Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain

Asma Ben Abacha
NLM/NIH
benabachaa@nih.gov

Yassine Mrabet
NLM/NIH
mrabety@mail.nih.gov

Yuhao Zhang
Stanford University
zyh@stanford.edu

Chaitanya Shivade
Amazon
shivadc@amazon.com

Curtis Langlotz
Stanford University
langlotz@stanford.edu

Dina Demner-Fushman
NLM/NIH
ddemner@mail.nih.gov

Abstract

The MEDIQA 2021 shared tasks at the BioNLP 2021 workshop addressed three tasks on summarization for medical text: (i) a question summarization task aimed at exploring new approaches to understanding complex real-world consumer health queries, (ii) a multi-answer summarization task that targeted aggregation of multiple relevant answers to a biomedical question into one concise and relevant answer, and (iii) a radiology report summarization task addressing the development of clinically relevant impressions from radiology report findings. Thirty-five teams participated in these shared tasks with sixteen working notes submitted (fifteen accepted) describing a wide variety of models developed and tested on the shared and external datasets. In this paper, we describe the tasks, the datasets, the models and techniques developed by various teams, the results of the evaluation, and a study of correlations among various summarization evaluation measures. We hope that these shared tasks will bring new research and insights in biomedical text summarization and evaluation.

1 Introduction

Text summarization aims to create natural language summaries that represent the most important information in a given text. Extractive summarization approaches tackle the task by selecting content from the original text without any modification (Nallapati et al., 2017; Xiao and Carenini, 2019; Zhong et al., 2020), while abstractive approaches extend the summaries' vocabulary to out-of-text words (Rush et al., 2015; Gehrmann et al., 2018; Chen and Bansal, 2018).

Several past challenges and shared tasks have focused on summarization. The Document Understanding Conference[1] (DUC) organized seven challenges from 2000 to 2007 and the Text Analysis Conference[2] (TAC) ran four shared tasks (2008-2011) on news summarization. The last TAC 2014 summarization task tackled biomedical article summarization with referring sentences from external citations. Recent efforts in summarization have focused on neural methods (See et al., 2017; Gehrmann et al., 2018) using benchmark datasets compiled from news articles, such as the CNN-DailyMail dataset (CNN-DM) (Hermann et al., 2015). However, despite its importance, fewer efforts have tackled text summarization in the biomedical domain for both consumer and clinical text and its applications in Question Answering (QA) (Afantenos et al., 2005; Mishra et al., 2014; Afzal et al., 2020).

While the 2019 BioNLP-MEDIQA[3] edition focused on question entailment and textual inference and their applications in medical Question Answering (Ben Abacha et al., 2019), MEDIQA 2021[4] addresses the gap in medical text summarization by promoting research on summarization for consumer health QA and clinical text. Three shared tasks are proposed for the summarization of (i) consumer health questions, (ii) multiple answers extracted from reliable medical sources to create one answer for each question, and (iii) textual clinical findings in radiology reports to generate radiology impression statements.

For the first two tasks, we created new test sets for the official evaluation using consumer health questions received by the U.S. National Library of Medicine (NLM) and answers retrieved from reliable sources using the Consumer Health Question Answering system CHiQA[5]. For the third task, we created a new test set by combining public radiology reports in the Indiana Univer-

[1] www-nlpir.nist.gov/projects/duc

[2] tac.nist.gov/tracks

[3] sites.google.com/view/mediqa2019

[4] sites.google.com/view/mediqa2021

[5] chiqa.nlm.nih.gov

Proceedings of the BioNLP 2021 workshop, pages 74–85
June 11, 2021.

sity dataset (Demner-Fushman et al., 2016) and newly released chest x-ray reports from the Stanford Health Care.

Through these tasks, we focus on studying:

- The best approaches according to the summarization task objective and the language/vocabulary (consumers' questions, patient-oriented medical text, and professional clinical reports);
- The impact of medical data scarcity on the development and performance of summarization methods in comparison with open-domain summarization;
- The effects of different summary evaluation measures including lexical metrics such as ROUGE (Lin, 2004), embedding-based metrics such as BERTScore (Zhang et al., 2019), and hybrid ensemble-oriented metrics such as HOLMS (Mrabet and Demner-Fushman, 2020).

2 MEDIQA 2021 Task Descriptions

2.1 Consumer Health Question Summarization (QS)

Consumer health questions tend to contain peripheral information that hinders automatic Question Answering (QA). Empirical studies based on manual expert summarization of these questions showed a substantial improvement of 58% in QA performance (Ben Abacha and Demner-Fushman, 2019a). Effective automatic summarization methods for consumer health questions could therefore play a key role in enhancing medical question answering. The goal of this task is to promote the development of new summarization approaches that address specifically the challenges of long and potentially complex consumer health questions. Relevant approaches should be able to generate a condensed question expressing the minimum information required to find correct answers to the original question (Ben Abacha and Demner-Fushman, 2019b).

2.2 Multi-Answer Summarization (MAS)

Different answers can bring complementary perspectives that are likely to benefit the users of QA systems. The goal of this task is to promote the development of multi-answer summarization approaches that could solve simultaneously the aggregation and summarization problems posed by multiple relevant answers to a medical question (Savery et al., 2020).

