Overview of the MEDIQA-Sum Task at ImageCLEF 2023: Summarization and Classification of Doctor-Patient Conversations*

Wen-wai Yim[1,*], Asma Ben Abacha[1], Griffin Adams[2], Neal Snider[3] and Meliha Yetisgen[4]

[1] *Microsoft Health AI, Redmond, 98052, USA*
[2] *Columbia University, New York, 10027 USA*
[3] *Nuance Communications, Burlington, 01803 USA*
[4] *University of Washington, Seattle, 98109 USA*

Abstract

This paper presents the overview of the MEDIQA-Sum task at ImageCLEF 2023. MEDIQA-Sum 2023 includes three subtasks, in which a doctor-patient dialogue source is given, and participants were tasked with (A) dialogue2topic classification, e.g. classifying the conversation into one of twenty section header categories, (B) dialogue2note snippet generation, e.g. generating clinical note section text additionally given the clinical section header, and (C) dialogue2note full note summarization, e.g. generating a full clinical note. Twelve teams participated with a total of 48 runs. The best teams achieved 0.8 Accuracy on topic classification (subtask A) and ROUGE-1 scores of 0.43 and 0.49 F1, for subtasks B and C, respectively.

Keywords

Dialogue Summarization, Clinical Note Generation, Natural Language Generation, Doctor-patient Conversations

1. Introduction

To date, large language models (LLM) pre-trained with massive amounts of data have lead to surprisingly large out-of-the-box gains across all sectors of machine learning. This is true in tasks that these models were not trained for, e.g. classification tasks, complex tasks that require special syntax and domain knowledge, e.g. generating code based on the functional description, and even creative tasks, e.g. generating original poems given a subject prompt. One specific area to test such technology is the problem of clinical note generation from doctor-patient conversations. As LLM are built to generate, this is a very natural task; on the other hand, note creation in the health care space is a critical ubiquitous burdensome task for health-care professionals[1].

Note generation from doctor-patient conversations is a daily occurrence accompanying a doctor-patient encounter. The clinical note, like meeting notes, highlight important discussion

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.
yimwenwai@microsoft.com (W. Yim); abenabacha@microsoft.com (A. Ben Abacha); griffin.adams@columbia.edu (G. Adams); neal.snider@nuance.com (N. Snider); melihay@uw.edu (M. Yetisgen)

points, relevant history and future planned tests and treatments. While clinical notes are generated natively, doctor-patient conversations are not routinely recorded. Therefore acquisition and testing of such datasets present a prohibitive hurdle[2]. Other domain challenges which increase the difficulty of the task include (a) the existence of clinical note format varieties, as well as their semi-structured technical writing, conditioned on provider preferences, specialties, and institutions[1]; (b) the high diversity and topic spread of the doctor-patient conversation, depending on regional practices, socioeconomic origin, speech characteristics, and meeting discussion preferences; and (c) the length of the generated notes, which are often longer than the typical generation tasks.

To investigate the state-of-the-art performances in this space, we have conducted the MEDIQA-Sum 2023 task as part of IMAGECLEF 2023[3], a pilot task for multi-modal summarization. In the tradition of MEDIQA tasks that began in 2019[4]–hosting various tasks related to clinical language inference, consumer health question answering entailment and retrieval ranking, as well as clinical findings and consumer health question-answering summarization–this year's edition tackles a summarization task that spans clinical dialogue as a source and the clinical note as the target. An overlapping dataset was part of the related ACL 2023 ClinicalNLP challenge MEDIQA-Chat 2023[5].

In the following sections, we introduce the tasks, describe the evaluation, present the participating teams' results, as well as provide some insight on future directions.

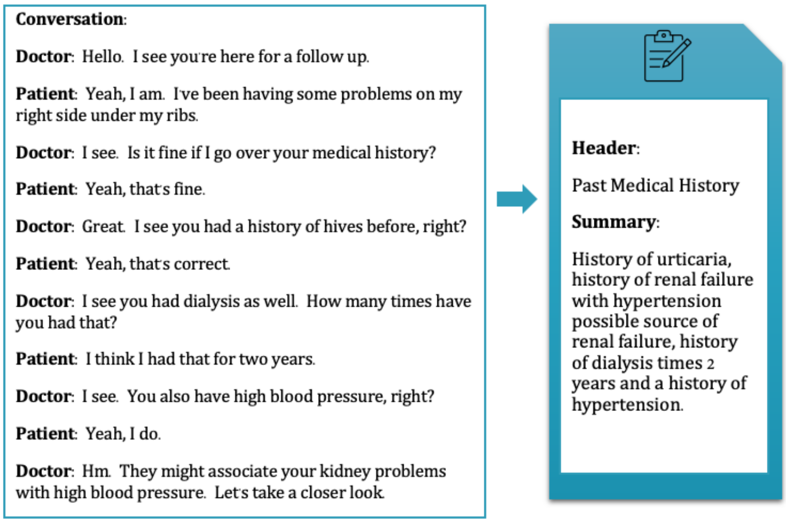
2. Task Description

The MEDIQA-Sum 2023 overall task comprises three sub-tasks: (A) dialogue2topic (section header) classification, (B) dialogue2note summarization given the target section header, and (C) full-encounter dialogue2note summarization. Although it's possible to perform each task in series, with one model or data being utilized for the next, each task could be participated in independently.

