HuskyScribe at MEDIQA-Sum 2023: Summarizing Clinical Dialogues with Transformers

Bin Han[1,*,†], Haotian Zhu[1,†], Sitong Zhou[1,†], Sofia Ahmed[1], Md. Mushfiqur Rahman[2], Fei Xia[1] and Kevin Lybarger[2]

[1] *University of Washington*
[2] *George Mason University*

Abstract

Documentation burnout contributes to clinician job dissatisfaction, and clinical notes often omit salient information. The automatic generation of notes from doctor-patient conversations using a computerized medical scribe, referred to as a Digital Scribe, provides an alternative, potentially time-saving documentation process. Generating notes from the transcribed patient interviews requires reorganizing utterances by topical note sections, identifying the clinically significant information, and generating a medical language summary. The MEDIQA-Sum task of ImageCLEF 2023 explores the development of Digital Scribe through the generation of clinical note summaries of transcribed patient visits. We participated in all three Subtasks and made contributions related to note subsection classification and dialogue-note alignment. We achieved high classification accuracy (81.5%) for Subtask A by fine-tuning T5-large, which ranked 2ND among 10 participants. We explored the capabilities of state-of-the-art large language models in the Subtask B summarization task. For Subtask C, we manually annotated the alignment between dialogue transcripts and clinical notes for a subset of training examples, to assist in learning the mapping from dialogue content to clinical note subsection.

Keywords

Digital Scribe, Dialogue Summarization, Transformers, Clinical NLP

1. Introduction

Clinicians spend approximately half of their time interacting with the Electronic Health Record (EHR) and performing other desk work [1, 2, 3]. A majority of this desktop medicine involves creating clinical notes, which accounts for over a third of clinician time [3]. This documentation burden contributes to clinician job dissatisfaction, burnout, and attrition [4, 5]. Additionally, clinical notes frequently contain errors and omit clinically significant information [6, 7], since most clinical notes are manually typed by clinicians [5]. Dictation using automatic speech recognition is successfully implemented in domains, like radiology, where the note content is not derived from a patient interview. Medical scribes reduce clinician documentation time, as a majority of the note is derived from the patient interview [8]; however, medical scribes are cost prohibitive in most healthcare settings [9]. A *Digital Scribe* is a computerized medical scribe that automatically generates clinical notes from the the patient

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece
[*] Corresponding author.
[†] These authors contributed equally.

visit by transcribing the dialogue and summarizing visit [10]. Digital Scribe technology may offer similar benefits as medical scribes at a reduced cost [11, 12]. Generating notes from the transcribed patient interviews requires reorganizing utterances by topical note sections, identifying the clinically significant information, and generating a medical language summary.

The MEDIQA-Sum task [13] of ImageCLEF 2023 [14] explores Digital Scribe development through the three Subtasks (A, B, and C) presented in Table 2. We participated in all three Subtasks and made the following contribution: 1) we achieved high classification accuracy (81.5%) for Subtask A, which ranked 2ND among 10 teams and was only slightly lower (0.5%) than the top ranking submission; and 2) we proposed an annotation schema to align the subsections in the clinical notes with dialogue exchanges between doctors and patients; 3) we developed two classifiers for dialogue exchanges, and the union of both predictions was used to map the doctor-patient dialogue exchanges to topical note subsections.

Table 1: A dialogue snippet and its summary

Dialogue:
Doctor: When did the nausea and vomiting start?
Patient: About a few hours ago. I can't seem to stomach anything.
Doctor: How many eposodes of vomiting have you had?
Patient: At least four.
Doctor: Any abdominal pain, fever, chill, or other symptoms?
Patient: Just nausea and vomiting. It's been so terrible.
Doctor: I'll order you some Zofran to help bring the nausea to bay. One moment while put the order in.
Subsection: Chief Complaint.
Summary: Intractable nausea and vomiting.

2. Related Work

General domain summarization, such as meeting dialogue summarization [15] and dialogue state tracking [16], is an active area of research. Summarization datasets that are frequently used in general summarization tasks include CNN/DailyMail[17], XSum[18], SamSum[19], DialogSum[20] etc. Unlike the general domain summarization task, where the full dialogues are relatively short and can be modeled directly to generate the full summaries [21, 20, 19, 22], clinical dialogues are frequently lengthy and beyond the input limit of most models.