2.3 Radiology Report Summarization (RRS)

The task of radiology report summarization aims to promote the development of clinical summarization models that are able to generate the concise impression section (i.e., summary) of a radiology report conditioned on the free-text findings and background sections (Zhang et al., 2018). The resulting systems have significant potential to improve the efficiency of clinical communications and accelerate the radiology workflow. While state-of-the-art techniques in language generation have enabled the generation of fluent summaries, these models occasionally generate spurious facts limiting the clinical validity of the generated summaries (Zhang et al., 2020b). It is therefore important to develop systems that are able to summarize the radiology findings in a consistent manner.

3 Data Description

3.1 QS Datasets

The MeQSum dataset of consumer health questions and their summaries (Ben Abacha and Demner-Fushman, 2019b) was suggested as a training dataset. It consists of 1,000 consumer health questions and their associated summaries. Participants were encouraged to use available external resources including, but not limited to, medical QA datasets and question focus and type recognition datasets. For instance, the Consumer Health Questions dataset (Kilicoglu et al., 2018) contains annotations of medical entities, focus, and type of the MeQSum questions and additional NLM questions[6].

The new QS validation and test sets[7] cover a wide range of topics and question types such as *Treatment*, *Information*, *Side effects*, *Cause*, *Effect*, *Person-Organization*, *Diet-Lifestyle*, *Complications*, *Contraindications*, *Diagnosis*, *Usage*, *Interaction*, *Ingredients*, *Prognosis*, *Susceptibility*, *Transmission*, and *Toxicity*. They consist of manually de-identified consumer health questions received by the U.S. National Library of Medicine and gold summaries created by medical experts. The validation set includes 50 NLM questions and

[6] https://bionlp.nlm.nih.gov/CHIQAcollections/CHQA-Corpus-1.0.zip

[7] https://github.com/abachaa/MEDIQA2021/tree/main/Task1

Example 1 (QID: 139) **NLM Question:** *did anyone have this and does it require surgery? my mri says forminal stenosis from bone spurs c4,5,6. my nerve test shows severe nerve compression c7,8. i'm in so much pain, mostly my arm and shoulder and leg. waiting to see the pain specialist to see what's next. would love to know what you guys think is required.* **Question Summary:** *How can I get rid of pain caused by foraminal stenosis and nerve compression?*
Example 2 (QID: 111) **NLM Question:** *covid-19 how long to quarantine after being positive how long are you contagious if i tested positive for covid-19. how long before i can safely return to work after a positive covid 19 test* **Question Summary:** *How long will I remain contagious after testing positive for COVID-19?*

Table 1: Test set examples for the QS task.

their summaries with additional annotations of the question focus and type. The test set contains 80 consumer health questions. Table 1 presents two examples from the QS test set.

3.2 MAS Datasets

The MEDIQA-AnS dataset (Savery et al., 2020) was suggested as a training set for the MAS task. Participants were allowed to use available external resources (e.g. existing medical QA datasets) as well as data creation, selection, and augmentation methods. To create the MAS validation and test sets[8], we used 130 consumer health questions received by NLM. In order to retrieve more accurate answers, we created question summaries that we used to query the medical QA system CHiQA that searches for answers from only trustworthy medical information sources (Ben Abacha and Demner-Fushman, 2019c; Demner-Fushman et al., 2020).

The answer summaries were manually created by medical experts. We provided both extractive and abstractive gold summaries, and encouraged the use of all types of summarization approaches (extractive, abstractive, and hybrid). The MAS validation set contains 192 answers to 50 medical questions. The test set contains 303 answers to 80 medical questions. Each question has at least two answers, one extractive multi-answer summary, and one abstractive multi-answer summary. Table 2 presents an example from the test set.