2.1. Subtask A - Section Header Topic Classification

In speech language processing, text classification is often used for categorizing dialogue acts, domains, and intents as for dialogue systems[6, 7, 8]; as well as for topic clustering for further speech language processing[9, 10]. We pose the task of dialogue text topic classification as a means of identifying whether the clinical information relevant in a dialogue relates to certain parts of the clinical note. Such a task can be viewed as one step in a multi-step processing of a long dialogue (e.g. clustering similar information) or may be used to get classification information for a short dialogue. Although we simplify the task here so that each dialogue snippet is one of several headers, in real conversations, the same snippets of text may pertain to multiple sections or may be included in different overlapping relevant text windows. Previous work in the area includes the classification of patient dialogue encounters at a sentence level to SOAP format or other categories[11, 12]. Meanwhile the task of clinical section header identification is a well-studied task in clinical NLP[13, 14, 15].

In this subtask, section headers were one of the following 20: Family History/Social History (fam/sochx), History of Present Illness (genhx), Past Medical History (pastmedicalhx), Chief



Conversation:

Doctor: Hello. I see you're here for a follow up.

Patient: Yeah, I am. I've been having some problems on my right side under my ribs.

Doctor: I see. Is it fine if I go over your medical history?

Patient: Yeah, that's fine.

Doctor: Great. I see you had a history of hives before, right?

Patient: Yeah, that's correct.

Doctor: I see you had dialysis as well. How many times have you had that?

Patient: I think I had that for two years.

Doctor: I see. You also have high blood pressure, right?

Patient: Yeah, I do.

Doctor: Hm. They might associate your kidney problems with high blood pressure. Let's take a closer look.

Header:

Past Medical History

Summary:

History of urticaria, history of renal failure with hypertension possible source of renal failure, history of dialysis times 2 years and a history of hypertension.

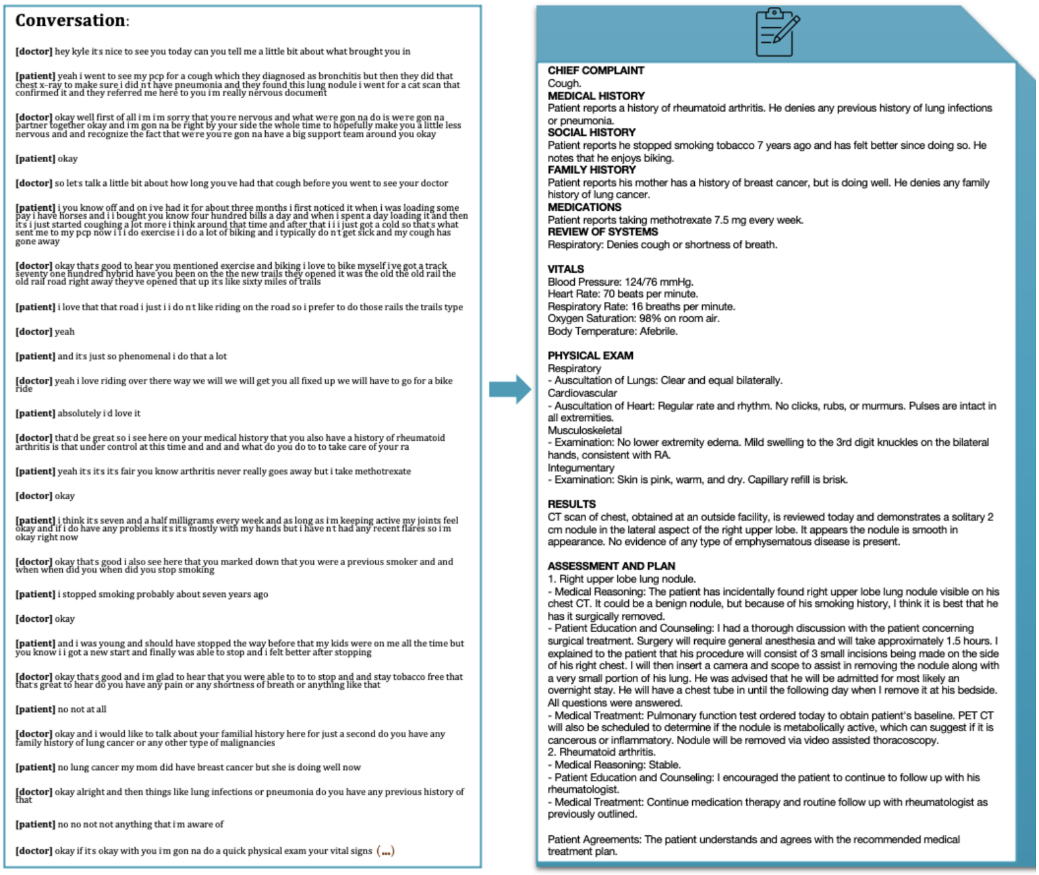
Figure 1: In Subtask A, short doctor-patient conversation can be categorized to a relevant section header. In Subtask B, The same conversation and a given section header can be used to generate a short clinical text snippet (example from the MTS-Dialog dataset).

