Question Generation on Clinical Texts in EMRs

Xinliang Frederick Zhang, Xiang Yue, Huan Sun
The Ohio State University

INTRODUCTION

Neural Question Generation

- Neural question generation (NQG) is defined as: given an input sentence **x,** the goal is to generate a natural question **y** related to information in the input sentence and **y** can be a sequence of an arbitrary length [1].

$$\bar{\boldsymbol{y}} = \arg\max P(\boldsymbol{y}|\boldsymbol{x})$$

- Neural question generation, as an important auxiliary task of QA (Question Answering), can provide large-scale QA pairs to help train QA systems in order to develop advanced downstream QA systems in the long run [2].

EmrQA

- A newly released clinical-specific large-scale QA dataset which is template-based and generated from existing annotated clinical notes [3].
- The EmrQA dataset contains 5 subsets: Medication, Relation, Heart Disease, Obesity and Smoking, all of which are generated from i2b2 challenge datasets.

Task Definition

- The task of question generation on clinical texts is defined as: given a patient's clinical note C and clinical sentence/evidence A from the clinical note, generate a quality and diverse clinical question set S, which might contain more than one valid question Q, by integrating domain-specific knowledge K. Mathematically put, QG on clinical texts is defined as searching for the "best set" S of Q(s) such that:

$$\bar{Q} = \arg\max P(Q|C, A, K)$$
$$S = \{\bar{Q}_i\} \quad where\ |S| \geq 1$$

Challenges

- Clinical data is messy and contains awkward acronyms (which needs expert knowledge).
- EmrQA contains lots of noises, and its template-based construction approach leads to redundancy.
- Unlike POS, NER lexical features, effectively incorporating domain-specific knowledge is difficult.
- Due to the limitation of current NQG (i.e. seq2seq-based) architecture, it is impossible to directly generate diverse questions if the evidence is fixed.

AIM

Project Overview

In this project, we aim at implementing an EMR-based Automatic Question Generation (QG) algorithm by using the EmrQA Clinical Dataset. Based on the developed QG system, we will then generate an EMR-based QA dataset.

Goals

- Aim at efficiently generating as many valid questions as possible for each clinical evidence.
- Aim at generating quality and coherent questions which can capture the semantic information encoded in the evidence.
- Ultimate goal is to generate a large-scale, quality and divisive EMR-based QA dataset to further improve the performance of downstream Clinical-QA systems.

DATASET & QA PAIRS

Dataset Overview

Table 1: Statistics of EmrQA [2]

Datasets	QA pairs	QL pairs	# Notes
i2b2 relations	1,322,789	1,008,205	425
i2b2 medications	226,128	190,169	261
i2b2 heart disease risk	49,897	35,777	119
i2b2 smoking	4,518	14	502
i2b2 obesity	354,503	336	1,118
EmrQA (total)	1,957,835	1,225,369	2,425

Table 2: Statistics of Medication and Relation datasets

	Medication	Relation
# QA pairs	222,957	904,592
# Notes	261	423
Question: avg. length	8.01	7.91
Evidence: avg. length	9.47	10.41
Note: avg. length	1062.66	889.23
Question: vocab size	1836	5142
Evidence: vocab size	5655	10381

QA pairs

Clinical Note	..... MEDICATIONS ON ADMISSION: 1. Aspirin q.d. 2. Enalapril 20 mg b.i.d. 3. Cardizem 300 mg q.d. 4 . Insulin mixed 70/30 with 60 units in the morning and 30 in the evening. 5. Atenolol 50 mg q.d. .....
Evidence	4 . insulin mixed 70/30 with 60 units in the morning and 30 in the
Questions	**what** is her current dose of insulin mixed 70/30 **what** was the dosage prescribed of insulin mixed 70/30 **has** patient ever been prescribed insulin mixed 70/30 **has** this patient ever been on insulin mixed 70/30 **has** the pt . ever been on insulin mixed 70/30 before **what** is the patient 's current dose does the patient take of her insulin mixed 70/30 **what** is the dosage of insulin mixed 70/30
NQG Question	**what** is the current dose of the 70/30
QW-NQG Question	**has** the patient had previous insulin mixed 70/30 **what** is the current dose of the patient 's insulin mixed 70/30

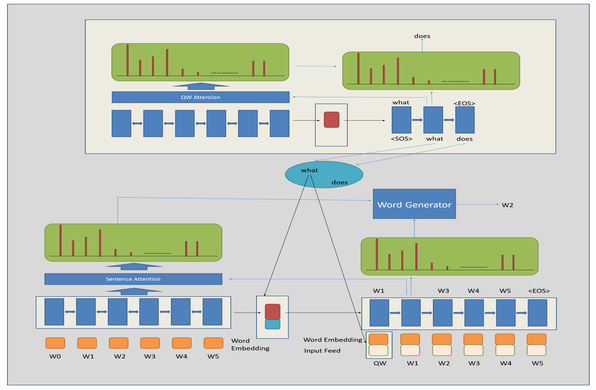
Table 3: QA pair where QW-NQG generates a QW set that exactly matches EmrQA