Original NLM question: *I have dementia like symptoms and wanted to know where is the best source to be tested for diagnosis? I have been prescribed Anticholinergic medicine since 2008...since I have been diagnosed with, Celiac disease and Obstructive Sleep Apnea. I think I have Frontal Temporal lobe atrophy. I'm going to try to get tested...any references on which process is easiest would be much appreciated. I can't take my Nasalcrom allergy spay any more nor, valium or prozac, benadryl and glutamate additives in meats because it sends me straight into cognitive emergency state and irrational thinking*
Question summary used in answer retrieval: *What tests are used to diagnose dementia?*
CHiQA's Answer #1: *Dementia is not a specific disease. It is a descriptive term for a collection of symptoms that can be caused by a number of disorders that affect the brain. People with dementia have significantly impaired intellectual functioning that interferes with normal activities and relationships. They also lose their ability to solve problems and maintain emotional control, and they may experience personality changes and behavioral problems, such as agitation, delusions, and hallucinations (...).*
CHiQA's Answer #2: *To diagnose dementia, doctors first assess whether a person has an underlying treatable condition such as depression, abnormal thyroid function, normal pressure hydrocephalus, or vitamin B12 deficiency. Early diagnosis is important, as some causes for symptoms can be treated. In many cases, the specific type of dementia a person has may not be confirmed until after the person has died and the brain is examined.A medical assessment for dementia generally includes:- Patient history (...) - Physical exam (...) - Neurological tests (...).*
Reference Extractive Summary: *Dementia is not a specific disease. It is a descriptive term for a collection of symptoms that can be caused by a number of disorders that affect the brain. Doctors diagnose dementia only if two or more brain functions - such as memory and language skills – are significantly impaired without loss of consciousness. To diagnose dementia, doctors first assess whether a person has an underlying treatable condition such as depression, abnormal thyroid function, normal pressure hydrocephalus, or vitamin B12 deficiency. Early diagnosis is important, as some causes for symptoms can be treated. In many cases, the specific type of dementia a person has may not be confirmed until after the person has died and the brain is examined.A medical assessment for dementia generally includes:- Patient history (...) - Physical exam (...) - Neurological tests (...).*
Reference Abstractive Summary: *Dementia could be caused by many different diseases of the brain. it is diagnosed if at least two brain functions are effected, for example, if people experience memory loss and changes in behavior and personality. Diagnostic tests for dementia include family history, physical examination, and neurological tests to asses balance, sensory functions, reflexes, vision, eye movements, and cognitive functions. In many cases, the type of dementia is confirmed after the person dies.*

Table 2: Test set example for the MAS task (QID:105).

[8] https://github.com/abachaa/MEDIQA2021/tree/main/Task2

3.3 RRS Datasets

We focus on the summarization of chest radiography reports for the RRS task, since chest radiography represents the most common study type in radiology, and public resources for chest studies are easily accessible. For training, we sampled a collection of 91,544 reports from the MIMIC-CXR chest X-ray report dataset[9] based on simple criteria such as the acceptable length of each section. For validation, we combined another 2,000 reports from the MIMIC-CXR dataset and 2,000 reports from the Indiana University chest X-ray dataset[10](Demner-Fushman et al., 2016). We sampled the reports such that there is no overlapping patients in the validation and training sets.

For the official test set, we used a combination of 300 reports from the Indiana dataset and 300 newly released chest X-ray reports drawn from the Stanford Health Care system. We intentionally designed the test set to be partially from a hospital system different from the training set (out-of-domain) to test the generalizability of the participating systems.

4 Evaluation

4.1 Evaluation Measures

Several new metrics for evaluating text generation systems were studied in recent years (Mao et al., 2020; Bhandari et al., 2020a,b; Zhang et al., 2019; Sellam et al., 2020), with a focus on evaluating text generation based on deep and contextualized representations. To understand these metrics in the context of summarization, Fabbri et al. (2020) have compared 34 traditional and recent model-based metrics on a manually annotated subset from the CNN-DM dataset. Although the study relied only on one correlation factor (Kendall's Tau) and one dataset, it highlighted the (continued) general relevance of ROUGE variants (Lin, 2004) and the challenge of designing or determining the best measure to use. Specifically, the study found that a different measure obtained the best score in each of the four considered evaluation dimensions: *coherence*, *consistency*, *fluency*, and *relevance*, with substantial discrepancies in rankings.

In parallel, HOLMS was recently proposed as an ensemble measure combining both contextualized similarity and a lexical ROUGE component through a multi-dimensional Gaussian function (Mrabet and Demner-Fushman, 2020). HOLMS was evaluated on multiple DUC and TAC datasets, and three correlation factors (Pearson's, Spearman's, and Kendall's), and was shown to benefit from the complementary strengths of lexical and language model-based similarity measurements for evaluating summarization systems.

In this shared task, we chose ROUGE-2 as our official ranking metric following its superiority observed by Owczarzak et al. (2012) on multiple TAC summarization datasets, and by Bhandari et al. (2020c) on the CNN-DM dataset.

We chose two additional metrics for the three tasks: (1) BERTScore for its wider adoption as a language model-based text generation metric, and (2) HOLMS for its hybrid and ensemble-oriented approach. For the RRS task we also considered an additional evaluation metric based on the hamming similarity on the labels produced by the CheXbert labeler (Smit et al., 2020) when applied to both the system and reference summaries, similar to the approach by Zhang et al. (2020b).

4.2 Baseline Systems

Our baseline system for the QS task relied on a distilled PEGASUS model (Zhang et al., 2020a) trained on the CNN-DM dataset and fine-tuned on a combination of biomedical answer-to-question data and question summarization data from MeQSum, LiveQA-Med data (Ben Abacha et al., 2017), a collection of clinical questions (Ely et al., 2000), and Quora question pairs dataset (Iyer et al., 2017). For the Quora and clinical questions datasets, we extracted only the question pairs with a minimum token reduction ratio of 33%.

Our extractive baseline for the MAS task relied on sentence clustering and selection. We used our fine-tuned question summarization model to generate a short question from each sentence, and then clustered the sentences using a word-based cosine distance between the generated questions and a distance threshold set to 0.7. Intersecting clusters were merged. For each cluster, we selected the sentence that was the best cumulative TF-IDF answer to all other sentences as a representative.

