TREDENCE at MEDIQA-Sum 2023: Clinical Note Generation from Doctor Patient Conversation using Utterance Segmentation and Question-Answer Driven Abstractive Summarization

Notebook for the ImageCLEF Lab at CLEF 2023

Vaibhav Adwani [1], Mohammed Sameer Khan [1] and Ankush Chopra [1]

[1] *Tredence, Whitefield, Bengaluru, India*

Abstract

This paper describes our submission to MEDIQA-Sum [2] shared task for automatic classification and summarization of doctor-patient conversations. These doctor-patient transcripts present many challenges: limited training data, significant domain shifts and noisy transcripts. Here in this paper, we explore the possibility of using pretrained transformer models to automatically summarize and generate clinical note from full doctor-patient transcript. We propose a two-step approach, first step is the use of pretrained encoder models like Biomedical-ROBERTA [3] to get dialogues belonging to similar category (section headers) together from the full transcript, these are known as conversation snippets (parts of full conversation) belonging to some section header like chief-complaint, to achieve this we propose two methods, Fixed window size and Variable window size, both these methods proved to be effective in getting good conversation snippets from the full conversation. In the second step, we summarize these conversation snippets, for this, we propose a QA-based summarization approach using transformer-based summarization models where we pass questions corresponding to the conversation category along with the conversation to make the model focus on important aspects of the conversation and then output the summary in the form of an answer.

Keywords

BERT, ROBERTA, BART, PEGASUS, T5, MEDIQA-Sum, ImageCLEF

1. Introduction

In recent years, the widespread adoption of electronic health records (EHRs) has generated an enormous amount of textual data, including doctor-patient conversations. These conversations contain valuable insights and critical information about patients' medical conditions, diagnoses, treatments, and outcomes. MEDIQA-Sum [2] focuses on the automatic summarization and classification of doctor-patient conversations through three subtasks. Effectively classifying full conversation is the primary step towards achieving good clinical notes, for this, we propose a window size-based approach using transformer-based classification models pretrained on medical data. The next step is the generation of good-quality summaries, for this, we propose the use of abstractive summarization models pretrained on medical data using Question-Answering (QA) approach. By using QA based approach, we frame the summarization task as an answer generation process, where questions correspond to important aspects of the conversation, and the answers are the abstractive summaries themselves.

[1]CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
EMAIL: vaibhav.adwani@tredence.com (V. Adwani); mohammed.sameerkhan@tredence.com (M.S. Khan); ankush.chopra@tredence.com (A. Chopra)
ORCID: 0009-0009-5225-2641 (V. Adwani); 0009-0006-9706-5602 (M.S. Khan); 0000-0002-9970-8038 (A. Chopra)

CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Works

Summarization is a well-known problem in NLP however recent years have seen big improvements in the field with models like BART [7], T5 [17] etc. Zhang et al. (2021) [15] finetune a BART-Large model to summarize conversations belonging to distinct sections such as History of Present Illness, etc. Similar models can also be used for clinical note generation. Krishna et al. (2020) [16] use multilabel B-LSTM to identify sentences belonging to different subsections like Allergy, Chief complaint etc. and then cluster all sentences belonging to each subsection together and then the use T5-Base [17] model to summarize the sentence clusters of each subsection.

3. Methods

In this section, we describe various tasks, data used, methodologies adopted, and experimentations carried out in ImageCLEF 2023 [1] MEDIQA-Sum [2] competition.

3.1 Task-A

This task [2] aims to identify the topic (section header) of the conversation between the doctor and patient. The section header will be one of the twenty normalized common section labels [E].

3.1.1 Dataset

We used Dialogue to Topic Classification dataset provided by MEDIQA-Sum organizers [2], dataset consisted of conversation ids, conversation snippets and section headers for each snippet. Train, validation, and test set consisted of 1200, 100 and 200 instances respectively.

Dataset Description & Preprocessing: We truncated conversations exceeding the sequence limit of 512 to 512 because of the sequence length limit (512) for base version BERT [11] and ROBERTA [10].