Complaint [cc], Past Surgical History (pastsurgical), allergy, Review of Systems (ros), medications, assessment, exam, diagnosis, disposition, plan, Emergency Department Course (edcourse), immunizations, imaging, Gynecologic History (gynhx), procedures, other_history, and labs. An example of this problem is shown in Figure 1, where the header is the target output.

2.2. Subtask B - Short Dialogue2Note Summarization

Dialogue summarization encompasses variety of tasks, including spoken conversation and text chatting. Typical English open domain datasets are related to news headline summarization[16, 17]. Related dialogue summarization datasets include MedDialog[18], a dataset of online medical chats and their final treatment summaries, and SAMSUM dataset, a corpus of chat dialogues with manually created summaries[19]. Dialogue2note generation from doctor-patient conversations for short dialogue has been the subject of previous work[20], however their datasets are not open to the public. As shown in Figure 1, here available input includes the dialogue as well as the relevant section header from a short dialogue and the target output is the summary. Specifying the desired header as input is a realistic scenario, as the same dialogue snippet may be relevant to several sections; moreover, different note sections may require different language patterns.

Subtask A and B use the same test set. After Subtask A was closed, the gold standard section header was released so that it would be available as input to Subtask B.



Conversation:

[doctor] hey kyle it's nice to see you today can you tell me a little bit about what brought you in

[patient] yeah i went to see my pcp for a cough which they diagnosed as bronchitis but then they did that chest x-ray to make sure i did n't have pneumonia and they found this lung nodule i went for a cat scan that confirmed it and they referred me here to you i'm really nervous document

[doctor] okay well first of all i'm i'm sorry that you're nervous and what we're gon na do is we're gon na partner together okay and i'm gon na be right by your side the whole time to hopefully make you a little less nervous and and recognize the fact that we're you're gon na have a big support team around you okay

[patient] okay

[doctor] so let's talk a little bit about how long you've had that cough before you went to see your doctor

[patient] i you know off and on i've had it for about three months i first noticed it when i was loading some pay i have horses and i i bought you know four hundred bills a day and when i spent a day loading it and then it's i just started coughing a lot more i think around that time and after that i i i just got a cold so that's what sent me to my pcp now i i i do exercise i i do a lot of biking and i typically do n't get sick and my cough has gone away

[doctor] okay that's good to hear you mentioned exercise and biking i love to bike myself i've got a track seventy one hundred hybrid have you been on the the new trails they opened it was the old the old rail the old rail road right away they've opened that up it's like sixty miles of trails

[patient] i love that that road i just i i do n't like riding on the road so i prefer to do those rails the trails type

[doctor] yeah

[patient] and it's just so phenomenal i do that a lot

[doctor] yeah i love riding over there way we will we will get you all fixed up we will have to go for a bike ride

[patient] absolutely i'd love it

[doctor] that'd be great so i see here on your medical history that you also have a history of rheumatoid arthritis is that under control at this time and and and what do you do to to take care of your ra

[patient] yeah it's it's it's fair you know arthritis never really goes away but i take methotrexate

[doctor] okay

[patient] i think it's seven and a half milligrams every week and as long as i'm keeping active my joints feel okay and if i do have any problems it's it's mostly with my hands but i have n't had any recent flares so i'm okay right now

[doctor] okay that's good i also see here that you marked down that you were a previous smoker and and when when did you when did you stop smoking

[patient] i stopped smoking probably about seven years ago

[doctor] okay

[patient] and i was young and should have stopped the way before that my kids were on me all the time but you know i i got a new start and finally was able to stop and i felt better after stopping

[doctor] okay that's good and i'm glad to hear that you were able to to to stop and and stay tobacco free that that's great to hear do you have any pain or any shortness of breath or anything like that

[patient] no not at all

[doctor] okay and i would like to talk about your familial history here for just a second do you have any family history of lung cancer or any other type of malignancies

[patient] no lung cancer my mom did have breast cancer but she is doing well now

[doctor] okay alright and then things like lung infections or pneumonia do you have any previous history of that

[patient] no no not not anything that i'm aware of

[doctor] okay if it's okay with you i'm gon na do a quick physical exam your vital signs (...)

CHIEF COMPLAINT
Cough.
MEDICAL HISTORY
Patient reports a history of rheumatoid arthritis. He denies any previous history of lung infections or pneumonia.
SOCIAL HISTORY
Patient reports he stopped smoking tobacco 7 years ago and has felt better since doing so. He notes that he enjoys biking.
FAMILY HISTORY
Patient reports his mother has a history of breast cancer, but is doing well. He denies any family history of lung cancer.
MEDICATIONS
Patient reports taking methotrexate 7.5 mg every week.
REVIEW OF SYSTEMS
Respiratory: Denies cough or shortness of breath.

VITALS
Blood Pressure: 124/76 mmHg.
Heart Rate: 70 beats per minute.
Respiratory Rate: 16 breaths per minute.
Oxygen Saturation: 98% on room air.
Body Temperature: Afebrile.

PHYSICAL EXAM
Respiratory
- Auscultation of Lungs: Clear and equal bilaterally.