For the RRS task, we prepared three baselines: a base pointer-generator model without modeling the background section of a radiology report, a full pointer-generator model with background model-

[9] https://physionet.org/content/mimic-cxr/2.0.0/

[10] openi.nlm.nih.gov/faq#collection

ing (Zhang et al., 2018), and a zero-shot T5-base summarization model (Raffel et al., 2020).

5 Official Results

We published three AIcrowd projects (one for each task) to release the datasets and manage team registration, submission, and leaderboard ranking[11].

5.1 Participating Teams

In total, 35 teams participated in the MEDIQA shared tasks and submitted 310 individual runs (with a limit of ten runs per team per task). Table 3 presents the participating teams with accepted working notes papers. The results of all 35 teams are available on AIcrowd and on the MEDIQA 2021 website.

5.2 Summarization Approaches & Results

A vast majority of the approaches submitted to the QS and RRS tasks were abstractive and relied on fine-tuning of pre-trained generative language models and encoders-decoders architectures. For the MAS task, most submitted approaches were extractive and used a wide spectrum of sentence selection techniques.

Question Summarization. Table 4 presents the official results of the teams with accepted working notes papers from the 22 teams that participated in the QS task.

All approaches submitted to the question summarization task were abstractive methods relying on the fine-tuning of pretrained transformer models (Vaswani et al., 2017). A wide variety of fine tuning, knowledge-based, and ensemble methods was investigated by the participating teams to achieve higher performance (Mrini et al., 2021; Xu et al., 2021; Zhu et al., 2021; Sänger et al., 2021; Lee et al., 2021b; Balumuri et al., 2021; Yadav et al., 2021; He et al., 2021; Lee et al., 2021a). A first interesting insight from the overview is that building ensemble models with deep neural networks such as discriminators is not a trivial task, and achieves results that stay on par with the best single model (Sänger et al., 2021). In contrast, heuristic, downstream ensembles of the models outputs led to substantial improvements when compared to its components/single models (He et al., 2021). The best performing approach relied on such an ensemble by ranking the outputs of PEGASUS, T5, and BART models according to hand-picked features based on the contents of the input question and lengths of the outputs. Spell checking was also a performance boost factor in the question summarization task with some teams using a knowledge base to correct misspelling errors in the original long questions (He et al., 2021), and others relying on third party tools such as CSpell (Yadav et al., 2021; Lu et al., 2019). The datasets used for transfer learning or fine-tuning also played a major role in the achieved performance as demonstrated, for instance, by the combination of datasets from HealthCareMagic, question entailment recognition and question summarization in (Mrini et al., 2021). Moving forward, we think that the overview of the question summarization task revealed two key challenges that need to be addressed to enhance the relevance and performance of existing systems:

1. a relevant learning-based ensemble method that could rely either on the textual outputs or the logits of single models.

2. a more systemic way to select the most relevant datasets for both pretraining and fine tuning.

Multi-Answer Summarization. Both extractive and abstractive approaches were used by the 17 teams that submitted runs to MAS task (Zhu et al., 2021; Can et al., 2021; Xu et al., 2021; Mrini et al., 2021; Yadav et al., 2021; Le et al., 2021; Lee et al., 2021a). Table 5 and Table 6 present official results of the teams with extractive and abstractive systems when evaluated, respectively, on extractive gold summaries and abstractive gold summaries.

The best MAS run (Zhu et al., 2021) relied on an ensemble method and a recent multi-document summarization approach (Xu and Lapata, 2020) using a Roberta model to rank locally the candidate sentences and a Markov chain to evaluate them globally. A similar approach was also used by the ChicHealth team (Xu et al., 2021) without a downstream ensemble method. Participating teams used transfer learning (e.g. (Mrini et al., 2021)) as well as answer sentence selection methods. Sentence selection was used in building extractive summaries (e.g. (Can et al., 2021)) and as an intermediate step in abstractive summarization to provide more concise inputs to generative models (e.g. (Le et al., 2021)). Different models, such

[11] www.aicrowd.com/challenges/mediqa-2021

Team	**Institution**	**QS**	**MAS**	**RRS**
BDKG (Dai et al., 2021)	Baidu, Inc			✓
ChicHealth (Xu et al., 2021)	Chic Health		✓	✓
damo_nlp (He et al., 2021)	Alibaba Group	✓		✓
IBMResearch (Mahajan et al., 2021)	IBM Research			✓
MNLP (Lee et al., 2021a)	George Mason University	✓	✓	
NCUEE-NLP (Lee et al., 2021b)	National Central University	✓		
NLM (Yadav et al., 2021)	U.S. National Library of Medicine	✓	✓	
optumize (Kondadadi et al., 2021)	Optum			✓
paht_nlp (Zhu et al., 2021)	ECNU & Pingan Health Tech	✓	✓	✓
QIAI (Delbrouck et al., 2021)	Stanford University	✓		✓
SB_NITK (Balumuri et al., 2021)	National Institute of Technology Karnataka	✓		
UCSD-Adobe (Mrini et al., 2021)	UC San Diego & Adobe Research	✓	✓	
UETfishes (Le et al., 2021)	VNU University of Engineering and Technology		✓	
UETrice (Can et al., 2021)	VNU University of Engineering and Technology		✓	
WBI (Sänger et al., 2021)	Humboldt University of Berlin	✓		

