MULTIMODAL QUESTION ANSWERING IN THE MEDICAL DOMAIN

Dr. Asma Ben Abacha

U.S. National Library of Medicine
Lister Hill National Center for Biomedical Communications

Friday, April 24, 2020

Carnegie Mellon University
Language Technologies Institute

Carnegie Mellon University
Language Technologies Institute

ASMA BEN ABACHA

NATIONAL INSTITUTES OF HEALTH
NATIONAL LIBRARY OF MEDICINE

Dr. Asma Ben Abacha is a staff scientist at the U.S. National Institutes of Health (NIH), National Library of Medicine (NLM), Lister Hill National Center for Biomedical Communications. Prior to joining the NLM in 2015, she was a researcher at the Luxembourg Institute of Science and Technology and lecturer at the University of Lorraine, France. Dr. Ben Abacha received a Ph.D. in computer science from Paris 11 University, France, a research master's degree from Paris 13 University, and a software engineering degree from the National School of Computer Sciences (ENSI), Tunisia. She is currently working on medical question answering, visual question answering, and NLP-related projects in the medical domain.

Multimodal Question Answering in the Medical Domain

Artificial intelligence (AI) is playing an increasingly important role in our access to information. However, a one-fits-all approach is suboptimal, especially in the medical domain where health-related information is more sensitive due to its potential impact on public health, and where domain-specific aspects such as technical language and case or context-based interpretation have to be taken into account. Bridging the gap between several research areas such as AI, NLP, medical informatics, and computer vision is a promising way to achieve reliable and efficient access to medical information. In this talk, I will discuss some of my recent projects on multimodal Question Answering (QA) including NLP methods for textual QA and Visual Question Answering (VQA). In particular, I'll present the lessons learned from working on QA from trusted answer sources and alternative NLP approaches such as recognizing question entailment and question summarization. In a second part, I'll address the task of VQA from radiology images and potential solutions to support the creation of large-scale training data through visual question generation. Throughout the talk, I'll present our recent efforts in creating relevant datasets and new approaches as well as the challenges that we organized to promote research in multimodal question answering.

FRIDAY, APRIL 24, 2:30 - 4 PM
CONDUCTED VIA ZOOM
HTTPS://CMU.ZOOM.US/J/208848796

Abstract

Artificial intelligence (AI) is playing an increasingly important role in our access to information. However, a one-fits-all approach is suboptimal, especially in the medical domain where health-related information is more sensitive due to its potential impact on public health, and where domain-specific aspects such as technical language and case or context-based interpretation have to be taken into account. Bridging the gap between several research areas such as AI, NLP, medical informatics, and computer vision is a promising way to achieve reliable and efficient access to medical information.

In this talk, I will discuss some of my recent projects on multimodal question answering (QA) including NLP methods for textual QA and visual question answering (VQA). In particular, I will present the lessons learned from working on QA from trusted answer sources and alternative NLP approaches such as recognizing question entailment and question summarization. In a second part, I will address the task of VQA from radiology images and potential solutions to support the creation of large-scale training data through visual question generation. Throughout the talk, I will present our recent efforts in creating relevant datasets and new approaches as well as the challenges that we organized to promote research in multimodal question answering.

Disclaimer

The views and opinions expressed do not necessarily state or reflect those of the U.S. Government, and they may not be used for advertising or product endorsement purposes.

Plan

Introduction

Question Answering from Trusted Sources

1. Answering Questions about Medications
2. Summarization of Consumer Health Questions
3. Recognizing Question Entailment

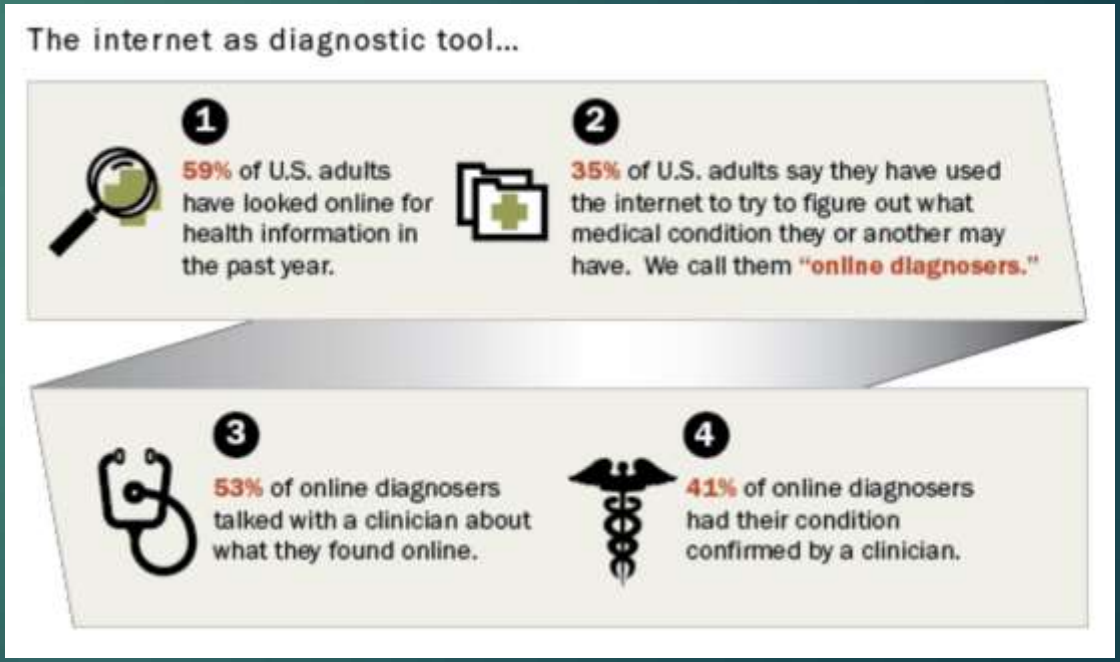
VQA from Radiology Images

1. Visual Question Answering (VQA)
2. Visual Question Generation (VQG)

Discussion

INTRODUCTION

Information Retrieval & Question Answering (QA) in the Medical Domain



https://www.pewresearch.org/internet/2013/01/15/health-online-2013/

INFORMATION SOURCES (1/2)

Posted by u/Drestos 2 months ago

reddit

0

How to prevent coronavirus ?

question

So I'm going to South Korea on February 2 with my family but I heard there's cases of the virus in South Korea and going to the airport is a bad idea since there's people with virus there but is there anyways to prevent myself and my family from getting the virus . (Sorry for the bad grammars)

