derived why-QAs were much more diverse (see Table 2). There were 60 sentences used by both.

Two reasons that the original 280 hit sentences dropped to the 242 distinct annotated sentences were: 1) there were 3 sentences actually containing both cues, 2) 35 of the sentences were not usable to generate a QA pair because of anaphora. Note that it is possible for a double-cue sentence to generate two separate questions because of different why-anchors. As for the prevalence of the two cues, Table 1 shows that more than 30% (100% – 65.7%) of the study notes had at least one cue, with as many as 7 cue-containing sentences within one note.

The full categorization and distribution of the annotated why-QAs are shown in Table 2, while the distributions aggregated by the question anchors and answer reason types are in Table 3 and Table 4 respectively.

Example contexts of some noteworthy why-QA categories as follows:
[abnormal manifestation → disease-caused]
>> Why did his arm show poor motor movement? → loss of sensation
[procedure disposition → clinical indication]
>> Why was ultrafiltration fluid removal done at each dialysis session? → volume overload
[medication administered → clinical indication]
>> Why was he given levofloxacin? → gram-positive cocci
[consultation/admission, discharge, or transfer event → clinical indication]
>> Why was she admitted? → cholangitis
[procedure avoidance → disease interaction]
>> Why was the dobutamine stress test deferred? → patient having fever and hypotension
[procedure unsuccessful → patient attribute]
>> Why the GI PEG placement failed? → difficult anatomy
[procedure avoidance → patient preference]
>> Why the patient refused transesophageal echo? → did not want to swallow the probe
[medication avoidance → procedure interaction]
>> Why was metformin held temporarily? → CT with contrast

Why-question anchor	# of QAs (%)
abnormal manifestation	81 (33.1)
procedure disposition	40 (16.3)
consultation, admission, discharge, or transfer event	37 (15.1)
medication avoidance	26 (10.6)
procedure avoidance	24 (9.8)
procedure unsuccessful	20 (8.2)
medication administered	13 (5.3)
patient interpretation	1 (0.4)
procedure effective	1 (0.4)
social background	1 (0.4)
nonmedical treat	1 (0.4)

Table 3: Distribution of QA pairs aggregated by the why-question anchor types

Answer reason type	# of QAs (%)
clinical indication	85 (34.7)
disease-caused	52 (21.2)
adverse effect	35 (14.3)
disease interaction	29 (11.8)
patient attribute	13 (5.3)
patient preference	9 (3.7)
manifestation elaborated	6 (2.4)
environment factor	5 (2.0)
procedure interaction	5 (2.0)
disease attribute	4 (1.6)
patient assessment	1 (0.4)
family factor	1 (0.4)

Table 4: Distribution of QA pairs aggregated by the answer reason types

5 Discussion

Although the explicit cues contributed a relatively small set of why-QAs, they exhibit a wealth of subject contours for further investigation. The majority of the emrQA why-questions correspond to the two anchor categories procedure disposition and medication administered, together covering only 21.6% among the various anchors in Table 3. Notably, the top anchor category abnormal manifestation (33.1%) concurs with the most commonly asked why-equivalent questions surveyed by (Ely et al., 1999), i.e., *What is the cause of a symptom or finding?* This concordance implies clinicians tend to explicitly document reasons on certain topics they feel like inquiring in practice as well. Moreover, annotations of medication avoidance and procedure avoidance (together making 20.4% of the anchors) host rich knowledge that is worth capturing systematically.

For example, procedure interaction and disease interaction (e.g., risk from comorbidity) are typical reasons in avoiding certain intervention.

Even though the annotations involve only simple cues and single-sentence contexts, they should benefit the training of QA systems. It is known that such instances of atomic and regular structure can help stabilize/smooth the behavior of statistical models. The other possible route is to use the annotations as seed examples and train a question-generation model that automatically asks why-questions as additional training data. Although the study was short of resource to include experimental validation, it is hoped that at least as a self-contained descriptive analysis the results can be informative to the clinical NLP community.

The representativeness of the study was limited by using only discharge summaries and the two specific cues. The annotations with the complete answer available within one sentence do not touch upon complex scenarios that require synthesizing cross-sentence information. The questions from rephrasing sentences may lack natural intent and diversity, which was a limitation likely shared by repurposing NLP challenge annotations as done in emrQA. This study used only one annotator, which would introduce subjectivity especially in categorizing the QAs.

The annotations by this study are available at https://github.com/Jung-wei/ClinicalWhyQA

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.

I would like to thank the anonymous reviewers for their constructive feedback.

References

Abacha, A. B., & Zweigenbaum, P. (2015). MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies. *Information processing & management, 51*(5), 570-594.

Cairns, B. L., Nielsen, R. D., Masanz, J. J., Martin, J. H., Palmer, M. S., Ward, W. H., & Savova, G. K. (2011). The MiPACQ clinical question answering system. *AMIA Annu Symp Proc, 2011*, 171-180.

Cao, Y., Liu, F., Simpson, P., Antieau, L., Bennett, A., Cimino, J. J., . . . Yu, H. (2011). AskHERMES: An online question answering system for complex clinical questions. *J Biomed Inform, 44*(2), 277-288. doi:10.1016/j.jbi.2011.01.004

Dulceanu, A., Le Dinh, T., Chang, W., Bui, T., Kim, D. S., Vu, M. C., & Kim, S. (2018). PhotoshopQuiA: A corpus of non-factoid questions and answers for why-question answering. *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Ely, J. W., Osheroff, J. A., Ebell, M. H., Bergus, G. R., Levy, B. T., Chambliss, M. L., & Evans, E. R. (1999). Analysis of questions asked by family doctors regarding patient care. *BMJ, 319*(7206), 358-361.

Higashinaka, R., & Isozaki, H. (2008). Corpus-based question answering for why-questions. *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I*.

Mrozinski, J., Whittaker, E., & Furui, S. (2008). Collecting a why-question corpus for development and evaluation of an automatic QA-system. *Proceedings of ACL-08: HLT*, 443-451.

Pampari, A., Raghavan, P., Liang, J., & Peng, J. (2018). emrQA: A large corpus for question answering on electronic medical records. *arXiv preprint arXiv:1809.00732*.

Prasad, R., & Joshi, A. (2008). A discourse-based approach to generating why-questions from texts. *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge*, 1-3.

Raghavan, P., Patwardhan, S., Liang, J. J., & Devarakonda, M. V. (2018). Annotating electronic medical records for question answering. *arXiv preprint arXiv:1805.06816*.

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Uzuner, O., South, B. R., Shen, S., & DuVall, S. L. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc, 18*(5), 552-556. doi:10.1136/amiajnl-2011-000203

Verberne, S. (2006). Developing an approach for why-question answering. *Proceedings of the 11th*

Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, 39-46.

Verberne, S., Boves, L., Oostdijk, N., & Coppen, P. (2006). Data for question answering: the case of why. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*.