Table 3: Participating teams with accepted working notes papers at BioNLP-MEDIQA 2021

Rank	**Team**	**ROUGE-2**	**ROUGE-1**	**ROUGE-L**	**HOLMS**	**BERTScore**
1	damo_nlp	**0.1608**	**0.3514**	0.3131	0.5677	0.6898
2	WBI	0.1599	0.3340	**0.3149**	0.5767	0.6996
3	NCUEE-NLP	0.1597	0.3352	0.3090	**0.5787**	0.6960
4	NLM	0.1514	0.3556	0.3110	0.5649	0.6892
5	UCSD-Adobe	0.1414	0.3463	0.3065	0.5586	0.6942
6	ChicHealth	0.1398	0.3403	0.2962	0.5551	0.6810
7	SB_NITK	0.1393	0.3331	0.3077	0.5663	**0.7025**
–	*QS Baseline*	0.1373	0.3203	0.2962	0.5672	0.6277
8	MNLP	0.1114	0.2840	0.2587	0.5455	0.6732
9	paht_nlp	0.0935	0.2486	0.2331	0.5428	0.6591
10	QIAI	0.0385	0.1514	0.1356	0.4898	0.5101

Table 4: Official results of the MEDIQA-QS task.

Rank	**Team**	**ROUGE-2**	**ROUGE-1**	**ROUGE-L**	**HOLMS**	**BERTScore**
1	paht_nlp	**0.5076**	0.5848	0.4354	0.7047	**0.8038**
2	UETrice	0.5040	**0.6110**	**0.4412**	0.7383	0.7958
3	ChicHealth	0.4893	0.5776	0.4261	0.7033	0.7916
4	UCSD-Adobe	0.4720	0.6073	0.4289	**0.7612**	0.7753
5	NLM	0.4677	0.5470	0.3276	0.6575	0.7645

Table 5: Official results of the MEDIQA-MAS task (1): **Extractive Approaches**.

Team	**Rank**	**ROUGE-2**	**ROUGE-1**	**ROUGE-L**	**HOLMS**	**BERTScore**
paht_nlp	**1**	**0.5076**	0.5848	**0.4354**	0.7047	**0.8038**
	(1)	**0.1621**	0.3215	0.1910	0.4220	**0.6528**
UETfishes	**2**	0.4698	0.5720	0.4001	0.6970	0.7821
	(3)	0.1495	0.3124	0.1885	0.4213	0.6466
UCSD-Adobe	**3**	0.4595	**0.5921**	0.4170	**0.7502**	0.7689
	(2)	0.1604	**0.3843**	**0.2117**	0.4937	0.6326
MNLP	4	0.2594	0.4220	0.2954	0.6568	0.6479
	(4)	0.1167	0.3490	0.2047	**0.5269**	0.5763

Table 6: Official results of the MEDIQA-MAS task (2): **Abstractive Approaches**. Ranks in bold and in parenthesis correspond to evaluation on extractive gold summaries and on abstractive gold summaries, respectively.

Rank	Team	R-2	R-1	R-L	HOLMS	BERTScore	CheXbert
1	BDKG	**0.4362**	**0.5572**	**0.5365**	**0.7402**	**0.7184**	**0.6927**
2	IBMResearch	0.4082	0.5328	0.5134	0.7185	0.7115	0.6774
3	optumize	0.3918	0.5185	0.4957	0.7087	0.6975	0.6773
4	QIAI	0.3778	0.4954	0.4793	0.7132	0.5328	0.5565
5	ChicHealth	0.3236	0.4606	0.4410	0.6822	0.6768	0.6261
6	damo_nlp	0.2763	0.4329	0.4115	0.6604	0.6576	0.6343
–	*baseline (PG-full)*	0.2734	0.4182	0.4041	0.6647	0.6194	0.6014
–	*baseline (PG-base)*	0.2639	0.4026	0.3885	0.6553	0.6103	0.5537
7	paht_nlp	0.1987	0.3400	0.3053	0.5915	0.5985	0.6705
–	*baseline (T5)*	0.0945	0.2108	0.1831	0.4432	0.4921	0.5245

Table 7: Official results of the MEDIQA-RRS task on the full test set.

Rank	Team	ROUGE-2		CheXbert	
		Stanford	Indiana	Stanford	Indiana
1	BDKG	**0.2768**	**0.5955**	0.6547	**0.7052**
2	ChicHealth	0.2690	0.3781	0.6291	0.5873
3	damo_nlp	0.2687	0.2839	0.6645	0.5517
4	optumize	0.2654	0.5182	0.6474	0.6592
5	QIAI	0.2516	0.5039	0.5508	0.4970
6	paht_nlp	0.2491	0.1483	**0.6834**	0.6148
–	*baseline (PG-full)*	0.2414	0.3054	0.6216	0.5466
–	*baseline (PG-base)*	0.2408	0.2870	0.5892	0.4754
7	IBMResearch	0.2283	0.5880	0.6472	0.6937
–	*baseline (T5)*	0.1280	0.0610	0.5067	0.5609

Table 8: Official results of the MEDIQA-RRS task on the Stanford and Indiana test splits.

as BART and T5, and datasets (e.g. MEDIQA-AnS, MSMARCO, MEDIQA-2019) have been used for single and multiple answer summarization (Yadav et al., 2021; Mrini et al., 2021; Zhu et al., 2021; Can et al., 2021).