Cardiovascular
- Auscultation of Heart: Regular rate and rhythm. No clicks, rubs, or murmurs. Pulses are intact in all extremities.

Musculoskeletal
- Examination: No lower extremity edema. Mild swelling to the 3rd digit knuckles on the bilateral hands, consistent with RA.

Integumentary
- Examination: Skin is pink, warm, and dry. Capillary refill is brisk.

RESULTS
CT scan of chest, obtained at an outside facility, is reviewed today and demonstrates a solitary 2 cm nodule in the lateral aspect of the right upper lobe. It appears the nodule is smooth in appearance. No evidence of any type of emphysematous disease is present.

ASSESSMENT AND PLAN
1. Right upper lobe lung nodule.
- Medical Reasoning: The patient has incidentally found right upper lobe lung nodule visible on his chest CT. It could be a benign nodule, but because of his smoking history, I think it is best that he has it surgically removed.
- Patient Education and Counseling: I had a thorough discussion with the patient concerning surgical treatment. Surgery will require general anesthesia and will take approximately 1.5 hours. I explained to the patient that his procedure will consist of 3 small incisions being made on the side of his right chest. I will then insert a camera and scope to assist in removing the nodule along with a very small portion of his lung. He was advised that he will be admitted for most likely an overnight stay. He will have a chest tube in until the following day when I remove it at his bedside. All questions were answered.
- Medical Treatment: Pulmonary function test ordered today to obtain patient's baseline. PET CT will also be scheduled to determine if the nodule is metabolically active, which can suggest if it is cancerous or inflammatory. Nodule will be removed via video assisted thoracoscopy.

2. Rheumatoid arthritis.
- Medical Reasoning: Stable.
- Patient Education and Counseling: I encouraged the patient to continue to follow up with his rheumatologist.
- Medical Treatment: Continue medication therapy and routine follow up with rheumatologist as previously outlined.

Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.

Figure 2: Task B: summarize each doctor-patient conversation to generate a full note with all relevant sections (example from the ACI-Bench dataset).

2.3. Subtask C - Full Dialogue2Note Summarization

In the full encounter summarization subtask, Subtask C, the objective is to generate a complete clinical note for each doctor-patient conversation, as shown in Figure 2. A similar task to this is the meeting summarization task, which includes long dialogues of multiple speakers as well as technical, at times semi-structured meeting notes[21, 22]. One challenge in this space is the long-document nature of notes. Similar issues arise in PubMed and arXiv scientific paper summarization, as well as BigPatent, BillSum, and GovReport summarizations[23, 24]. Previous works summarizing full doctor-patient conversations[25, 26] have not focused on full note generation and evaluation, rather splits full note generation targets into smaller note parts rather than the creation of the full note.

3. Dataset

The 2023 MEDIQA-Sum challenge includes data from two collections: MTS-Dialog[27] and ACI-BENCH[28]. Subtasks A and B consisted of 1,201 pairs of conversations and associated section headers and contents; 100 examples in validation, and 200 pairs in test. Subtask C includes full encounters with 67 examples in training, 20 in validation, and 40 in test.

The train and validation set for all subtasks were featured in the MEDIQA-Chat 2023 task[5]; however with different test sets.

4. Evaluation Methodology

Subtask A topic classification was evaluated using accuracy. The subtask B snippet summarization was evaluated using the mean of BLEURT, BERTscore(microsoft/deberta- xlarge-mnli), and ROUGE-1; metrics found to be correlated to human evaluation in several independent health summarization datasets [29]. Full-encounter summarization in Subtask C used two metrics: (1) a full-note ROUGE-1 score and (2) an equally weighted division-based (subjective, objective_exam, objective_results, assessment_and_plan) aggregate score of the BLEURT, BERTscore, and ROUGE-1 metric[28].

Code repositories were required at submission. This was done to encourage high quality submissions as well as encourage participants to release code after the challenge. The organizers checked outputs of code against submitted runs and documented each team's code replicability status as defined here:

1. Code runs and exactly reproduces
2. Code runs with minor differences
3. Results unstable due to non-deterministic components (e.g., generative API calls)
4. Results unstable
5. Code does not run under our configurations

We provided feedback on the shared codes and their outputs/errors to the participants.

5. Results and Discussion

Overall 12 teams participated with a total of 48 runs. Subtask A included 23 valid submissions among 11 teams. Subtask B included 16 submissions among 7 teams. Subtask C included 9 submissions among 4 teams. At most three runs were allowed per team in each subtask. With the exception of 1 team, all teams participated in Subtask A. Four teams participated in two subtasks. Three teams participated in all three subtasks. Table 1 shows the full breakdown.