11 Comments Share Save Hide Report 25% Upvoted

idkwhyimonredditlol 0 points · 2 months ago
Do you have a mask ? If so what

Drestos 1 point · 2 months ago
I'm wearing two mask when I go to the airport

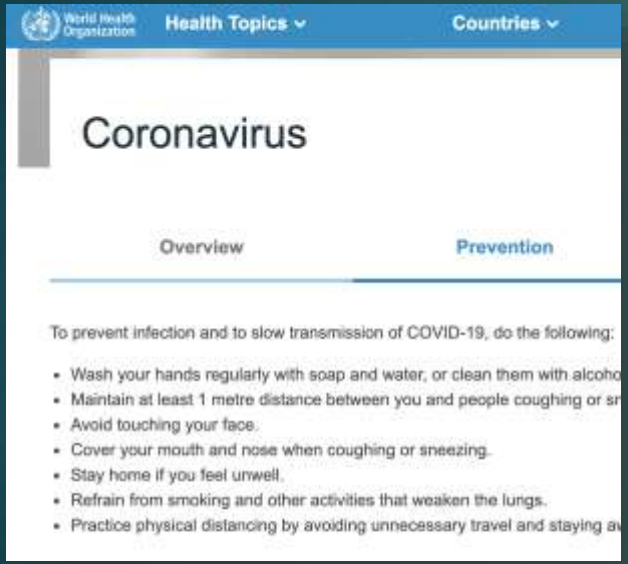
idkwhyimonredditlol 1 point · 2 months ago
What kind of mask

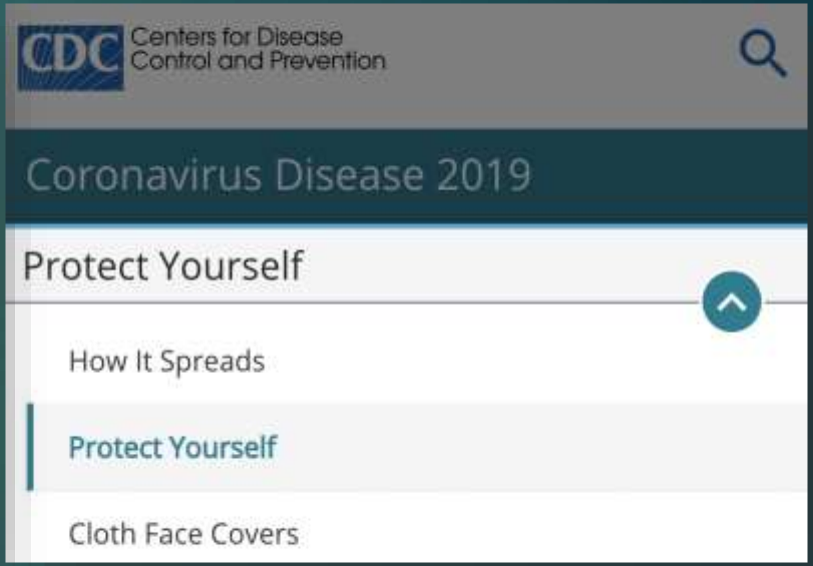
Drestos 1 point · 2 months ago
The first layer is normal the second layer is the cloth ones

idkwhyimonredditlol 1 point · 2 months ago
Well they're saying surgical masks are ineffective and the n95 masks with the little filters work

Drestos 1 point · 2 months ago
That mask is just too big

INFORMATION SOURCES (2/2)





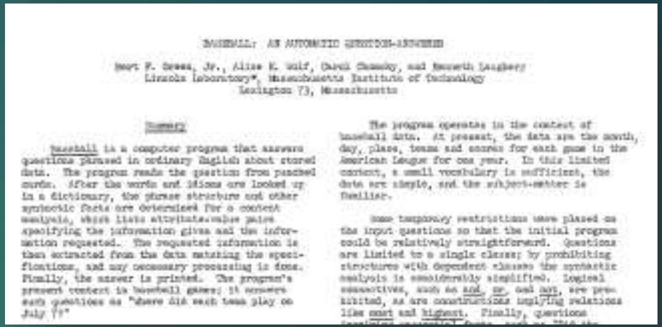
Part I:

QA from Trusted Answer Sources

- Problem
- Methods & Results
- Lessons Learned

Early History of Question Answering

BASEBALL (Green et al.,1961): built to answer questions about American baseball games.



Summary

Baseball is a computer program that answers questions phrased in ordinary English about stored data. The program reads the question from punched cards. After the words and idioms are looked up in a dictionary, the phrase structure and other syntactic facts are determined for a content analysis, which lists attribute-value pairs specifying the information given and the information requested. The requested information is then extracted from the data matching the specifications, and any necessary processing is done. Finally, the answer is printed. The program's present context is baseball games; it answers such questions as "Where did each team play on July 7?"

The program operates in the context of baseball data. At present, the data are the month, day, place, teams and scores for each game in the American League for one year. In this limited context, a small vocabulary is sufficient, the data are simple, and the subject-matter is familiar.

Some temporary restrictions were placed on the input questions so that the initial program could be relatively straightforward. Questions are limited to a single clause; by prohibiting structures with dependent clauses the syntactic analysis is considerably simplified. Logical connectives, such as and, or, and not, are prohibited, as are constructions implying relations like most and highest. Finally, questions

LUNAR (Woods, 1973): built to answer scientists' questions about the Apollo 11 moon rocks.

QA @ TREC-8 (Voorhees, 1999).

QA @ CLEF Starting from 2003.

Early History of Medical QA

- Pierre Zweigenbaum. **Question Answering in Biomedicine**. Workshop on NLP for QA, EACL 2003,

- Dina Demner-Fushman & Jimmy Lin. **Answer Extraction, Semantic Clustering, and Extractive Summarization for Clinical Question Answering**. COLING/ACL 2006.

- **BioASQ** challenges on biomedical semantic indexing and question answering (started in 2012).
- **CLEF QA4MRE** Alzheimer's task (2012).

A summary of multimodal QA datasets and systems for the medical domain:
https://github.com/abachaa/Existing-Medical-QA-Datasets

1) Answering Questions about Medications

Medical Question Answering about Medications

- Manual creation of the first gold standard corpus of consumer health questions about medications with associated **answers** and **annotations**.