Radiology Report Summarization. 14 teams participated in the RRS task. Table 7 presents the official results of the teams (with accepted papers) on the full test set, and Table 8 presents the results on the Stanford and Indiana subsets of the test set.

Similar to the previous tasks, participating teams for the RRS task have extensively used pretrained transformer models: out of the 7 teams that submitted papers describing their systems, 6 reported the use of pretrained language models such as BART or PEGASUS in their submissions (Xu et al., 2021; Zhu et al., 2021; Kondadadi et al., 2021; Dai et al., 2021; Mahajan et al., 2021; He et al., 2021). Among them, Xu et al. (2021); Zhu et al. (2021); Dai et al. (2021) reported that best results were achieved with pretrained PEGASUS models, while Kondadadi et al. (2021) reported better results from BART. Xu et al. (2021) and Zhu et al. (2021) reported that using PEGASUS models pretrained on the PubMed corpus yielded worse results than using the general PEGASUS models, potentially due to the domain difference of the RRS task with the PubMed text.

In addition to the use of pretrained models, the highest-ranked systems from Dai et al. (2021) made effective use of a dedicated domain adaptation module, an ensemble module, and text normalization heuristics. Zhu et al. (2021) reported that freezing the embedding layer in the pretrained models helps the model generalize at test time. Kondadadi et al. (2021) reported that adding the background section as input improves performance at validation time, but not test time, suggesting that the model performance is sensitive to the different text styles of the background sections from different splits. Mahajan et al. (2021) focused their study on the factual consistency of generated summaries, and proposed a specialized fact-aware re-ranking approach based on the predicted disease values from the findings section with a transformer model. As a result, their submissions achieved competitive rankings under the CheXbert metric. Lastly, Delbrouck et al. (2021) studied the use of image features for the RSS task: they retrieved and linked images for each study to the report at training and validation time, and combined a visual encoder with a text encoder for the summarization task. They found that at validation time the multi-modal setting is beneficial to the summarization of MIMIC reports, but not to the Indiana reports, potentially due to the distribution shift in the images.

6 Correlations among the Evaluation Measures

In this section, we discuss correlations between the different evaluation metrics that we used in the challenge. Table 9 shows Pearson correlations between the F1 scores of the three lexical measures (ROUGE-1, ROUGE-2, and ROUGE-L) and the two language model-based and ensemble-based measures (i.e., HOLMS and BERTScore).

Over all three tasks the HOLMS metric had a better Pearson correlation with ROUGE, ranging from 0.734 to 0.755, while also maintaining a high correlation of 0.736 with BERTScore. This observation supports the findings from the experiments in (Mrabet and Demner-Fushman, 2020), which suggested that lexical measures such as ROUGE and language model-based measures bring different and complementary perspectives to summary-evaluation.

Table 10 shows Pearson correlations for the RRS task. HOLMS is substantially closer than CheXbert and BERTScore in its correlation with ROUGE for the RRS task, while maintaining high correlation of respectively 0.645 and 0.702 with CheXbert and BERTScore.

In contrast, BERTScore is substantially closer than HOLMS in its correlation with the ROUGE metrics for both the MAS task (cf. table 11) and the QS task (see Table 12). Two factors that could explain these correlations are (i) the predominance of extractive runs in the MAS task and (ii) the sequential n-gram-based modeling in HOLMS that takes into account the order of the n-grams, while BERTScore relies on a cosine distance between two given sets of token embeddings.

Both language model-based measures had positive correlations with ROUGE for the QS task, but the level of correlation was substantially lower when compared to the MAS and RRS tasks, going from a Pearson coefficient range between 0.663 and 0.958 to a range between 0.193 and 0.372. As all submitted QS runs were described as abstractive or hybrid approaches, this discrepancy might be due to a stronger disagreement on summary assessment due to semantically-close but lexically distant summaries. It is also likely that the lexical distance between paraphrases was more pronounced due to the lengths of the question summaries, which are shorter than the summaries in the MAS task.

7 Conclusion

We presented an overview of the MEDIQA 2021 shared tasks on summarization in the medical domain. We presented the results for the three tasks on Question Summarization, Multi-Answer Summarization and Radiology Reports Summarization, and discussed the impact of summarization approaches and automatic evaluation methods. We find that pre-trained transformer models, fine-tuning on the carefully selected domain-specific text and ensemble methods worked well for all three summarization tasks. The results encourage future research to include in-depth exploration of ensemble methods, systematic approaches to selection of datasets for pre-training and fine-tuning, as well as a thorough assessment of the quality and relevance of different evaluation measures for summarization. We hope that the MEDIQA 2021 shared tasks will encourage further research efforts in medical text summarization and evaluation.

