Development of a Clinical Question-Answering Corpus with Realistic Multi-Answer Challenges

Sungrim Moon, PhD [1], Huan He PhD [1], and Jungwei W. Fan, PhD [1,2]

[1] Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, Minnesota, USA
[2] Kern Center for the Science of Health Care Delivery, Mayo Clinic, Rochester, Minnesota, USA

Abstract

We generated a clinical question-answering (QA) corpus that involves realistic, multi-answer cases by converting the concept-relation annotations from the i2b2 Adverse Drug Event extraction challenge dataset. A total of 91,440 QA entries were derived, representing a diverse array of 1-to-0, 1-to-1, N-to-1, and 1-to-N relations between the question and answer(s).

Introduction

Medical question-answering (QA) research using artificial intelligence (AI) is necessary

- Clinician's thought processes and decision-making often involve a cascade of questions and answers.
- Achieving human-like QA capability is highly regarded. For example, finding the answer within a given clinical document to enable patient-specific QA task.

Challenge of medical QA research

- Need the "state-of-the-art" AI research in the medical field
- QA training data should be
  - Representing the target scenario of QA training corpora
  - High annotation quality
- Currently, simplified QA task, a one-answer-one-document scheme only along with other issues observed in existing medical QA corpora[1], is common.
- Necessity of medical QA can naturally have multiple-answers-one-document

Methods

Dataset: The 2018 i2b2 Adverse Drug Event (ADE) challenge[2] dataset

- originally organized for extracting ADEs and various drug-related entities in clinical documents
- It contains a gold standard annotation of 83,869 concepts and 59,810 relations in 505 discharge summaries.

Methods:

- We focus on generating QA pairs from the subset of annotated Reason-Drug relations
- A generated QA: a question from a Drug and an answer from a Reason

Source	Example
Original text	"The patient received **morphine** for *pain* as needed
Annotation	Reason "*pain*" – Drug "**morphine**"
Generated QA	Q: "Why **morphine** was prescribed to the patient?" A: "*pain*"

- We derived N-to-1 or 1-to-N QA entries from many-to-many Reason-Drug relations.
- We derived unanswerable questions from no Reason-to-one-Drug relations.
- We injected paraphrastic questions (9 different question types) for enhancing the generalizability [3]

Results and Discussion

Our generated dataset

- a total of 91,440 QA entries, of which half (51%) are unanswerable.
- The answerable 45,162 QAs entries include 1-to-1 (24%), N-to-1 (15%) and 1-to-N (10%) QAs, and each has more than 9,000 entries.
- The results are formatted into Stanford Question Answering Dataset (SQuAD) 2.0 JSON, except that we now allow each 1-to-N QA to have a list of reasons under the answer block.

QA Category	Frequency (%)	Example	Processing method
No reason annotated (1-to-0 QA)	46,278 (51%)	**Mirtazapine** 15 mg PO QHS [*only the drug is mentioned but no reason is documented*]	Make it an unanswerable QA entry
1 reason 1 drug (1-to-1 QA)	21,906 (24%)	The patient received **morphine** for *pain* as needed	Make into a 1-to-1 QA entry
1 reason N drugs (N-to-1 QA)	14,004 (15%)	*Hypertension*: Severely elevated blood pressure. Started **amlodipine**, **metoprolol**, and **isorbide**.	Break into N separate 1-to-1 QA entries
N reasons 1 drug (1-to-N QA)	9,252 (10%)	**albuterol sulfate** 90 mcg… Puff Inhalation Q4H for *sob* or *wheeze*	List the N reasons under answer block to form a 1-to-N QA entry

Major contribution

- Generated a clinical QA corpus containing realistic, multi-answer cases by converting the concept-relation annotations from an existing AI challenge dataset.
- diverse and sizable annotations represent more realistic train/test data for developing robust clinical QA systems.
- that will transcend the conventional, limited single-answer task definition.

Access and Contact

- The dataset is available through https://portal.dbmi.hms.harvard.edu/projects/n2c2-du/ upon meeting the requirements for data access.
- A manuscript that describes the study details will be coming soon.
- You may contact Moon.Sungrim@mayo.edu or Fan.Jung-wei@mayo.edu for questions related to the study.

References

1. Yue X, Jimenez Gutierrez B, Sun H, **Clinical reading comprehension: a thorough analysis of the emrQA dataset**. *Proceedings of the ACL*. 2020:4474-86.
2. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. **The 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records**. *J Am Med Inform Assoc. 2020* Jan;27(1):3-12.
3. Moon S, Fan J. **How you ask matters: the effect of paraphrastic questions to BERT performance on a clinical SQuAD dataset**. *Proceedings of the 3rd Clinical Natural Language Processing Workshop.* 2020:111-6.