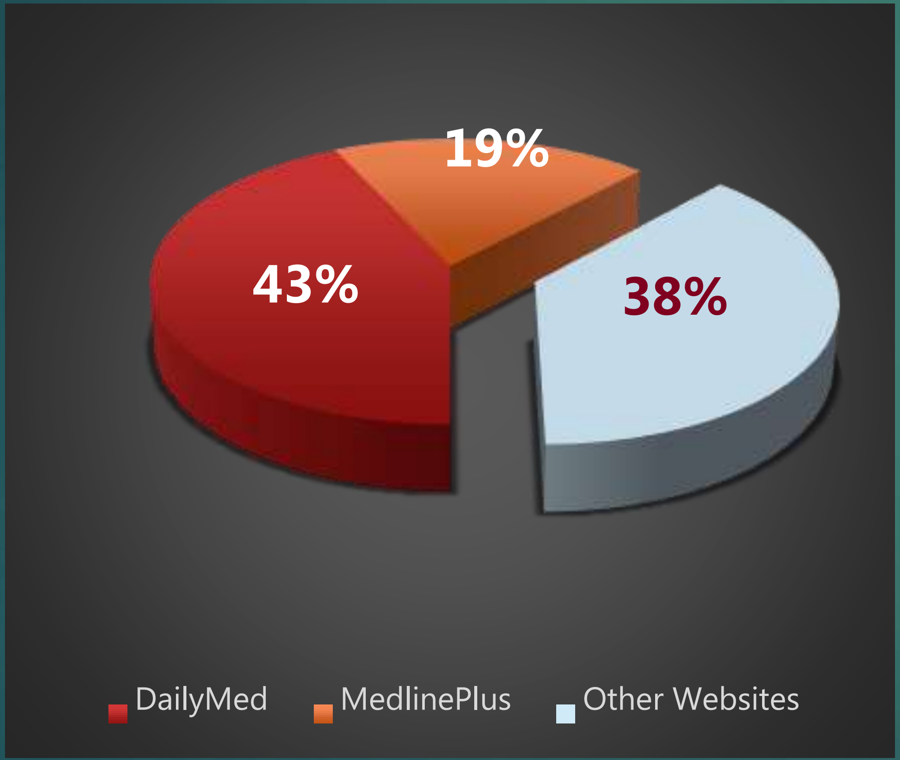
- Deep learning experiments on:
  - **question type identification** (CNN),
  - **focus recognition** (Bi-LSTM-CRF),
  - **question answering** (CHiQA).

"Bridging the Gap between Consumers' Medication Questions and Trusted Answers". Ben Abacha et al. MEDINFO 2019

- 1) <u>Questions:</u> The corpus contains **674** consumer questions submitted to MedlinePlus.

Question Type	#	Example
Information	112	what type of drug is **amphetamine**?
Dose	70	what is a daily amount of **prednisolone eye drops** to take?
Usage	61	how to self inject **enoxaparin sodium**?
Side Effects	60	does **benazepril** aggravate hepatitis?
Indication	55	why is **pyridostigmine** prescribed?
Interaction	51	can i drink **cataflam** when i drink **medrol**?

- **2) <u>Answer Sources</u>:** 43% of the questions were answered from **DailyMed** and 19% from **MedlinePlus**.



- Other websites were used to answer **38%** of the questions (e.g. cdc.gov, mayoclinic.org, PubMed).
- External resources (e.g. eHealthMe) were needed for specific types of questions such as Interactions.

Results

Bi-LSTM-CRF Results for Focus Recognition:

Results (%)	F1	P	R
Exact entity match	74.07	78.12	70.42
Partial entity match	90.37	95.31	85.92

CNN Results for Question Type Identification:

Accuracy	75.7%

Results

Answer Retrieval:

- CHiQA found the **correct** answer in the top four results in **35%** of the cases, only **related** answers for **35%** of them and **irrelevant** answers for the remaining **30%**.
- Our independent observations also hint that classical QA systems may not be the best fit for **medication** questions.

Several Challenges towards Automatic QA

Question Analysis:

- Linguistic ambiguity due to misspellings, wrong grammar, etc.
- The use of too many words or general terms to express a medical term.
- The requested medical information was not formalized or written online.

Answer retrieval:

- Conditional answers (e.g. depending on the manufacturer or on the disease).
- Distributed answers formed only by combining different text snippets from different sources.
- Answers that could not be found without expert inference/knowledge.

Solutions for the Automation of Medication QA

- Medical text translation, simplification, and inference are often needed to find relevant answers and/or make the answers readable for non-expert users.

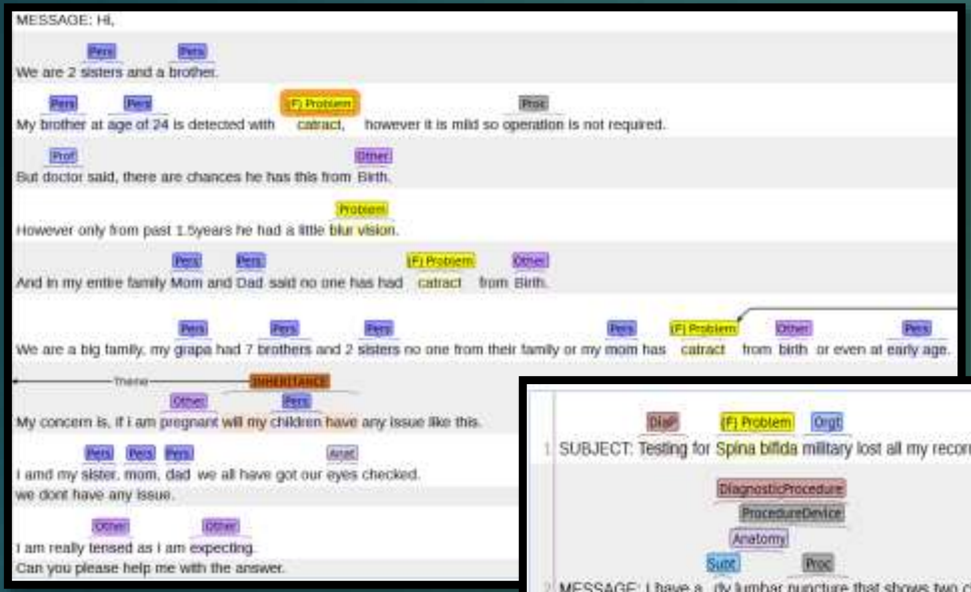
- More data and resources are needed to cover information about drug interactions and usage guidelines (e.g. scientific literature).

- Conditional answers require different solutions such as providing a list of answers or interacting with the user in a dialogue-based approach.

"Bridging the Gap between Consumers' Medication Questions and Trusted Answers". Ben Abacha et al. MEDINFO 2019

2) Summarization of consumer health questions

Consumer Health Questions



Question Summarization

Question: polymicrogyria. My 16 month old son has this. Does not sit up our crawl yet but still trying and is improving in grabbing things etc etc. Have read about other cases that seem 10000 time worse. It's it possible for this post of his brain to grown to normal and he grow out of it?

⇨ **Summary:** What is the **prognosis** for **polymicrogyria**?