The summarization of doctor-patient conversations using Digital Scribes is a relatively new area of research. Prior Digital Scribe work has explored various approaches to generate clinical notes from doctor-patient encounter dialogues, which can be relatively long. Most studies apply transformer-based models — Enarvi et al. [23] compared hierarchical recurrent neural network (RNN) encoder with transformer-based model in a sequence-to-sequence approach. They found that transformer-based model was faster to train and outperformed RNN. Zhang et al. [24] leveraged pre-trained a BART model [25] and fine-tuned it with limited clinical dialogue data. Michalopoulos et al. [26] applied a transformer-based sequence-to-sequence architecture and integrated medical domain knowledge. Joshi et al. [27] proposed a variant of a pointer generator network [28], which incorporated local structures in patients' medical history. The model captured important properties of medical conversations, such as medical knowledge coming from standardized medical ontology. Our method aligned closely with Yim and Yetisgen [29], which broke down whole-note summarization into two stages: dialogue-to-note alignment and dialogue snippet summarization. They did not utilize pre-trained transformer models to enhance generalization.

3. Subtasks & Data

In this section, we describe the three MEDIQA-Sum 2023 Subtasks and present an exploratory analysis on the provided datasets. Table 2 summarizes the task type and system inputs and outputs for each Subtask. The input for all Subtasks is transcribed doctor-patient dialogue, where the speaker roles (doctor vs. patient) are indicated at the beginning of each dialogue turn. Table 1 presents an example dialogue snippet with the associated subsection header and clinical note summary.

In this paper, we distinguish between sections and subsections: sections are the four main sections (i.e., "Subjective", "Objective Exam", "Objective Results" and "Assessment and Plan") that are evaluated by the MEDIQA-Sum organizers. Subsections refer to topical subheadings in the clinical notes (e.g. "Exam", "Chief Complaint", "History of Present Illness", etc.), which were identified using pattern matching.

Table 2
Descriptions of the three Subtasks with their corresponding inputs and outputs. The mean and standard deviation of word count in the dialogue snippet or full-encounter dialogue, from the training and validation set, are in the parenthesis.

Subtask	Subtask Type	Input (Mean±Std.)	Output (Mean±Std.)
A	Classification	Dialogue snippet (104.5±116.2)	Subsection header (N.A.)
B	Summarization	Dialogue snippet (104.5±116.2)	Subsection note (40.2±65.8)
C	Summarization	Full-encounter dialogue (1082.9±387.8)	Full note (423.1±136.1)

3.1. Subtask A — Dialogue2Topic Classification

Clinical notes are typically semi-structured and organized by topical subsections. Subtask A is a classification task, where the input is a snippet from a doctor-patient dialogue (dialogue snippet) and the output is a subsection header. A set of 20 subsection headers (e.g. Assessment, Medications, etc.) was predefined by the challenge, as shown in Table 6 in the Appendix.

3.2. Subtask B — Dialogue2Note Summarization

Subtask B is a summarization task, where the input is a dialogue snippet, similar to Subtask A, and the output is a summary of the snippet, corresponding to the prose of a subsection of a clinical note.

3.3. Subtask C — Full-Encounter Dialogue2Note Summarization

Unlike Subtask B which focus on single subsection summarization, Subtask C requires summarizing a full doctor-patient encounter dialogue through complete clinical note with multiple subsections. Table 7 in the Appendix shows the distribution of the subsection headers for Subtask C training and validation sets. There is a mismatch between subsection headers of subtask

A & B and those of subtask C. For example, the subsection header “GYNHX” (which stands for "Gynecological History") is a label for Subtask A, but is not used in SubTask C. Similarly, the subsection header “VITALS” is exclusive to Subtask C. Additionally, there is inconsistency in the phrasing of subsection header names. For instance, Subtask A uses a subsection header “EXAM”, and Subtask C uses “EXAM”, “PHYSICAL EXAM”, and “PHYSICAL EXAMINATION.” To address this problem, we map the original 20 subsection headers in Subtask A and 26 subsection headers in Subtask C to 12 canonical headers, as shown in Table 8 in the Appendix. To create this mapping, we merge the original subsection headers that have similar contents in the clinical notes. For example, in Subtask A and B, subsections “DIAGNOSIS”, “LABS” and “IMAGING” all describe the results of some medical tests or labs; therefore, we merge them into the ‘RESULTS’ subsection. We also used the section tagger code provided by the task organizers, where subsections such as “PLAN”, “ASSESSMENT” and “DISPOSITION” are merged into “ASSESSMENT AND PLAN”. We use the canonical subsection headers in Subtask C.