Table 1
MEDIQA-Sum 2023: Participating teams, number of runs (with a limit of three runs/task), submitted codes, and working notes papers.

	team	affiliation	subtask	runs	code	paper
1	Cadence	Cadence Solutions, USA	A	1	2	[30]
2	ds4dh	University of Geneva, Switzerland	A	1	1	
3	HuskyScribe	University of Washington, USA	A,B,C	4	5,1,1	[31]
4	MLRG-JBTTM	Sri Sivasubramaniya Nadar College of Engineering, India	A	3	1	[32]
5	PULSAR	ASUS AICS / University of Manchester, Singapore, UK	B,C	5	1	[33]
6	SKKU-DSAIL	Department of Applied Artificial Intelligence, Sungkyunkwan University, South Korea	A,B	2	1	
7	SSNdhanyadivyakavitha	Sri Sivasubramaniya Nadar College of Engineering, India	A	3	1	[34]
8	SSNSheerinKavitha	Sri Sivasubramaniya Nadar College of Engineering, India	A,B	6	5	[35]
9	StellEllaStars	University of Michigan , USA	A	3	1	[36]
10	SuryaKiran	Optum, India	A,B	4	1	[37]
11	Tredence	Tredence Inc, India	A,B,C	7	1	[38]
12	uetcorn	University of Engineering and Technology, VNUH, Vietnam	A,B,C	9	1	[39]

Table 2
Performance of the participating teams in the MEDIQA-Sum 2023 Subtask A on topic classification.

team	run	accuracy	rank	code_status
Cadence	run1	0.820	1	2
HuskyScribe	run1	0.815	2	5
Tredence	run2	0.800	3	1
Tredence	run1	0.800	3	1
StellEllaStars	run1	0.765	5	1
Tredence	run3	0.755	6	1
SSNSheerinKavitha	run3	0.740	7	1
SSNSheerinKavitha	run2	0.735	8	1
SuryaKiran	run1	0.735	8	1
SSNdhanyadivyakavitha	run1	0.720	10	1
ds4dh	run1	0.710	11	1
uetcorn	run3	0.710	11	1
SKKU-DSAIL	run1	0.700	13	1
StellEllaStars	run2	0.695	14	1
SSNdhanyadivyakavitha	run2	0.680	15	1
StellEllaStars	run3	0.675	16	1
uetcorn	run1	0.670	17	1
MLRG-JBTTM	run1	0.665	18	1
SSNdhanyadivyakavitha	run3	0.660	19	1
uetcorn	run2	0.625	20	1
MLRG-JBTTM	run2	0.570	21	1
MLRG-JBTTM	run3	0.565	22	1
SSNSheerinKavitha	run1	0.140	23	1

The best teams achieved 0.8 Accuracy on Subtask A topic classification (Table 2) and an aggregate score of 0.43 for Subtask B (Table 3). The top two systems for Subtask C achieved ROUGE-1 at 0.49 F1 (Table 4) and aggregated scores at 0.44 (Table 5).

Subtask A submissions included classic machine learning algorithms as well as neural network based models. Specifically, for each category:
classical models

- SVM/Logistic regression: MLRG-JBTTM, SSNdhanyadivyakavitha, SSNSheerinKavitha, StellEllaStars
- KNN: MLRG-JBTTM
- Random Forest: SSNdhanyadivyakavitha

pretrained models

- CBOW (with custom network): StellEllaStars
- general models (bert, roberta, t5, longformer, bart): SSNSheerinKavitha, HuskyScribe, SKKU-DSAIL
- biomedical models (bioroberta, clinicalbert, bioclinicalbert, biomedical-roberta, clinical-longformer pubmedbert, clinicalT5): HuskyScribe, StellEllaStars, SuryaKiran, Tredence

Pre-processing steps for the classical models included stop word removal, lower-casing, TF-IDF, and lemmatization. Eight out of 23 submissions either used additional training data or adjusted data sampling. The top team, Cadence, used bart-large and additionally augmented the training set with data produced by GPT3.5[30]. The second best system by the HuskyScribe team[31], used a T5 large model and fine-tuned on the training data. The two tied third best system by Tredence used a Clinical-Longformer and Biomedical-ROBERTA[38].

Table 3
Performance of the participating teams in the MEDIQA-Sum 2023 Subtask B on dialogue2note summarization.

team	run	aggregate_score	rank	code_status
SuryaKiran	run3	0.573	1	1
PULSAR	run2	0.569	2	1
PULSAR	run1	0.565	3	1
Tredence	run1	0.559	4	1
SuryaKiran	run2	0.559	5	1
SuryaKiran	run1	0.550	6	1
PULSAR	run3	0.538	7	1
HuskyScribe	run1	0.529	8	1
Tredence	run2	0.508	9	1
uetcorn	run1	0.481	10	1
uetcorn	run2	0.480	11	1
uetcorn	run3	0.479	12	1
SKKU-DSAIL	run1	0.461	13	1
SSNSheerinKavitha	run1	0.419	14	5
SSNSheerinKavitha	run2	0.419	14	5
SSNSheerinKavitha	run3	0.279	16	5

Subtask B primarily consisted of pre-trained sequence-to-sequence models fine-tuned on the training and validation sets. Eight out of 16 submissions used the gold standard section headers released from Subtask A. Teams used similar families of models as shown below.

pretrained model families

- T5: HuskyScribe, PULSAR, SSNSheerinKavitha, SuryaKiran
- bart: SKKU-DSAIL, Tredence, SSNSheerinKavitha, SuryaKiran, UETCorn
- llama: PULSAR

The SSNSheerinKavitha team also experimented with a rule-based extractive system by selecting dialogue sentences based on scores related to word frequencies. The UETCorn team experimented with a mixture of conditioned reading comprehension extraction, using hand-crafted section-specific queries with rule-based processing. The best performing system was an ensemble method by the SuryaKiran team fined-tuned several BioBART-V2-large LoRA models (fine-tuned on different training folds) with both the dialogue and section header as inputs[37]. The best summary was selected using a semantic similarity approach. The second and third ranked systems by PULSAR, used a FLAN-T5 model and a FLAN-T5 model additionally pre-trained using a MIMIC III note term extraction objective.