QA Results w/(o) Manual Question Summarization

Measures	Original* Questions	Question Summaries
AvgScore (0-3)	**0.711**	**1.125**
Succ@2+	0.442	0.663
Succ@3+	0.192	0.317
Succ@4+	0.077	0.144
Prec@2+	0.46	0.663
Prec@3+	0.2	0.317
Prec@4+	0.08	0.144

- **Summarizing the questions leads to a substantial improvement.**

* Benchmark from the Medical Question Answering Competition @ LiveQA 2017 (Ben Abacha et al., TREC 2017).

"On the Role of Question Summarization and Information Source Restriction in Consumer Health Question Answering".
Ben Abacha & Demner-Fushman. AMIA 2019 Informatics Summit.

Automatic Question Summarization: New Datasets

Method	Type	Example
#1 MeQSum Dataset	Consumer Health Question	I suffered a massive stroke on [DATE] with paralysis on my left side of my body, I'm home and conduct searches on the internet to find help with recovery, and always this product called neuroaid appears claiming to restore function. to my knowledge it isn't approved by the FDA, but it sounds so promising. do you know anything about it and id there anything approved by our FDA, that does help?
	Summary	What are treatments for stroke paralysis, including neuroaid?
#2 Augmentation with Clinical Data	Clinical Question	55-year-old woman. This lady has epigastric pain and gallbladder symptoms. How do you assess her gallbladder function when you don't see stones on the ultrasound? Can a nonfunctioning gallbladder cause symptoms or do you only get symptoms if you have stones?
	Summary	Can a nonfunctioning gallbladder cause symptoms or do you only get symptoms if you have stones?
#3 Augmentation with Semantic Selection	Medical Question	Is it healthy to ingest 500 mg of vitamin c a day? Should I be taking more or less?
	Summary	How much vitamin C should I take a day?

"On the Summarization of Consumer Health Questions". Ben Abacha & Demner-Fushman. ACL 2019.

Automatic Question Summarization: Results

Method	Training Set	ROUGE-1	ROUGE-2	ROUGE-L
Seq2seq Attentional Model	#1	24.80	13.84	24.27
	#2	28.97	18.34	28.74
	#3	27.62	15.70	27.11
Pointer Generator (PG) *See et al., ACL 2017*	#1	35.80	20.19	34.79
	#2	42.77	25.00	40.97
	#3	**44.16**	**27.64**	**42.78**
PG + Coverage	#1	39.57	23.05	38.45
	#2	40.00	24.13	38.56
	#3	41.76	24.80	40.50

- ***#1** MeQSum Dataset*
- ***#2** Augmentation with Clinical Data*
- ***#3** Augmentation with Semantic Selection*

➢ **Data augmentation from question datasets improves the overall performance.**

Automatic Question Summarization: Results

Currently displaying: attn_vis_data.json

Question

dvt . can a birth control called ocella cause dvt ? my daughter experiences pains cramping , redness and swelling in her thigh and also really bad huge blood clots during her menstrual cycles after she was prescribed osella for birth control . also these __syntoms__ worsened after she gave birth . this has been happening for a year now should she see discuss this with her doctor right away ?

Reference summary

can birth control drug ocella cause deep vein thrombosis .

Generated summary (highlighted = high generation probability)

can ocella cause dvt ?

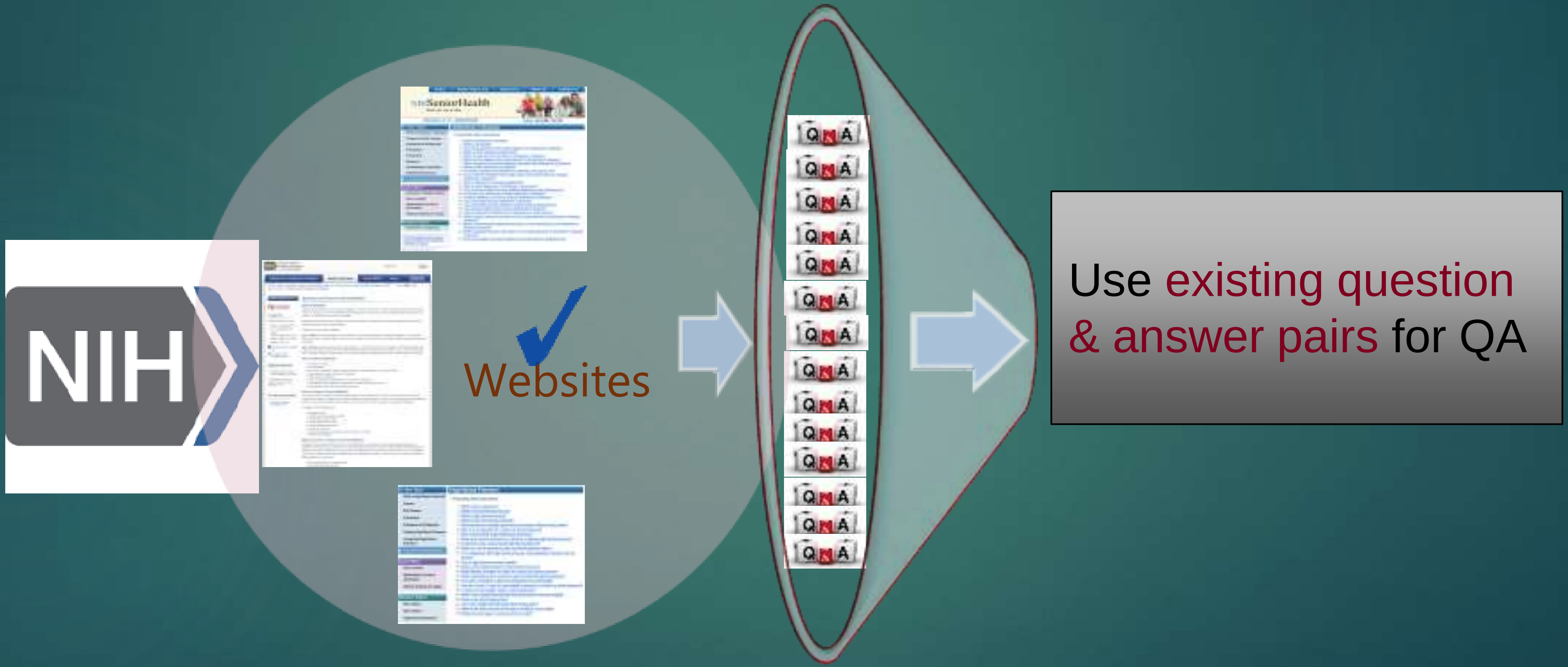
- **The best performance of 44.16% is comparable to the state-of-the-art results in open-domain using 8K training pairs (2.5% of the size of the CNN-DailyMail dataset).**
- **Promising results considering the low-frequency nature of most medical entities.**
- **Importance of data selection and augmentation.**

"On the Summarization of Consumer Health Questions". Ben Abacha & Demner-Fushman. ACL 2019.