Table 1:
Corpus statistics for Task-A dataset

	Max	Min	Average	Std	97%	95%
Conversation length	1509	6	105	117	392	339
Tokenized conv. length (ROBERTA)	2417	14	169	183	611	517
Tokenized conv. length (BERT)	2218	11	147	162	526	463

3.1.2 Method

To solve this task, we have used the transformer-based models from hugging face [9] pretrained on biomedical and clinical data like PubMed [D] and MIMIC III [D]. We further finetuned these models to achieve state-of-the-art performance. To deal with long conversations we have also tried a long sequence encoder model pretrained on clinical data and then further finetuned it on our dataset.

3.1.3 Experimentations

We performed three experiments (runs) using three different models.

3.1.3.1 Run – 1

In this run, we used Biomedical-ROBERTA [3], base version, pretrained on Biomedical data from Semantic Scholar [D]. The model was finetuned for conversation classification on train and validation data together by adding a classification head consisting of two linear layers and a 0.1 dropout layer in between two linear layers.

3.1.3.2 Run-2

In this run, we used Clinical-Longformer [4] model pretrained using MIMIC-III [D] clinical notes. The Longformer model has a sequence limit of 4096 tokens, this helped us to train the model on full conversations even if the conversation length is large. We then finetuned the model for conversation classification on train and validation data together by adding a classification head consisting of two linear layers and a 0.1 dropout layer in between two linear layers.

3.1.3.3 Run-3

In this run, we used Bio-Clinical BERT [5] model pretrained on all notes from MIMIC III [D]. We further finetuned the models on train and validation data together for final submission by adding two linear layers and a 0.1 dropout layer between the two linear layers on top of Bio-Clinical BERT.

Table 2:
Hyperparameters for Run-1, Run-2 & Run-3 (Task-A) found after multiple rounds of tuning.

	Learning Rate	Train Batch Size	Epochs	L.R schedular	Optimizer	Weight decay	Warmup Steps	Gradient Accumulation Steps
Run-1	5e-5	20	11	Linear	Adam	0.01	100	1
Run 2	5e-5	2	13	Linear	Adam	0.01	100	5
Run-3	5e-5	20	5	Linear	Adam	0.01	100	1

3.2 Task-B

This task [2] aims to summarize the conversation snippet between the doctor and the patient. The conversation belongs to a particular topic identified by one of the 20 section headers. The summary contains abstract information related to the specified section header.

3.2.1 Dataset

Dataset is the same as provided in Task-A but contains one additional column called section text containing summaries for every conversation snippet.

Dataset Description & Preprocessing: The important step in preprocessing is to prepare inputs for the model to finetune models using Question-Answering based approach as described in section 3.2.2, also because of resource constraints during training, we have truncated the input sequence length to the model to 400 in every experiment.

Table 3: Corpus statistics for Task-B dataset

	Max	Min	Average	Std
Tokenized conversation length	2417	14	169	182
Tokenized summary length	1450	2	55	88

3.2.2 Method

To solve this summarization task, we used transformer models from hugging face [9] pretrained on biomedical or clinical data with pretraining objective of abstractive summarization. We used Question-Answering (QA) based approach to finetune these models on our dataset for abstractive summarization. Although this is fundamentally a summarization task, but we observed that summaries of different sections had different styles. We therefore modelled this as a QA task where the question is used to inculcate the distinct style of a particular section while summarization.

To train a QA model we require question, context, and an answer. In this use case, answer is the expected output summary, context is the conversation snippet to be summarized and the question is a kind of instruction given to the model to look for specific information in the conversation during summarization.

Figure 1 explains how models were finetuned for QA-based summarization [12] by passing the question corresponding to the section header along with the conversation snippet as input to the model. The conversation shown in the below figure belongs to CHIEF COMPLAINT [CC] section header, we picked up question corresponding to CHIEF COMPLAINT and passed it along with the conversation to tell the model what to summarize, inputs to the model are passed in the specific format, as shown in figure 1, both during finetuning as well as during inference.