4. Methods

4.1. Subtask A — Dialogue2Topic Classification

We formulate the topical subsection labeling task as a text generation problem using T5 [30], an encoder-decoder model. In the generative T5 approach, the input consists of a dialogue snippet, a question about the topic, and a list of the subsection headers in the form of “{DIALOGUE SNIPPET} Question: what is the section topic among categories below? topic categories: general_history | medications |…| gynecological_history”. The output is a single predicted subsection header. The model is initialized with T5-large and fine-tuned on Subtask A training data, for 2000 iterations with a batch size of 4.

4.2. Subtask B — Dialogue2Note Summarization

Subtask B involves summarizing dialogue snippets to generate individual clinical note subsection text. We explored Subtask B as a traditional text-in-text-out summarization task using the following generative pre-trained language models (links for models provided in Table 9 of the Appendix):

- Bidirectional and Auto-Regressive Transformers(BART) [25]: a transformer-based model with a bidirectional encoder and an autoregressive decoder; we used a version that was fine-tuned on the Samsum dataset [19].
- Fine-tuned LAnguage Net(Flan)-Text-To-Text Transfer Transformer(T5)-Base [31]: an instruction fine-tuned T5 [30] model; we used a version that was fine-tuned on the Samsum dataset.
- Alpaca-LoRA: a LLaMA [32] model that is instruction fine-tuned using low-rank adaptation (LoRA) [33] on Standford Alpaca dataset [34].
- Generative Pretrained Transformer (GPT) [35]: we applied GPT-3.5, an autoregressive language model with 175 billion parameters. We conducted two-shot experiments with GPT-3.5 through OpenAI’s Chat Completion Application Programming Interface (API). The prompt we used is presented in §A.3 in the Appendix.

The language models above are pre-trained on general domain text or general conversations. To improve the models' understanding of the clinical dialogues in our tasks, we fine-tuned BART-Large, T5-Large[1], and Flan-T5-Large on the provided training data for subtask B. We did not fine-tune GPT-3.5 because it is a proprietary model that cannot be fine-tuned. We were unable to fine-tune Alpaca-LoRA within the time constraints of the shared task.

4.3. Subtask C — Full-Encounter Dialogue2Note Summarization

Subtask C summarizes full encounter dialogues into comprehensive clinical notes with topical subsection headers such as "Chief Complaint" and "Review of Systems". We approach it in three steps. First, we split the full dialogue into dialogue exchanges (§4.3.1). Second, we build two classifiers which label each dialogue exchange with one or more subsection headers (§4.3.2). Third, dialogue exchanges with the same subsection header are aggregated (concatenated) and the concatenated string is sent to Subtask B summarizer models to generate the associated note for that subsection.[2] We repeat this process for every Subtask C canonical subsection header, grouping the generated note text (with their subsection headers) by note section and concatenating all section text to form the full clinical note.

4.3.1. Splitting the Full Dialogue into Dialogue Exchanges

Since the full dialogue transcripts are often too long to be used as a single input sequence with many language models (see mean and standard deviation statistics in Table 2), for Subtask C, we operated on shorter dialogue *exchanges*. The full dialogue was split at the start of each doctor turn using the speaker role tags (e.g., "Doctor: " and "[Doctor]"), such that each *exchange* consists of a single doctor turn and zero or more patient turns. Qualitatively, an exchange frequently constitutes a relatively independent question-answer (QA) pair, where the doctor asks a question and the patient responds.

4.3.2. Classifying Dialogue Exchanges

After the full dialogue is split into dialogue exchanges, each exchange is labeled with one or more Subtask C canonical subsection headers. Like Subtask A, we formulate this subsection labeling task as a text generation problem using T5 [30] and fine-tuned T5-large with training data. Our input template for Subtask C is slightly different from the one for Subtask A, consisting of the exchange to be labeled, two exchanges preceding and two following the current exchange, a question about the topic, and a list of subsection headers (see Appendix A.5). The neighboring exchanges are included in the input because they would provide contextual information to help classify the current exchange. The main challenge for fine-tuning T5 is that no labeled data exist for this classification task. We resolve this issue in two ways, resulting in two classifiers, **Classifier-I** and **Classifier-II**.