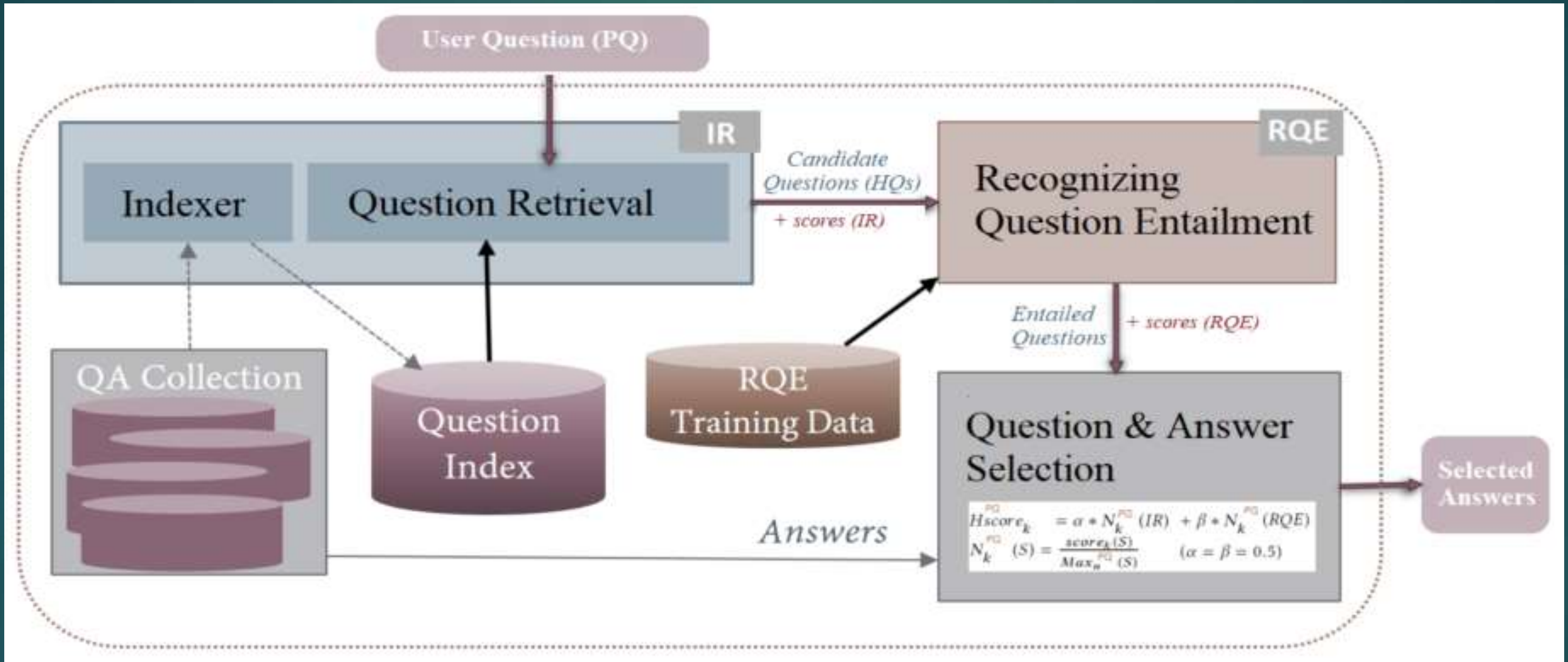
3) Recognizing Question Entailment

Our Goal

- **Answering new questions by retrieving entailed questions with existing answers.**

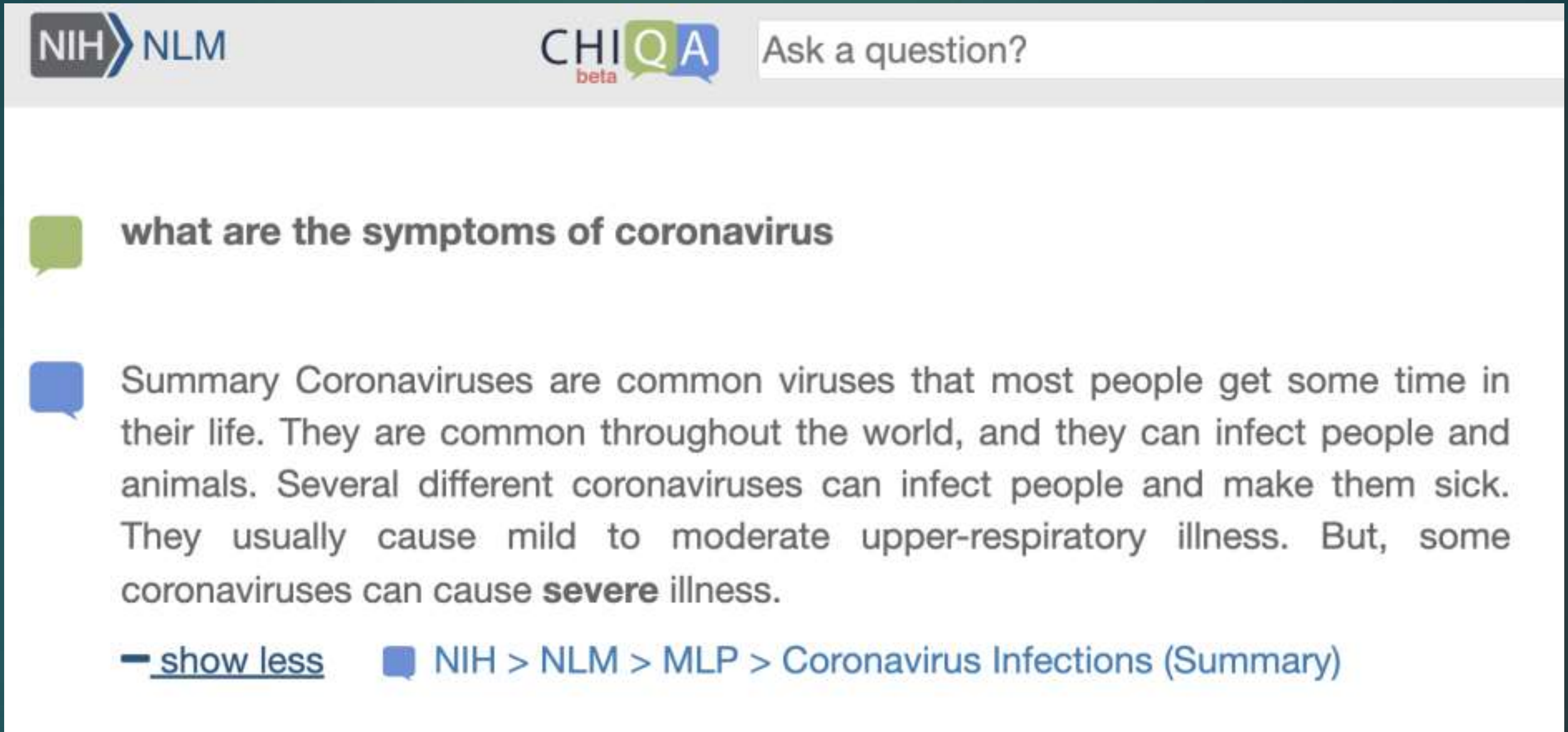


Recognizing Question Entailment (RQE)