Questions for each section header were empirically selected after multiple experimentations, please refer Appendix section for a full table containing questions for every section header.

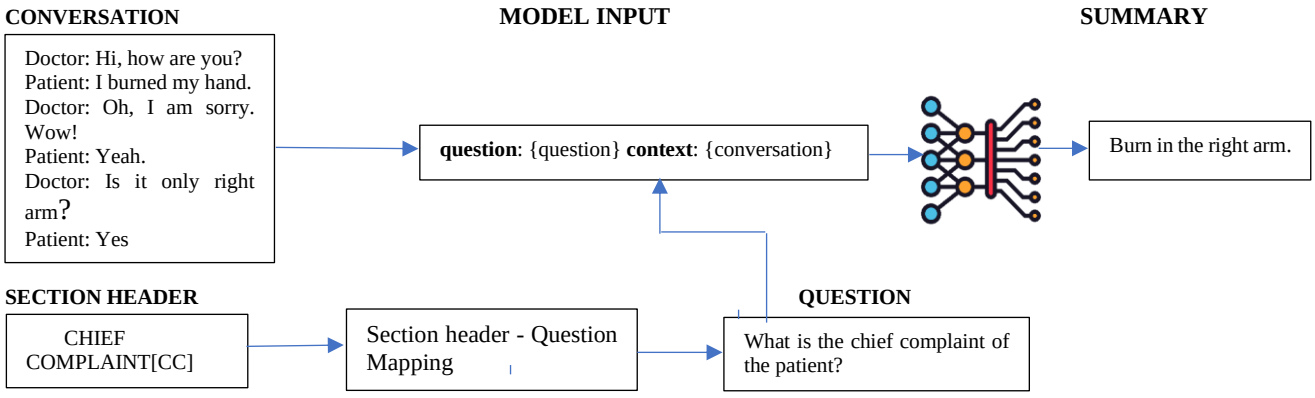

Figure 1: Figure showing the way models were finetuned for QA Task to generate summaries.

3.2.3 Experimentations

We experimented with multiple summarization models. Two runs have been submitted for evaluation, other experiments which were not submitted are described under Other Experimentations section.

3.2.3.1 Run-1

In this run, we have used BART-large [7] model pretrained on biomedical data (PubMed). We finetuned the model using Question-Answering based approach by adding the language modelling head on top of BART-large [7]. The model was finetuned on train and validation data together for final submission.

3.2.3.2 Run-2

In this run, we have used BART-base [8] model pretrained on biomedical data. To finetune model for the summary generation we added language modelling on top of BART-base [8] and finetuned it

using the Question-Answering approach, as described in section 3.2.2, on train and validation data together for final submission.

Table 4:
Hyperparameters for Run-1 and Run-2 (Task-B) found after multiple rounds of hyperparameter tuning

Learning Rate	Train Batch Size	Epochs	L.R Scheduler	Optimizer	Weight decay	Warmup Steps	Gradient Accumulation Steps
5e-5	5	4	Linear	Adam	0.01	60	4

3.2.3.3 Other Experimentations

We tried two abstractive summarization models, PEGAUS [14] and T5 [17] base version, that were already finetuned for QA-based summarization but not on medical data, we have also tried a T5 [17] base version that was finetuned on BIOASQ [D] dataset for QA based summarization. We further finetuned each of these models, using the QA approach by adding a language modelling head on top of the base model.

Table 5:
Hyperparameters for experimented models

	Learning Rate	Train Batch Size	Epochs	L.R Scheduler	Optimizer	Weight decay	Warmup Steps	Gradient Accumulation Steps
PEGASUS	5e-5	5	15	Linear	Adam	0.01	10	4
Valhalla T5	5e-5	5	15	Linear	Adam	0.01	10	4
Bio-T5	5e-5	5	17	Linear	Adam	0.01	10	4

3.3 Task-C

This task [2] aims to generate a full clinical note summarizing the full encounter conversation between the doctor and patient. Clinical notes should contain abstract information about every topic (section header) being discussed in the full conversation.