Acknowledgments

This research was partially supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health. We thank Anna Ripple (NLM/NIH) for her help with the manual annotation and Soumya Gayen (NLM/NIH) for his help with the summarization interface.

References

Stergos D. Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from medical documents: A survey. *Artif Intell Med*, 33(2):157–77.

Muhammad Afzal, Fakhare Alam, Khalid Mahmood Malik, and Ghaus M Malik. 2020. Clinical context–aware biomedical text summarization using deep

Measure	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**HOLMS**	**BERTScore**
ROUGE-1	1.000				
ROUGE-2	0.966	1.000			
ROUGE-L	0.813	0.762	1.000		
HOLMS	**0.734**	**0.722**	**0.755**	1.000	
BERTScore	0.546	0.519	0.409	**0.736**	1.000

Table 9: Pearson correlations between metrics aggregated over all three tasks. For ROUGE and BERTScore we use their F1 scores. Best correlations with the ROUGE metrics are highlighted in bold.

Measure	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**CheXbert**	**HOLMS**	**BERTScore**
ROUGE-1	1.000					
ROUGE-2	0.970	1.000				
ROUGE-L	0.998	0.975	1.000			
CheXbert	0.777	0.667	0.749	1.000		
HOLMS	**0.951**	**0.938**	**0.958**	0.645	1.000	
BERTScore	0.752	0.663	0.743	**0.719**	**0.702**	1.000

Table 10: Pearson correlations between metrics for the RRS task. For ROUGE and BERTScore we used the F1 scores. Best correlations with the lexical measures are highlighted in bold.

Measure	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**HOLMS**	**BERTScore**
ROUGE-1	1.000				
ROUGE-2	0.960	1.000			
ROUGE-L	0.951	0.946	1.000		
HOLMS	0.812	0.823	0.873	1.000	
BERTSCore	**0.913**	**0.924**	**0.889**	**0.784**	1.000

Table 11: Pearson correlations between metrics for the MAS task. Extractive runs were evaluated on extractive gold summaries. Abstractive runs were evaluated on both extractive and abstractive gold summaries. All evaluation scores were concatenated to compute correlations. For ROUGE and BERTScore we used the F1 scores. Best correlations with the lexical measures are highlighted in bold.

Measure	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**HOLMS**	**BERTScore**
ROUGE-1	1.000				
ROUGE-2	0.951	1.000			
ROUGE-L	0.944	0.981	1.000		
HOLMS	0.193	0.204	0.259	1.000	
BERTSCore	**0.292**	**0.332**	**0.372**	0.972	1.000

Table 12: Pearson correlations between metrics for the QS task. For ROUGE and BERTScore we used the F1 scores. Best correlations with the lexical measures are highlighted in bold.

neural network: Model development and validation. *J Med Internet Res*, 22(10):e19810.

Spandana Balumuri, Sony Bachina, and Sowmya Kamath S. 2021. Sb_nitk at mediqa 2021: Leveraging transfer learning for question summarization in medical domain. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Asma Ben Abacha, Eugene Agichtein, Yuval Pinter, and Dina Demner-Fushman. 2017. Overview of the medical question answering task at trec 2017 liveqa. In *TREC 2017*.

Asma Ben Abacha and Dina Demner-Fushman. 2019a. On the role of question summarization and information source restriction in consumer health question answering. In *Proceedings of the AMIA 2019 Informatics Summit, San Francisco, CA, USA, 2019*.

Asma Ben Abacha and Dina Demner-Fushman. 2019b. On the summarization of consumer health questions. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2228–2234. Association for Computational Linguistics.

Asma Ben Abacha and Dina Demner-Fushman. 2019c. A question-entailment approach to question answering. *BMC Bioinform.*, 20(1):511:1–511:23.

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the MEDIQA 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task, BioNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 370–379. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, and Pengfei Liu. 2020a. Metrics also disagree in the low scoring range: Revisiting summarization evaluation metrics. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5702–5711. International Committee on Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020b. Reevaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 9347–9359. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020c. Reevaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Duy-Cat Can, Quoc-An Nguyen, Quoc-Hung Duong, Minh-Quang Nguyen, Huy-Son Nguyen, Cam-Van Thi Nguyen, Quang-Thuy Ha, and Mai-Vu Tran. 2021. Uetrice at mediqa 2021: A prosper-thy-neighbour extractive multi-document summarization model. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Yen-Chun Chen and Mohit Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 675–686. Association for Computational Linguistics.

Songtai Dai, Quan Wang, Yajuan Lyu, and Yong Zhu. 2021. Bdkg at mediqa 2021: System report for the radiology report summarization task. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Cassie Zhang, and Daniel Rubin. 2021. Qiai at mediqa 2021: Multimodal radiology report summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *J. Am. Medical Informatics Assoc.*, 27(2):194–201.