- "A Question-Entailment Approach to Question Answering". Ben Abacha & Demner-Fushman. BMC Bioinformatics 2019.

The RQE-based QA system is a component of the CHiQA system



https://chiqa.nlm.nih.gov

The RQE-based QA System: Data

MedQuAD dataset of 47k QA pairs:

- Created from **trusted** resources (**12 NIH websites).**
- Covers 36 question types about **diseases** and **drugs.**
- https://github.com/abachaa/MedQuAD

The RQE-based QA System: Results

Metrics	RQE-based QA System	LiveQA-Med Best	LiveQA-Med Median
Average Score	**0.827**	0.637	0.431
MAP@10	0.311	--	--
MRR@10	0.333	--	--

- **Approach:** Relying on the retrieval of entailed questions is a viable strategy to medical QA.
- **Data:** Limiting the number of answer sources could enhance the QA performance.

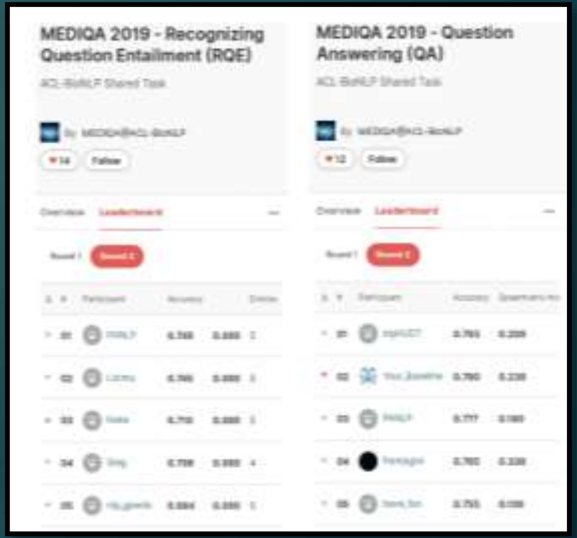
- "A Question-Entailment Approach to Question Answering". Ben Abacha & Demner-Fushman. BMC Bioinformatics 2019.

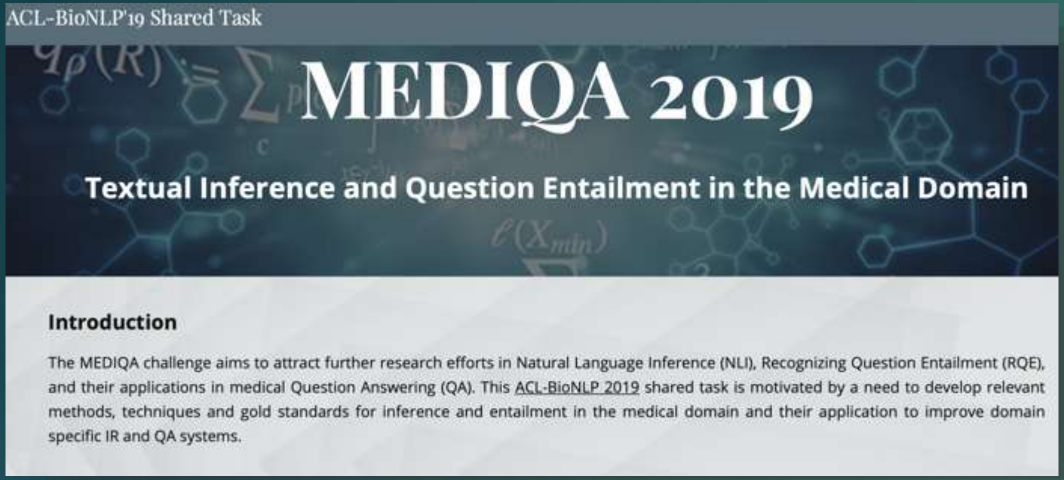
RQE Models

Methods	Training Datasets			
	SNLI	MultiNLI	Quora	Clinical-QE
Neural Network (NN)	48.94	54.59	52.35	48.71
NN + GloVe embeddings	49.41	54.82	52.82	57.18
Logistic Regression + Features	67.05	64.94	52.11	**73.18**

- Logistic regression outperformed neural networks when trained with traditional word embeddings such as glove and word2vec.

- Relying on recent language models such as Bert for pre-training led to a better performance (***MEDIQA-RQE 2019***).





Three tasks: NLI, RQE & QA
72 participating teams
20 published papers

Team	Task(s)
ANU-CSIRO (Nguyen et al., 2019)	NLI, RQE, QA
ARS_NITK (Agrawal et al., 2019)	NLI, RQE, QA
DoubleTransfer (Xu et al., 2019)	NLI, RQE, QA
Dr.Quad (Bannihatti Kumar et al., 2019)	NLI, RQE, QA
DUT-BIM (Zhou et al., 2019a)	QA
DUT-NLP (Zhou et al., 2019b)	RQE, QA
IITP (Bandyopadhyay et al., 2019)	NLI, RQE, QA
IIT-KGP (Sharma and Roychowdhury, 2019)	RQE
KU_ai (Cengiz et al., 2019)	NLI
lasigeBioTM (Lamurias and Couto, 2019)	NLI, RQE, QA
MSIT_SRIB (Chopra et al., 2019)	NLI
NCUEE (Lee et al., 2019b)	NLI
PANLP (Zhu et al., 2019)	NLI, RQE, QA
Pentagon (Pugaliya et al., 2019)	NLI, RQE, QA
Saama Research (Kanakarajan, 2019)	NLI
Sieg (Bhaskar et al., 2019)	NLI, RQE
Surf (Nam et al., 2019)	NLI
UU_TAILS (Tawfik and Spruit, 2019)	NLI, RQE
UW-BHI (Kearns et al., 2019)	NLI
WTMED (Wu et al., 2019)	NLI