3.3.1 Dataset

In our experiments, we used Full-Encounter Dialogue to Note Summarization dataset provided by MEDIQA-Sum organizers [2] consisting of full conversation between doctor and patient along with the clinical note. Train, validation, and test dataset consisted of 67, 20 and 40 instances respectively.

Dataset Description & Preprocessing: Data was mostly clean and didn't require much preprocessing, apart from removing extra spaces, we formatted [doctor] and [patient] to doctor: and patient: because data on which classification and summarization model was trained, contained this representation.

Table 6:
Corpus statistics for Task-C dataset

	Max	Min	Average	Std
Full conversation length	3050	628	1301	412
Clinical note length	884	135	420	129

3.3.2 Method

To create clinical note, we require two things, first is the conversation snippets pertaining to one of the available section headers in the full conversation and second is the summary of these conversation snippets. For summarization we have used best performing summarization model from Task-B. Since we have full conversation with us, we ran classification model from Task-A on parts of full conversation to get conversation snippets belonging to available section headers in the full conversation. After trying various approaches, we found 2 approaches that worked best for us in getting better conversation snippets from the full conversation. These approaches are discussed below.

Variable Window Size: In the variable window size approach, we require a fixed window size to start with. We keep on extending this window according to different conditions as mentioned in the below steps, but the minimum window size in this approach remains fixed.

Empirically minimum window size of 2 worked best for us. Once we have this minimum value, we start with the first utterance(dialogue) of the conversation to create starting conversation snippet containing 2(window size) dialogues. Once we have the starting conversation snippet, we followed the below steps to implement the logic.

1. Classify the conversation snippet and keep a note of the predicted section header and its classification probability.
2. Add another utterance (dialogue) to the present window(snippet) and then classify this new conversation snippet, in this step also keep a note of the predicted section header and its classification probability.
3. If the predicted section header from Step 1 and Step 2 are the same and the classification probability for this section header increase or remains the same in Step 2, keep this utterance (dialogue) added in Step 2 to the present conversation snippet and repeat step no. 2.
4. If the predicted section header from step 1 and step 2 are different or the predicted section header is the same but the classification probability in step 2 decreases after adding the utterance to the present window(snippet), then we drop the utterance(dialogue) added to the window and consider this snippet to belong to section header which was predicted before adding this new utterance(dialogue). After this, we create a new window of fixed size, as decided earlier, from the next utterance of the previous window and then repeat the same steps from 1 to 4.
5. Continue steps 1 to 4 until the full conversation is covered, at the end concatenate conversations belonging to the same section header together in the same order in which they were obtained from different parts of the full conversation.

Fixed Window Size: To implement this approach we need to have a fixed window size, after evaluating model performance using different window sizes, we empirically selected a window of size two. We followed the below steps to implement fixed window size logic.

1. Using a fixed window of a particular size, select the conversation snippet starting from the first utterance(dialogue) containing utterances which are equal to the window size in number and then classify it.
2. Use the same window to select another conversation snippet from the next dialogue of the previous window and then classify it, Repeat the process until the full conversation is covered.
3. Concatenate conversation snippets belonging to the same section header together in the same order in which they were obtained from different parts of the full conversation.

3.3.3 Experimentations

We performed two experiments implementing variable window size and fixed window size approach as discussed in section 3.3.2. In each experiment window size is the hyperparameter.

3.3.3.1 Run-1

In this run, we finetuned the best performing classification model from Task-A, biomed-ROBERTA [3], on train, validation, and test set of Task-A and used it to get conversation snippets belonging to different section headers from full conversation using fixed window size approach as discussed in section 3.2.2. To summarize these snippets, we used the best performing summarizing model from Task-B, BART-large [7]. Summarization is done using QA based approach.

3.3.3.2 Run-2

Here we used the Variable window size approach to get conversation snippets belonging to available section headers in the full conversation. Classification and summarization models as well as the approach remain the same as used in run 1.