Classifier-I is a single-label classifier trained with Subtask A training data. Since the input in Subtask A is a dialogue snippet not a dialogue exchange, we split each snippet into exchanges

[1] We did not find any pre-trained T5-Large for the conversation summarization task.

[2] If no exchanges are labeled with a subsection header, the text for that subsection will be programmatically set to "None".

as described in §4.3.1 and labeled each exchange with the subsection header of the snippet. To make the exchange sequences similar to the Subtask C data, we sample the Subtask A snippets following the common order of the subsections in the Subtask C training data. There are several problems with these synthesized data. First, a snippet belonging to a subsection does not necessarily mean that every exchange in the snippet belongs to the same subsection. Second, the synthesized data use Subtask A subsection headers, and there is a mismatch between those headers and the ones in Subtask C, as discussed in §3.3. Third, a dialogue exchange might align to zero or more subsections, whereas in the synthesized data, each exchange is assigned with exactly one subsection. To address these limitations and better understand the dialogue-note alignment, we manually annotated the alignment for 14 files randomly selected from the Subtask C training set. The detail of the annotation is in Appendix A.6. As shown in Table 12 in the Appendix, only 37.5-41.2% of exchanges align to exactly one subsection, 36.0-37.8% of exchanges do not align to any subsection, and the remaining 22.0-24.7% align to multiple subsections. **Classifier-II** is a multi-label classifier, trained with the 14 manual alignment files. It uses Subtask C canonical subsection headers as labels and can assign multiple labels to an exchange. To fine-tune T5-large, we remove the exchanges that are not aligned to any subsections. During test time, we applied both classifiers to the dialogue exchanges and took the union of the predictions as the classification result.

5. Results

This section presents the results and rankings of the final challenge submissions on the test dataset. For the three tasks, the task organizers use ensemble metrics that correlate well with human judgments. These ensemble metrics combine state-of-the-art evaluation metrics including ROUGE, BERTScore and BLEURT. Since the gold summaries for the Subtask B & C are not released by the task organizers, we provide additional results on the validation dataset to help readers understand our selection of models. For the official evaluation, each team could submit up to three submissions. We present the **team ranking** based on the **best** submission from each team. Our team name is HuskyScribe, which is underlined in all results tables.

Table 3: Subtask A, Team Ranking: Top 3 teams among 10 participants.

Team	Accuracy	Rank
Cadence	0.820	1
HuskyScribe	0.815	2
Tredence	0.800	3

5.1. Subtask A — Dialogue2Topic Classification

Table 3 presents the top 3 teams for Subtask A from the 10 total participants. We rank **2ND** overall, and the gap to the 1st place team is only $-\Delta 0.005$ accuracy.

5.2. Subtask B — Dialogue2Note Summarization

We experimented with several pre-trained and fine-tuned models for Subtask B, to identify the highest performing approach on the validation set. Table 10 in the Appendix presents the validation experimentation, which identified T5-Large as the highest performing system. All our Subtask B test submissions were generated using a fine-tuned T5-Large model. Table 4 presents the Subtask B test results and rankings. In total, the challenge has 7 participants, and we rank at 4TH overall. For the BERT-based metrics (BERT P, BERT R, BERT F1), there is a relatively small gap among the top-ranking teams, including our submission. The major difference comes from the low ROUGE scores.

Table 4
Subtask B, Team Ranking: test results and rankings of final challenge submissions. Our team ranks 4TH overall among 7 participants. Abbreviations: 1) RG = ROUGE. 2) BERT P: Bertscore Precision; 3) BERT R: BERTScore Recall; 4) BERT F1: Bertscore F1.

Team	RG-1	RG-2	RG-L	RG-LSUM	BERT P	BERT R	BERT F1	BLEURT	Agg. Score	Rank
SuryaKiran	0.4398	0.1844	0.3501	0.3501	0.7259	0.7275	0.7231	0.5567	0.5732	1
PULSAR	0.4299	0.2004	0.3569	0.3569	0.7301	0.7211	0.7218	0.5549	0.5689	2
Tredence	0.4244	0.1724	0.3530	0.3530	0.7381	0.7114	0.7207	0.5330	0.5594	3
HuskyScribe	0.3767	0.1504	0.3126	0.3126	0.7361	0.6858	0.7054	0.5037	0.5286	4
...										