Table 4
Performance of the participating teams in the MEDIQA-Sum 2023 Subtask C on dialogue2note summarization, ranked by ROUGE1.

team	run	rouge1	rank	code_status
Tredence	run2	0.500	1	1
uetcorn	run2	0.498	2	1
uetcorn	run3	0.497	3	1
Tredence	run1	0.486	4	1
uetcorn	run1	0.485	5	1
HuskyScribe	run1	0.470	6	1
HuskyScribe	run2	0.318	7	1
PULSAR	run2	0.294	8	1
PULSAR	run1	0.276	9	1

Table 5
Performance of the participating teams in the MEDIQA-Sum 2023 Subtask C on dialogue2note summarization, ranked by aggregate score.

team	run	agg_score	rank	code_status
Tredence	run1	0.455	1	1
Tredence	run2	0.454	2	1
uetcorn	run3	0.444	3	1
uetcorn	run2	0.443	4	1
uetcorn	run1	0.441	5	1
HuskyScribe	run1	0.413	6	1
HuskyScribe	run2	0.396	7	1
PULSAR	run2	0.305	8	1
PULSAR	run1	0.247	9	1

Subtask C featured a diverse set of systems that used creative means to circumvent a low-resource generation problem. Specifically, Uetcorn, HuskyScribe, and Tredence all divided the problem into multiple parts. Firstly, relevant parts of the dialogue were grouped together as related to particular sections. Each team used a different method to achieve this; the UETCorn team identified relevant parts of dialogue for specific note section key points (e.g. “chief complaint” or “medications”), using a similarity function between dialogue sentences and a hand-crafted section-specific description; afterwards, several note generation strategies were used for each key point. HuskyScribe built a model classifying smaller dialogue exchanges into the same categories, while Tredence classified dialogues chunked by various window sizes. In the second step, grouped dialogue chunks were sent through a text generator to produce parts of the note. The use of pre-trained models such as BART/BioBART and FLAN-T5 for the generation was typical. The Uetcorn and Tredence team included some section/key-point specific questions as part of the generation input, e.g. (e.g. input: "question: {question} context: {conversation}", output: summary). The Uetcorn team also experimented with a reading comprehension answer extraction based on specially designed key point query (e.g. "names of medication used") and post-processing as in their Subtask B system. The HuskyScribe team additionally used Subtask A data to generate additional synthetic data for training. Finally, the completed note was assembled through concatenation and post-processing. Unlike the other three groups, the PULSAR team employed an end-to-end approach, experimenting with FLAN-T5 and llama models with additional data created using MTSamples data processed through GPT3.5.

6. Discussion and Conclusions

This year’s MEDIQA 2023 tasks, ACL ClinicalNLP MEDIQA-Chat Shared Tasks [5] and this ImageCLEF MEDIQA-Sum task, hosted similar problems on an overlapping dataset. A striking difference between the participants in this edition was that there were no GPT4 submissions. As GPT4 access requires a subscription, we can view the solutions from this evaluation lab as a whole to be constrained to only using open-source or free models and data.

In general, with the exception of the full-encounter task, scores in the two 2023 editions were comparable. Suggesting that many current off-the-self methods are still very competitive for classification and shorter generation tasks whereas longer generation may require more powerful and massive LLM. In MEDIQA-Chat Task A header generation scores were at 0.35-0.78 accuracy; the corresponding similar MEDIQA-Sum subtask A had a overlapping but larger range of 0.14-0.82 accuracy. The comparable MEDIQA-Sum subtask B was similar to MEDIQA-Chat subtask A snippet summarization with snippet summarization scores at a range of 0.37-0.58 aggregate score. In MEDIQA-Sum Subtask B snippet summarization, the scores were at 0.28-0.57 aggregate score; again with similar ranges. Finally the full-encounter task was MEDIQA-Chat in Task B, full-encounter generation ROUGE1 was at 0.28-0.61 and 0.21-0.65 for aggregate scoring. In this editions’ Subtask C, the ranges were at 0.28-0.50 ROUGE1 and 0.25-0.46 aggregate scoring; which were slighlty lower than those in MEDIQA-Chat.

Classic meeting summarization systems have split the generation in several steps including topic identification, extractive summarization, and then abstractive summarization. In the MEDIQA-Sum challenges many of our systems followed this motif. Such a split may be the result of past models’ abilities to perform narrow tasks, as well as size constraints. With the latest LLM models as shown in MEDIQA-Chat, it is clear LLM can now perform the end-to-end task competitively. However, recent work on GPT4 has shown that prompting for chain-of-thought reasoning, means multi-step generation may not be obsolete but may instead take a new form. We can track the progress of the field by continuing to benchmark on open datasets and shared tasks.