4. Results

We present the results of our experimentations on train, validation, and test sets. We have test set results of only submitted runs since test sets for Task-B and Task-C were not made available. Our submission results and metrics used by the organizers for evaluation are highlighted for each task.

Note: Test set results are of models trained on train and validation set together.

4.1 Task-A

Dataset provided was highly skewed, 9 classes or categories had less than 15 examples, because of which models performed poorly in identifying these categories and overall accuracy is a bit low. For this task, our best performing runs (run-1 and run-2) were ranked 3rd among 23 submitted runs.

Table 7:
Task-A results

	Train-Accuracy (%)	Validation-Accuracy (%)	**Test-Accuracy** (%)
Biomedical ROBERTA (Run-1)	99.83	80	**80**
Clinical Longformer (Run-2)	99.92	80	**80**
Clinical BERT (Run-3)	95.75	77	**75.5**

4.2 Task-B

QA-based approach significantly improved the performance by making summaries more concise, improving the Rouge-1 score. Models pretrained or finetuned on domain data (Bio-BART) performed better as compared to those pretrained on general data (PEGASIS-QA, Valhalla-T5). For this task, our best performing run (run-1) was ranked 4th among 16 submitted runs.

Table 8:
Task-B results

		Rouge-1	Bleurt	BERT-Score			**Aggregate score**
				Precision	Recall	F1	
Bio-BART-Large (Run-1)	Train	0.51	0.60	0.78	0.75	0.76	0.62
	Validation	0.41	0.54	0.74	0.71	0.72	0.56
	Test	0.42	0.53	0.74	0.71	0.72	**0.56**

Bio-BART-Base (Run-2)	Train	0.46	0.55	0.77	0.72	0.74	0.58
	Validation	0.42	0.52	0.74	0.69	0.71	0.55
	Test	0.36	0.47	0.72	0.67	0.69	**0.51**
PEGASUS-QA	Train	0.38	0.52	0.72	0.68	0.69	0.53
	Validation	0.38	0.51	0.72	0.67	0.69	0.53
Valhalla-T5	Train	0.35	0.51	0.68	0.67	0.67	0.51
	Validation	0.36	0.49	0.71	0.66	0.68	0.51
Bio-T5	Train	0.33	0.49	0.65	0.64	0.64	0.49
	Validation	0.34	0.48	0.67	0.66	0.66	0.49

4.3 TASK-C

Performance of fixed window size approach significantly depend on window size as window size remains fixed throughout the experiment, variable window on the other hand does not depend much on window size because window size is dynamic throughout the experiment. To be on the safer side, lower window size can be chosen as the initial value. For this task, our best run (run-2) was the overall best run of the competition and run-2 was second best run of the competition.

Table 9:
Task-C results

		Rouge-1	Rouge-2	Rouge-L	Rouge-Lsum
Fixed Window size (Run-1)	Train	0.51	0.22	0.26	0.46
	Validation	0.52	0.21	0.26	0.45
	Test	**0.49**	0.19	0.24	0.44
Variable Window size (Run-2)	Train	0.51	0.21	0.26	0.46
	Validation	0.52	0.22	0.26	0.46
	Test	**0.50**	0.20	0.24	0.45

5. Conclusions & Future Work

In this paper, we present systems developed by the TREDENCE team for MEDIQA-Sum [2] task. The task focused on the automatic classification and summarization of doctor-patient conversations. To address this, we used pretrained transformer models for classifying full conversation into conversation snippets using the variable-window side and fixed-window size approaches. Both these approaches performed reasonably well to give good quality snippets. For summarizing the snippets, we used a QA-based approach to instruct the model to give summaries in the form of answers to questions. QA. Our experimentations during summarization reveal that instruction-based learning can significantly improve performance.

In future, we are looking at incorporating Data-Augmentation to improve classification performance. Secondly, we used the summarization model from Task-B to summarize conversation snippets from Task-C, but we found some differences between expected summaries for both these tasks in terms of elaborateness and presentation, we are working towards finetuning our summarization models on Task-C dataset to further improve the performance.