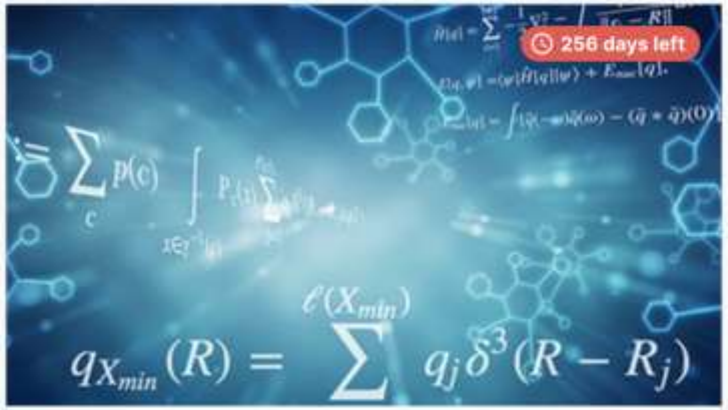
"Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering". Ben Abacha, Shivade & Demner-Fushman. ACL-BioNLP 2019.

MEDIQA – **Post** Challenge Round

AIcrowd

https://www.aicrowd.com/organizers/mediqa-acl-bionlp

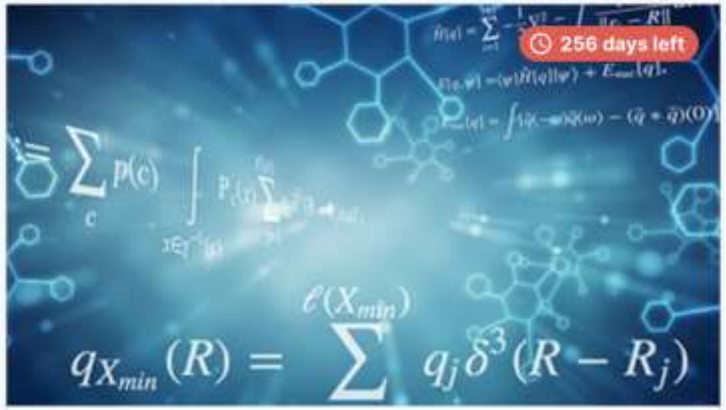
Submission open



MEDIQA 2019 - Natural Language Inference (NLI)

ACL-BioNLP Shared Task

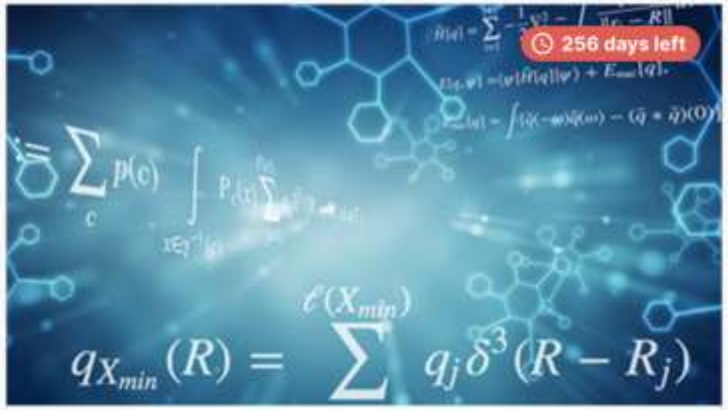
4755 120 171



MEDIQA 2019 - Recognizing Question Entailment (RQE)

ACL-BioNLP Shared Task

4434 142 290



MEDIQA 2019 - Question Answering (QA)

ACL-BioNLP Shared Task

4045 103 114

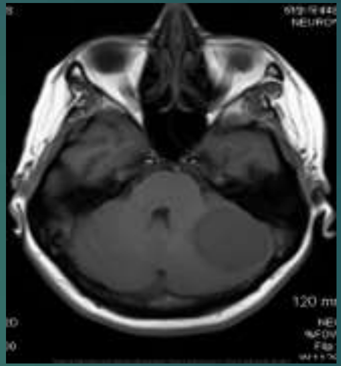
To Sum up:

- Relying on the retrieval of **entailed questions** is a viable strategy to answer consumer health questions
- **Limiting** the number of answer sources by using only **trusted sources** could enhance the QA performance.
- **Transfer learning & multi-task learning** improve the performance of RQE, NLI and QA models in the medical domain.
- **Summarizing the questions** leads to a substantial improvement.
- More efforts are needed for building large **datasets** (for training and testing) and efficient **methods** (e.g. medical text understanding and simplification).

Part II:

VQA & VQG from Radiology Images

Visual Question Answering in the Medical Domain

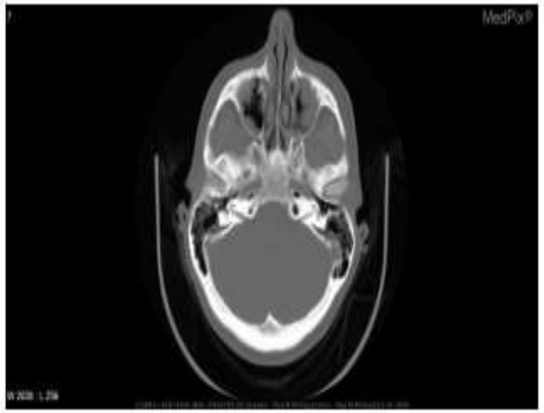


- VQA poses a challenging problem involving NLP and Computer Vision.
- Automatic understanding of **radiology images** and answering related questions could support clinical education, clinical decision making, and patient education.

VQA-Med @ ImageCLEF 2019: Data

- Four categories of questions: **Modality**, **Plane**, **Abnormality** & **Organ**
- Training, validation, and test sets created automatically:
  - Training set: 3,200 radiology images and 12,792 question-answer pairs.