The results in the MEDIQA challenges are exciting, however there were limitations to this work. Although this is the largest source of both short and full-encounter dialogue2note generation datasets, the data here is relatively small and limited to a single institution with only a handful of content creators. There are many areas to further explore and expand. In terms of dataset expansion, we allude to at least three frontiers: (a) expanding to a larger content creation force which will enable more linguistic patterns and more transcript variations (including length); (b) incorporation of structured data as additional input (e.g. past labs and vitals) and output (e.g. orders); and (c) additional gold standard references, including multiple note references using the same note structure as well as additional gold standard summaries using a variety of note formats. For modeling, the challenges and learnings from our tasks point to needs in several exciting directions of research including increased attention to long-text and medical natural language generation evaluation methods, as well as studying performance of multi-modal generation and partial-inputs generation. We hope that these shared tasks are the small beginnings that will inspire further widespread study into automatic clinical note generation; and that these efforts can be translated into integrated technologies that may improve the quality and outcomes for both doctors and patients.

References

[1] B. D. Tran, Y. Chen, S. Liu, K. Zheng, How does medical scribes' work inform development of speech-based clinical documentation technologies? a systematic review, Journal of the American Medical Informatics Association : JAMIA (2020).

[2] J. C. Quiroz, L. Laranjo, A. B. Kocaballi, S. Berkovsky, D. Rezazadegan, E. W. Coiera, Challenges of developing a digital scribe to reduce clinical documentation burden, NPJ Digital Medicine 2 (2019).

[3] B. Ionescu, H. Müller, A. Drăgulinescu, W. Yim, A. Ben Abacha, G. A. Neal Snider, M. Yetisgen, J. Rückert, A. Garcıa Seco de Herrera, C. M. Friedrich, L. Bloch, A. I. Raphael Brüngel, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, socialmedia and recommender systems applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer Lecture Notes in Computer Science LNCS, Thessaloniki, Greece, 2023.

[4] A. Ben Abacha, C. Shivade, D. Demner-Fushman, 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, Association for Computational Linguistics, 2019, pp. 370–379. URL: https://doi.org/10.18653/v1/w19-5039.

[5] A. Ben Abacha, W. Yim, G. Adams, N. Snider, M. Yetisgen, Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations, in: ACL-ClinicalNLP 2023, 2023.

[6] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. A. Bates, D. Jurafsky, P. A. Taylor, R. Martin, C. V. Ess-Dykema, M. W. Meteer, Dialogue act modeling for automatic tagging and recognition of conversational speech, Computational Linguistics 26 (2000) 339–373.

[7] P. Budzianowski, T.-H. Wen, B.-H. Tseng, I. Casanueva, S. Ultes, O. Ramadan, M. Gasic, Multiwoz - a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling, in: Conference on Empirical Methods in Natural Language Processing, 2018.

[8] D. Jurafsky, J. H. Martin, Speech and language processing, 3rd edition, 2023.

[9] J. Liu, Y. Zou, H. Zhang, H. Chen, Z. Ding, C. Yuan, X. Wang, Topic-aware contrastive learning for abstractive dialogue summarization, in: Conference on Empirical Methods in Natural Language Processing, 2021.

[10] L. Zhu, G. Pergola, L. Gui, D. Zhou, Y. He, Topic-driven and knowledge-aware transformer for dialogue emotion detection, in: Annual Meeting of the Association for Computational Linguistics, 2021.

[11] N. Wang, Y. Song, F. Xia, Studying challenges in medical conversation with structured annotation, in: Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, Association for Computational Linguistics, Online, 2020, pp. 12–21. URL: https://aclanthology.org/2020.nlpmc-1.3. doi:10.18653/v1/2020.nlpmc-1.3.

[12] B. Schloss, S. Konam, Towards an automated soap note: Classifying utterances from medical conversations, in: Machine Learning in Health Care, 2020.

[13] J. C. Denny, R. A. Miller, K. B. Johnson, A. Spickard, Development and evaluation of a clinical note section header terminology, AMIA ... Annual Symposium proceedings. AMIA Symposium (2008) 156–60.

[14] M. Tepper, D. Capurro, F. Xia, L. Vanderwende, M. Yetisgen-Yildiz, Statistical section segmentation in free-text clinical records, in: International Conference on Language Resources and Evaluation, 2012.

[15] P. Landes, K. Patel, S. S. Huang, A. Webb, B. Di Eugenio, C. Caragea, A new public corpus for clinical section identification: MedSecId, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 3709–3721. URL: https://aclanthology.org/2022.coling-1.326.

[16] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1073–1083. URL: https://aclanthology.org/P17-1099. doi:10.18653/v1/P17-1099.

[17] S. Narayan, S. B. Cohen, M. Lapata, Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 1797–1807. URL:

https://aclanthology.org/D18-1206. doi:10.18653/v1/D18-1206.

[18] G. Zeng, W. Yang, Z. Ju, Y. Yang, S. Wang, R. Zhang, M. Zhou, J. Zeng, X. Dong, R. Zhang, H. Fang, P. Zhu, S. Chen, P. Xie, MedDialog: Large-scale medical dialogue datasets, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 9241–9250. URL: https://aclanthology.org/2020.emnlp-main.743. doi:10.18653/v1/2020.emnlp-main.743.