6. References

[1] Bogdan Ionescu, Henning M\"uller, Ana-Maria Dr\u{a}gulinescu, Wen{-}wai Yim, Asma {Ben Abacha}, Neal Snider, Griffin Adams, Meliha Yetisgen, Johannes R\"uckert, Alba Garc\'{\i}a Seco de Herrera, Christoph M. Friedrich, Louise Bloch, Raphael Br\"ungel, Ahmad Idrissi-Yaghir, Henning Sch\"afer, Steven A. Hicks, Michael A. Riegler, Vajira Thambawita, Andrea Storås, Pål Halvorsen, Nikolaos Papachrysos, Johanna Schöler, Debesh Jha, Alexandra-Georgiana Andrei, Ihar Filipovich, Ioan Coman, Vassili Kovalev, Alexandru Stan, George Ioannidis, Hugo Manguinhas, Liviu-Daniel \c{S}tefan, Mihai Gabriel Constantin, Mihai Dogariu, J\'er\^ome Deshayes, Adrian Popescu, {Overview of ImageCLEF 2023}: Multimedia Retrieval in Medical, Social Media and Recommender Systems Applications in use: Experimental IR Meets Multilinguality, Multimodality, and Interaction, in: Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), 2023, Springer Lecture Notes in Computer Science LNCS, September 18-21, Thessaloniki, Greece.

[2] Wen{-}wai Yim and Asma {Ben Abacha} and Neal Snider and Griffin Adams and Meliha Yetisgen, Overview of the MEDIQA-Sum Task at ImageCLEF 2023: Summarization and Classification of Doctor-Patient Conversations in use: CLEF 2023 Working Notes, in: {CEUR} Workshop Proceedings, 2023, CEUR-WS.org, September 18-21, Thessaloniki, Greece.

[3] Suchin Gururangan and Ana Marasović and Swabha Swayamdipta and Kyle Lo and Iz Beltagy and Doug Downey and Noah A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, 2020, Proceedings of ACL.

[4] Li, Yikuan and Wehbe, Ramsey M and Ahmad, Faraz S and Wang, Hanyin and Luo, Yuan, A comparative study of pretrained language models for long clinical text, Journal of the American Medical Informatics Association 30 (2023) 340-347, Oxford University Press.

[5] Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, Matthew B. A. McDermott, Publicly Available Clinical BERT Embeddings. URL: https://doi.org/10.48550/arXiv.1904.03323

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining. URL: https://doi.org/10.48550/arXiv.1901.08746

[7] Hongyi Yuan and Zheng Yuan and Ruyi Gan and Jiaxing Zhang and Yutao Xie and Sheng Yu, BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model, 2022. URL: https://doi.org/10.48550/arXiv.2204.03905.

[8] Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, Sheng Yu, BioBART: Pretraining and Evaluation of A Biomedical Generative Language Model, 2022. URL: https://doi.org/10.48550/arXiv.2204.03905

[9] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush, HuggingFace's Transformers: State-of-the-art Natural Language Processing. https://doi.org/10.48550/arXiv.1910.03771.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach. URL: https://doi.org/10.48550/arXiv.1907.11692.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. URL: https://doi.org/10.48550/arXiv.1810.04805.

[12] Ping Chen, Fei Wu, Tong Wang, Wei Ding, A Semantic QA-Based Approach for Text Summarization Evaluation. URL: https://arxiv.org/ftp/arxiv/papers/1704/1704.06259.pdf.

[13] Iz Beltagy, Matthew E. Peters, Arman Cohan, Longformer: The Long-Document Transformer. URL: https://doi.org/10.48550/arXiv.2004.05150.

[14] Jingqing Zhang, Yao Zhao, Mohammad Saleh, Peter J. Liu, PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. URL: https://doi.org/10.48550/arXiv.1912.08777.