(1) **Q**: is this a contrast or non-contrast ct? **A**: contrast



(2) **Q**: which plane is the image shown in? **A**: axial

(3) **Q**: what is abnormal in the gastrointestinal image? **A**: gastric volvulus (organoaxial)

(4) **Q**: which organ system is shown in the ct scan? **A**: lung, mediastinum, pleura

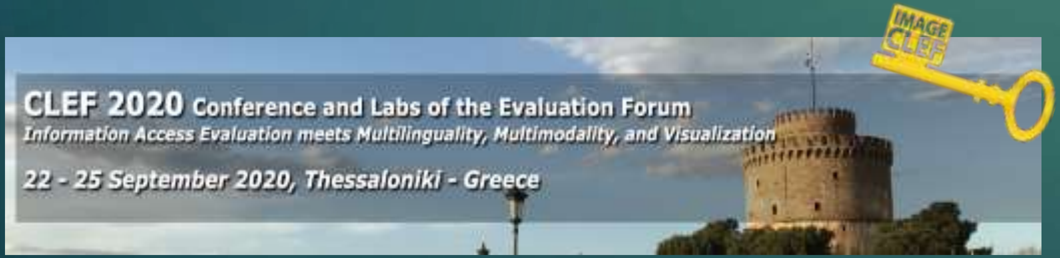
VQA-Med @ ImageCLEF 2019: Results

Team	Run ID	Modality	Plane	Organ	Abnormality	Overall
Hanlin	26889	0.202	0.192	0.184	0.046	0.624
yan	26853	0.202	0.192	0.184	0.042	0.620
minhvu	26881	0.210	0.194	0.190	0.022	0.616
TUA1	26822	0.186	0.204	0.198	0.018	0.606
UMMS	27306	0.168	0.190	0.184	0.024	0.566
AIOZ	26873	0.182	0.180	0.182	0.020	0.564
IBM Research AI	27199	0.160	0.196	0.192	0.010	0.558
LIST	26908	0.180	0.184	0.178	0.014	0.556
Turner.JCE	26913	0.164	0.176	0.182	0.014	0.536
JUST19	27142	0.160	0.182	0.176	0.016	0.534
Team_Pwc_Med	26941	0.148	0.150	0.168	0.022	0.488
Techno	27079	0.082	0.184	0.170	0.026	0.462
deepak.gupta651	27232	0.096	0.140	0.124	0.006	0.366
ChandanReddy	26884	0.094	0.126	0.064	0.010	0.294
Dear stranger	26895	0.062	0.140	0	0.008	0.210
abhishekthanki	27307	0.122	0	0.028	0.010	0.160
IITISM@CLEF	26905	0.052	0.004	0.026	0.006	0.088

- The best team achieved 0.624 accuracy and 0.644 BLEU score.
- Methods: transfer learning, multi-task learning, ensemble methods, and hybrid approaches combining classification models and answer generation methods.

“VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019”. Ben Abacha et al. CLEF 2019.

VQA-Med @ ImageCLEF 2020



Two Tasks:

1. Visual Question Answering (VQA)
2. Visual Question Generation (VQG)

- **Datasets:**
  - VQA training set: 4,000 images with 4,000 QA pairs.
  - VQG training set: 780 images with 2,156 questions.

- Submission open on April 22, 2020.
- **Run submission deadline** extended to: **June 5, 2020.**

Visual Question Generation (VQG)


nature > scientific data > data descriptors > article

SCIENTIFIC DATA

Data Descriptor | OPEN | Published: 20 November 2018

A dataset of clinically generated visual questions and answers about radiology images

Jason J. Lau, Soumya Gayen, Asma Ben Abacha & Dina Demner-Fushman

Scientific Data 5, Article number: 180251 (2018) | Download Citation

VQA-RAD dataset

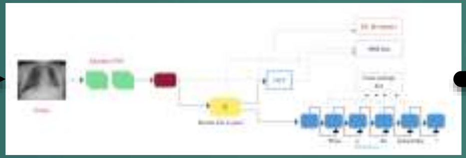
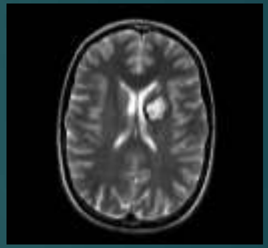


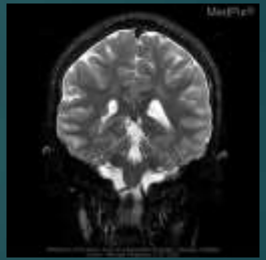
Image	Generated questions vs. ground truth
	what type of mri is used to acquire this image ? mri imaging modality used for this image?
	what is seen in the lung apices ? what abnormalities are in the lung apices ?
	is a ring enhancing lesion present in the right lobe of the liver? is the liver normal ?

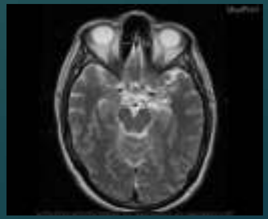
"Visual Question Generation from Radiology Images".
Sarrouti, Ben Abacha & Demner-Fushman. ACL-ALVR 2020.

Neuroimaging Collection (Ongoing)

- **11,500** radiology images of the brain and the head.
- **Several Tasks:** Classification, VQA, VQG, and caption generation.
- Annotators: neuroradiologists and radiologists. **Contact me if you are interested to participate.**







AI in the Medical Domain

Evaluation

- We cannot advance research without "**good" gold standard corpora** and efficient **evaluation metrics**.

Interdisciplinarity

- We need more collaborations and exchanges between **medical experts, AI, NLP & CV scientists**.

Ethics

- Data & Applications

Thank you for your Attention!

Email: asma.benabacha@nih.gov

Twitter: @AsmaBenAbacha

GitHub: abachaa

References

1. Asma Ben Abacha & Dina Demner-Fushman. On the Summarization of Consumer Health Questions. ACL 2019.
2. Asma Ben Abacha, Chaitanya Shivade & Dina Demner-Fushman. Overview of the MEDIQA 2019 Shared Task on Textual Inference, Question Entailment and Question Answering. ACL-BioNLP 2019.
3. Asma Ben Abacha & Dina Demner-Fushman. On the Role of Question Summarization and Information Source Restriction in Consumer Health Question Answering. AMIA 2019 Informatics Summit.
4. Asma Ben Abacha, Eugene Agichtein, Yuval Pinter & Dina Demner-Fushman. Overview of the Medical QA Task @ TREC 2017 LiveQA Track. TREC 2017.
5. Asma Ben Abacha & Dina Demner-Fushman. Recognizing Question Entailment for Medical Question Answering. AMIA 2016.
6. Asma Ben Abacha & Dina Demner-Fushman. A Question-Entailment Approach to Question Answering. BMC Bioinformatics 2019.
7. Asma Ben Abacha, Yassine Mrabet, Mark Sharp, Travis Goodwin, Sonya E. Shooshan & Dina Demner-Fushman. Bridging the Gap between Consumers' Medication Questions and Trusted Answers. MEDINFO 2019.
8. Jason J. Lau, Soumya Gayen, Asma Ben Abacha & Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. Scientific Data, Nature, 2018.
9. Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman & Henning Müller. VQA-Med: Overview of the Medical Visual Question Answering Task at ImageCLEF 2019. CLEF 2019.
10. Visual Question Generation from Radiology Images. Mourad Sarrouti, Asma Ben Abacha & Dina Demner-Fushman. ACL-ALVR 2020.