[19] B. Gliwa, I. Mochol, M. Biesek, A. Wawer, SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 70–79. URL: https://aclanthology.org/D19-5409. doi:10.18653/v1/D19-5409.

[20] W. Yim, M. Yetisgen, Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization, in: Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, Association for Computational Linguistics, Online, 2021, pp. 10–20. URL: https://aclanthology.org/2021.nlpmc-1.2. doi:10.18653/v1/2021.nlpmc-1.2.

[21] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, P. D. Wellner, The ami meeting corpus, 2005.

[22] A. L. Janin, D. Baron, J. Edwards, D. P. W. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, C. Wooters, The icsi meeting corpus, 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 1 (2003) I–I.

[23] B. Pang, E. Nijkamp, W. Kryscinski, S. Savarese, Y. Zhou, C. Xiong, Long document summarization with top-down and bottom-up inference, in: Findings, 2022.

[24] H. Y. Koh, J. Ju, M. Liu, S. Pan, An empirical survey on long document summarization: Datasets, models, and metrics, ACM Computing Surveys 55 (2022) 1 – 35.

[25] C. Grambow, L. Zhang, T. Schaaf, In-domain pre-training improves clinical note generation from doctor-patient conversations, in: Proceedings of the First Workshop on Natural Language Generation in Healthcare, Association for Computational Linguistics, Waterville, Maine, USA and virtual meeting, 2022, pp. 9–22. URL: https://aclanthology.org/2022.nlg4health-1.2.

[26] S. Enarvi, M. Amoia, M. Del-Agua Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, L. Rubini, M. Ruiz, G. Singh, F. Stemmer, W. Sun, P. Vozila, T. Lin, R. Ramamurthy, Generating medical reports from patient-doctor conversations using sequence-to-sequence models, in: Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, Association for Computational Linguistics, Online, 2020, pp. 22–30. URL: https://aclanthology.org/2020.nlpmc-1.4. doi:10.18653/v1/2020.nlpmc-1.4.

[27] A. Ben Abacha, W. Yim, Y. Fan, T. Lin, An empirical study of clinical note generation from doctor-patient encounters, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 2291–2302. URL: https://aclanthology.org/2023.eacl-main.168.

[28] W. Yim, Y. Fu, A. Ben Abacha, N. Snider, T. Lin, M. Yetisgen, Aci-bench: a novel ambi-

ent clinical intelligence dataset for benchmarking automatic visit note generation, 2023. arXiv:2306.02022.

[29] A. Ben Abacha, W.-w. Yim, G. Michalopoulos, T. Lin, An investigation of evaluation methods in automatic medical note generation, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2575–2588. URL: https://aclanthology.org/2023.findings-acl.161.

[30] A. Sharma, D. I. Feldman, Team cadence at mediqa-sum 2023: Using chatgpt as a data augmentation tool for classifying clinical dialogue, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[31] B. Han, H. Zhu, S. Zhou, S. Ahmed, M. Rahman, K. Lybarger, F. Xia, Huskyscribe at mediqa-sum 2023: Summarizing clinical dialogues with transformers, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[32] H. S. Palaniraj, K. Vinod, M. Adluru, B. Jayaraman, M. Tt, Mlrg-jbttm at mediqa-sum 2023: Dialogue2topic classification, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[33] V. Schlegel, H. Li, Y. Wu, A. Subramanian, T.-T. Nguyen, A. R. Kashyap, D. Beck, X. Zeng, R. T. Batista-Navarro, S. Winkler, G. Nenadic, Pulsar at imageclef 2023 medisum: Large language models augmented by synthetic dialogue convert patient dialogues to medical records, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[34] D. Krishnan, D. Srinivasan, K. Srinivasan, Ssndhanyadivyakavitha at mediqa-sum 2023: Medical dialogue summarization using linear support vector classification technique, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[35] S. S. N. Mohamed, K. Srinivasan, Ssn mlrg at mediqa-sum 2023: Automatic text summarization using support vector machine and roberta, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[36] C.-Y. Chang, J. Li, S. Kumar, V. V. Vydiswaran, Stellellastars at mediqa-sum 2023: Exploring transformer-based models for dialogue2topic classification, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[37] K. Suri, P. Mishra, S. Saha, A. Singh, Suryakiran at mediqa-sum 2023: Leveraging lora for clinical dialogue summarization, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[38] V. Adwani, M. S. Khan, A. Chopra, Tredence at mediqa-sum 2023: Clinical note generation from doctor patient conversation using utterance segmentation and question-answer driven abstractive summarization, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[39] D.-C. Can, Q.-A. Nguyen, B.-N. Nguyen, M.-Q. Nguyen, K.-V. Nguyen, T.-H. Do, H.-Q. Le, Uetcorn at mediqa-sum 2023: Template-based summarization for clinical note generation from doctor-patient conversation, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

7. Acknowledgments

We would like to thank Thomas Lin from Microsoft Health AI and the ImageCLEF organizers for their feedback and support for the MEDIQA-Sum 2023 shared tasks. We also thank our annotation team for preparing the data in time for the challenge and all the participating teams who contributed to the success of these shared tasks through their interesting approaches and experiments and strong engagement.