[15] Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, Matthew R. Gormley, Leveraging Pretrained Models for Automatic Summarization of Doctor-Patient Conversations URL: https://arxiv.org/abs/2109.12174

[16] Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary Lipton, Generating SOAP notes from Doctor-Patient Conversations Using Modular Summarization Techniques URL: https://arxiv.org/abs/2005.01795.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu, Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. URL: https://doi.org/10.48550/arXiv.1910.10683.

A. Question- Section header Mapping

Table 10 represent Questions corresponding to every possible section headers of the conversation snippet.

Table 10: Table representing mapping.

Section Header	Question
FAM/SOCHX [FAMILY/SOCIAL HISORY]	what all things patient mentions about his/her family or his/her social life?
GENHX [HISTORY OF PRESENT ILLNESS]	what all problems did patient mentions to the doctor?
PASTMEDICALHX [PAST MEDICAL HISTORY]	what is the past medical history of the patient?
CC [CHIF COMPLAINT]	what is the chief complaint of the patient?
ALLERGY	what allergies does patient mentions?
ROS [REVIEW OF SYSTEMS]	what review is done by the doctor on the patient?
PASTSURGICAL [PAST SURGICAL HISTORY]	what is the past surgical history of the patient?
MEDICATIONS	what medications are being discussed in the given conversation?
ASSESSMENT	what assessment did doctor do about the patient?
EXAM	what is the result of the exam carried out by the doctor on the patient?
DIAGNOSIS	doctor diagnosed patient with what disease?
DISPOSITION	what is the disposition status of the patient?
PLAN	what is the plan that patient has to follow?
EDCOURSE [EMERGENCY DEPARTMENT COURSE]	what is the condition of the patient in the emergency department?
IMMUNIZATIONS	what is the status of patient's immunization or vaccinations?
IMAGING	what is the result of patient's imaging report?
GYNHX [GYNECOLOGIC HISTORY]	what is the gynecologic history of the patient?
PROCEDURES	what procedures or surgeries did patient had?
OTHER_HISTORY	what is the other history of the patient mentioned in the conversation?
LABS	what are the findings from patient's lab report?

B. Hugging Face Models References for Task-A

- Biomedical- ROBERTA: https://huggingface.co/allenai/biomed_roberta_base (Run-1)
- Clinical-LongFormer: https://huggingface.co/yikuan8/Clinical-Longformer (Run-2)
- Clinical-BERT: https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT (Run-3)

C. Hugging Face Models References for Task-B

- Valhalla T5 model: https://huggingface.co/valhalla/t5-base-qa-qg-hl
- Pegasus-QA model: https://huggingface.co/tuner007/pegasus_qa
- Bio-T5 model: https://huggingface.co/ozcangundes/T5-base-for-BioQA

- Bio-BART base: https://huggingface.co/suryakiran786/5-fold-stratified-cv-biobart-v2-base-with-section-description-complete-data-1 (Run-1)
- Bio-BART large: https://huggingface.co/GanjinZero/biobart-large (Run-2)

D. Important Links

- BIOASQ website: http://participants-area.bioasq.org/datasets/
- PubMed website: https://pubmed.ncbi.nlm.nih.gov/
- MIMIC III database: https://www.nature.com/articles/sdata201635
- Semantic scholar: https://www.semanticscholar.org/

E. SECTION LABELS FOR TASK A & B

1. fam/sochx [FAMILY HISTORY/SOCIAL HISTORY]
2. genhx [HISTORY of PRESENT ILLNESS]
3. pastmedicalhx [PAST MEDICAL HISTORY]
4. cc [CHIEF COMPLAINT]
5. pastsurgical [PAST SURGICAL HISTORY]
6. allergy
7. ros [REVIEW OF SYSTEMS]
8. medications
9. assessment
10. exam
11. diagnosis
12. disposition
13. plan
14. edcourse [EMERGENCY DEPARTMENT COURSE]
15. immunizations
16. imaging
17. gynhx [GYNECOLOGIC HISTORY]
18. procedures
19. other_history
20. labs