John W. Ely, Jerome A. Osheroff, Paul N. Gorman, Mark H. Ebell, M. Lee Chambliss, Eric A. Pifer, and P. Zoe Stavri. 2000. A taxonomy of generic clinical questions: classification study. *British Medical Journal*, 321:429–432.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2020. Summeval: Reevaluating summarization evaluation. *arXiv preprint arXiv:2007.12626*.

Sebastian Gehrmann, Yuntian Deng, and Alexander M. Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4098–4109.

Yifan He, Mosha Chen, and Songfang Huang. 2021. damo_nlp at mediqa 2021: Knowledge-base preprocessing and coverage-oriented reranking for medical question summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. First quora dataset release: Question pairs.

Halil Kilicoglu, Asma Ben Abacha, Yassine Mrabet, Sonya E. Shooshan, Laritza Rodriguez, Kate Masterton, and Dina Demner-Fushman. 2018. Semantic annotation of consumer health questions. *BMC Bioinform.*, 19(1):34:1–34:28.

Ravi Kondadadi, Sahil Manchanda, Jason Ngo, and Ronan McCormack. 2021. Optum at mediqa 2021: Abstractive summarization of radiology reports using simple bart finetuning. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Hoang-Quynh Le, Quoc-An Nguyen, Quoc-Hung Duong, Minh-Quang Nguyen, Huy-Son Nguyen, Tam Doan Thanh, Hai-Yen Thi Vuong, and Trang M. Nguyen. 2021. Uetfishes at mediqa 2021: Standing-on-the-shoulders-of-giants model for abstractive multi-answer summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Jooyeon Lee, Huong Dang, Ozlem Uzuner, and Sam Henry. 2021a. Mnlp at mediqa 2021: Fine-tuning pegasus for consumer health question summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Lung-Hao Lee, Po-Han Chen, Yu-Xiang Zeng, Po-Lei Lee, and Kuo-Kai Shyu. 2021b. Ncuee-nlp at mediqa 2021: Health question summarization using pegasus transformers. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chris J Lu, Alan R Aronson, Sonya E Shooshan, and Dina Demner-Fushman. 2019. Spell checker for consumer language (cspell). *Journal of the American Medical Informatics Association*, 26(3):211–218.

Diwakar Mahajan, Ching-Huei Tsou, and Jennifer J Liang. 2021. Ibmresearch at mediqa 2021: Toward improving factual correctness of radiology report abstractive summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Yuning Mao, Liyuan Liu, Qi Zhu, Xiang Ren, and Jiawei Han. 2020. Facet-aware evaluation for extractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4941–4957. Association for Computational Linguistics.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene R. Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: A systematic review of recent research. *Journal of Biomedical Informatics*, 52:457–467. Special Section: Methods in Clinical Research Informatics.

Yassine Mrabet and Dina Demner-Fushman. 2020. HOLMS: alternative summary evaluation with large language models. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 5679–5688. International Committee on Computational Linguistics.

Khalil Mrini, Franck Dernoncourt, Seunghyun Yoon, Trung Bui, Walter Chang, Emilias Farcas, and Ndapa Nakashole. 2021. Ucsd-adobe at mediqa 2021: Transfer learning and answer sentence selection for medical summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, CA, USA.*, pages 3075–3081.

Karolina Owczarzak, John M. Conroy, Hoa Trang Dang, and Ani Nenkova. 2012. An assessment of the accuracy of automatic evaluation in summarization. In *Proceedings of Workshop on Evaluation Metrics and System Comparison for Automatic Summarization@NACCL-HLT 2012, Montrèal, Canada, June 2012, 2012*, pages 1–9. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 379–389.

Max E. Savery, Asma Ben Abachaand Soumya Gayen, and Dina Demner-Fushman. 2020. Question-driven summarization of answers to consumer health questions. *Scientific Data, Nature*, 7.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mario Sänger, Leon Weber, and Ulf Leser. 2021. Wbi at mediqa 2021: Summarizing consumer health questions with generative transformers. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3009–3019. Association for Computational Linguistics.

Liwen Xu, Yan Zhang, Lei Hong, Yi Cai, and Szui Sung. 2021. Chichealth @ mediqa 2021: Exploring the limits of pre-trained seq2seq models for medical summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Yumo Xu and Mirella Lapata. 2020. Coarse-to-fine query focused multi-document summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3632–3645.

Shweta Yadav, Mourad Sarrouti, and Deepak Gupta. 2021. Nlm at mediqa 2021: Transfer learning-based approaches for consumer question and multi-answer summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D. Manning, and Curtis P. Langlotz. 2018. Learning to summarize radiology findings. In *EMNLP 2018 Workshop on Health Text Mining and Information Analysis*.

Yuhao Zhang, Derek Merck, Emily Bao Tsai, Christopher D. Manning, and Curtis Langlotz. 2020b. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Wei Zhu, Yilong He, Ling Chai, Yunxiao Fan, Yuan Ni, GUOTONG XIE, and Xiaoling Wang. 2021. paht_nlp at mediqa 2021: Multi-grained query focused multi-answer summarization. In *Proceedings of the 20th SIGBioMed Workshop on Biomedical Language Processing, NAACL-BioNLP 2021*. Association for Computational Linguistics.