Clinical Note	..... OTHER TREATMENTS/PROCEDURES (NOT IN O.R.) ..... 3 . ENDO: pt was continued on home insulin regimen with coverage with insulin sliding scale. Pt was found have a TSH of 158 FT4 1.8 , FT3 56. Upon discussion w/ pt , she revealed that she had previously been diagnosed and begun on replacement but did not know what it was for and subsequently stopped it on her own. Pt started on synthroid to be f/u w/endocrine. This may be the etiology of her symptoms. .....
Evidence	3 . endo : pt was continued on home insulin regimen with coverage with
Questions	**is** there a mention of insulin regimen usage / prescription in the record **has** the patient had multiple insulin regimen prescriptions **has** there been a prior insulin regimen **has** the pt . ever been on insulin regimen before
NQG Question	**has** this patient ever been treated with insulin regimen
QW-NQG Question	**has** the patient had previous insulin regimen **was** the patient ever given medication for diabetic control

Table 4: QA pair where QW-NQG generates a unique QW which is not originally in EmrQA QW set

MODEL

Model Architecture



During generation, only one question word (QW) will be sampled from generated QW set at a time .

Question Word Module

- Motivation: For a given evidence, the most likely QW should be generated at 1st time step; then based off 1st QW and evidence, decide whether to finish the QW generation, if not, decide what is the next proper QW.
- Multi-label task: There could be more than one QW ($|S| \geq 1$) generated for an input clinical evidence.
- Benefit: Dynamically determine the best QW set (incl. size and elements) and capture the ordering relation among QWs.

RESULTS & ANALYSIS

Evaluation Results

Table 5: Automatic Evaluation Metric Results

	Bleu-1	Bleu-2	Bleu-3	Bleu-4	METEOR	ROUGE_L
NQG [1]	76.05	67.27	59.31	51.69	37.04	70.26
QW-NQG (20%)	75.97	66.73	58.37	50.67	37.43	69.58
QW-NQG (10%)	73.07	63.24	54.85	47.66	36.20	66.8
GOD-QW-NQG	80.99	71.99	64.53	58.00	42.23	75.01

Case Study

Table 6: Bleu-based QW-NQG analysis (refer to Table 3&4)

	has	has + what
Bleu 1	75.00	90.00
Bleu 2	65.47	77.46
Bleu 3	41.49	57.24
Bleu 4	0.00	0.00

	has	has + was
Bleu 1	85.71	56.25
Bleu 2	75.59	44.82
Bleu 3	61.14	32.23
Bleu 4	48.89	24.06

**Other evaluation metrics will be adopted to evaluate uniqueness/QW outside reference set (e.g. Distinct [4])*

RESULTS & ANALYSIS (Cont'd)

QW Analysis

In QW module, *Truncated Ratio* (pre-defined in training) governs how aggressive the learnt QW module will be.

Table 7: Effect of Truncate Ratio on QW Generation

Ratio	P=R	P⊂R	R⊂P	QW acc	pred: # QW	ref: # QW
20%	8.09	78.53	1.84	43.47	1.61	3.37
10%	10.64	58.73	9.33	56.19	2.27	3.37

*P: predicted/generated QW set; R: reference QW set

Tail Distribution Analysis

In EmrQA, the QW has an inherent tail distribution issue due to templates used in generating EmrQA. NQG tends to learn the safest question so it is trapped by this problem. QW-NQG (our proposed model) is more aggressive and can significantly alleviate this issue.

Table 8: Distribution of Generated QW under Different Systems

	what	when	has	was	why	how	is	did
NQG	19.70	0.08	72.10	0.00	8.12	0.00	0.00	0.00
QW-NQG*	33.38	0.21	40.33	2.75	9.30	1.55	12.24	0.24
Reference	26.89	0.66	44.89	5.34	9.42	3.58	7.46	1.77

*Number indicates the likelihood of a particular QW among QA-pairs

FUTURE PLAN

- Leverage entities in the evidence to better capture semantic information in order to generate more high-quality questions instead of merely generating "textual-related" questions.
- Effectively incorporate domain-specific/background knowledge into our proposed model so as to better utilize clinical-specific prior/expert knowledge.
- Target at submission to EMNLP 2020

REFERENCES

[1] Du, X.; Shao, J.; and Cardie, C. 2017. Learning to ask: Neural question generation for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, ACL, 1342–1352
[2] Duan, N.; Tang, D.; Chen, P.; and Zhou, M. 2017. Question generation for question answering. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, EMNLP, 866–874.
[3] Pampari, A.; Raghavan, P., Liang, J.; and Peng, J. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,* EMNLP, 2357–2368.
[4] Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, 110–119.

ACKNOWLEDGEMENTS

I would like to express my special thanks to Lumley Engineering Fund Scholarship by College of Engineering for supporting my undergraduate research work at OSU.

Ohio Supercomputer Center
An OH·TECH Consortium Member

Contact: zhang.9975@osu.edu