MedMCQA : A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering

Ankit Pal, Logesh Kumar Umapathi and Malaikannan Sankarasubbu
Saama AI Research Chennai, India

Abstract

- This paper introduces MedMCQA, a new large-scale, Multiple-Choice Question Answering (MCQA) dataset designed to address realworld medical entrance exam questions.
- More than 194k high-quality AIIMS & NEET PG entrance exam MCQs covering 2.4k healthcare topics and 21 medical subjects are collected with an average token length of 12.77 and high topical diversity.
- Each sample contains a question, correct answer(s), and other options which requires a deeper language understanding as it tests the 10+ reasoning abilities of a model across a wide range of medical subjects & topics.
- `medmcqa.github.io`

Contributions

In brief, the contributions of this study are as follows.

- **Diversity and difficulty** 2.4k healthcare topics and 21 medical subjects with an average token length of 12.77
- **Quality** Detailed statistics, analysis of the data, and fine-grained evaluation per medical subject are provided
- **Evaluation of quality** Extensive experiments are conducted using high-performance pretrained medical domain models.
- **Reproducible exam-based split** The dataset is split based on the exams instead of a question based split

Dataset collection

Sources of the dataset are from

- Mocktests and online test series
- AIIMS & NEET PG exam questions (1991-present)

Dataset Sample

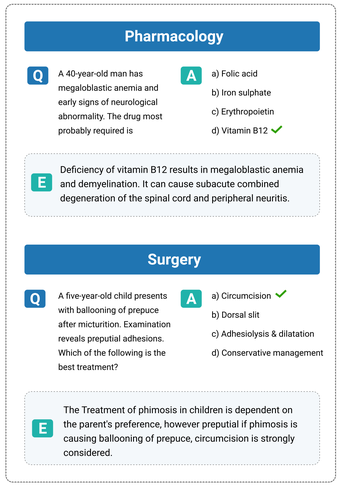

Figure: Samples from the MedMCQA dataset, along with the answer's explanation. (✓ : the correct answer)

Data statistics

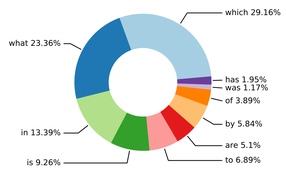

Figure: Relative sizes of Question Types in MedMCQA

	Train	Test	Dev	Total
Question #	182,822	6,150	4,183	193,155
Vocab	94,231	11,218	10,800	97,694
Max Q tokens	220	135	88	220
Max A tokens	38	21	25	38
Max E tokens	3,155	651	695	3,155
Avg Q tokens	12.77	9.93	14.09	12.71
Avg A tokens	2.69	2.58	3.19	2.70
Avg E tokens	67.52	46.54	38.44	66.22

Table 1: MedMCQA dataset statistics, where Q, A, E represents the Question, Answer, and Explanation, respectively

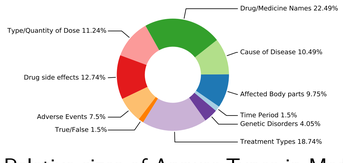

Figure: Relative sizes of Answer Types in MedMCQA

Experiments

Subject Name	Test	Dev
Anaesthesia	0.47	0.26
Anatomy	0.40	0.39
Biochemistry	0.48	0.49
Dental	0.43	0.36
ENT	0.47	0.52
FM	0.48	0.35
O&G	0.54	0.39
Medicine	0.49	0.47
Microbiology	0.50	0.44
Ophthalmology	0.60	0.51
Orthopaedics	-	0.33
Pathology	0.53	0.46
Pediatrics	0.39	0.45
Pharmacology	0.46	0.46
Physiology	0.47	0.47
Psychiatry	**0.67**	**0.56**
Radiology	0.42	0.31
Skin	0.50	0.29
PSM	0.44	0.35
Surgery	0.50	0.43
Unknown	0.44	1.0

Table 2: Fine-grained evaluation per medical subject in test and dev set

	w/o Context		Wiki		PubMed	
Model	**Test**	**Dev**	**Test**	**Dev**	**Test**	**Dev**
Bert$_{Base}$	0.33	0.35	0.33	0.35	0.37	0.35
BioBert	0.37	0.38	0.39	0.37	0.42	0.39
SciBert	0.39	0.39	0.38	0.39	0.43	0.41
PubMedBERT	**0.41**	**0.40**	**0.42**	**0.41**	**0.47**	**0.43**

Table 3: Performance of all baseline models in accuracy (%) on MedMCQA test-dev set

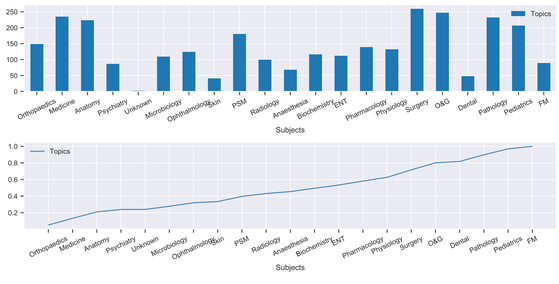

Figure: Distribution of topics per subject & Cumulative Frequency Graph for MedMCQA dataset.

Conclusion

1. In this work, MedMCQA, a new large-scale, Multi-Choice Question Answering (MCQA) dataset, is presented, which requires a deeper domain and language understanding as it tests the 10+ reasoning abilities of a model across a wide range of medical subjects & topics.
2. It is demonstrated that the dataset is challenging for the current state-of-the-art methods and domain-specific models, with the best baseline achieving only 47
3. It is expected that this dataset would facilitate future research in this direction.

Related Work

- `https://arxiv.org/abs/2009.13081` MedQA
- `https://aclanthology.org/P19-1092/` HEAD-QA

Full Paper & Contact info

Contact Information

- twitter: @aadityaura
- aadityaura@gmail.com