5.3. Subtask C — Full-Encounter Dialogue2Note Summarization

Table 5 presents the Subtask C test performance for the full notes. Subtask C has 4 participant teams, and we rank at 3RD overall. A performance breakdown by note section is included in Table 15 of the Appendix. Overall, our ROUGE scores are lower than the 1ST and 2ND ranking teams. If we look closer to the performances at the section level, we observe that our model generates better summaries for "Objective Exam" than team Trendence, and better summaries for "Objective Results" than team uetcorn. Our model does not summarize the "Assessment and Plan" section very well, thus lowering the overall performance.

Table 5
Subtask C, Team Ranking: test performances and ranking on the full note of final challenge submissions. Our team ranks 3RD among 4 participants.

Team	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM	Rank
Tredence	0.4998	0.2035	0.2430	0.4506	1
uetcorn	0.4976	0.2331	0.2467	0.4653	2
HuskyScribe	0.4697	0.1931	0.2228	0.4260	3
PULSAR	0.2941	0.1160	0.1918	0.2608	4

6. Error Analysis

6.1. Subtask A — Dialogue2Topic Classification

We plot the confusion matrix on Subtask A validation set in Figure 1 of the Appendix, where the accuracy among 100 examples is 74%. We observe the model is unable to detect some minority classes in the Subtask A training set, such as "LABS", "EDCOURSE" and "IMAGING", possibly as a result of class imbalance present in the training set. "GENERAL HISTORY" examples can be misclassified as other specific types of history or chief complaints ("CC"), which may be due to the fact that "GENERAL HISTORY" overlaps significantly with "CC" with respect to similarity in contents. Additionally, given the generative setting, our model hallucinates and generates labels that are not in the training set in rare occasion (i.e. "EVALUATION_REPORT").

6.2. Subtask B — Dialogue2Note Summarization

We manually checked some summaries on the validation set and noticed a few issues — 1) In several (3%) cases, our summarizer generates "None," while the gold standards are not. However, the semantics of gold standards are similar to "None". For example, *essentially unchanged from last visit* or *has not had previous history*. Therefore, word-based evaluation metrics such as ROUGE will be low in such cases. 2) We are using the same summarizer for different subsections that have different styles (e.g., "CC" is usually short, while "History of Present Illness" is generally longer). For future work, we plan to incorporate the subsection header to make the model aware of such differences.

6.3. Subtask C — Full-Encounter Dialogue2Note Summarization

For subtask C, we trained two classifiers to predict the subsection for a given exchange. Based on qualitative analysis conducted on the validation set by matching the dialogue to the system output, we observed multiple sources of error: The generated "CC" subsections are usually correlated to the main concern being addressed in the current visit but are not succinct enough compared to the gold standard. Subsections such as "ASSESSMENT AND PLAN" usually summarizes previous subsections and synthesizes information from the dialogue in parts pertinent to the specific instructions given by the doctor, which suggests that using the same summarizer for all subsections may fail to distinguish between subsections of different styles. The "VITALS" and "RESULTS" subsections do not appear in the training data in Subtask A & B. Therefore, our summarizer trained on Subtask A & B data may not be able to assign such subsections. In future work, we could have a separate component that extracts vital sign information from the dialogue.

7. Conclusions

In this work, we presented our approaches for the three Subtasks in 2023 MEDIQA-Sum challenge. For Subtask A, we achieved a high classification accuracy (81.5%) with a fine-tuned T5-large model, which ranked 2ND among 10 participants. For Subtask B, we assessed the capabilities of several state-of-the-art language models in the summarization task. Our final system yielded moderate performances with a fine-tuned T5-Large model. We identified several next steps for improving performance. Incorporating the subsection label could potentially improve the results, given that different subsections have various summary styles. GPT-3.5 has the best validation performance among the all pre-trained language models we evaluated, indicating its great potential in the summarization task. We anticipate that fine-tuning GPT-3.5 could boost the performance, if possible. For Subtask C, we built two dialogue exchange classifiers with 14 manually aligned files and Subtask A training data and use them and the Subtask B summarizers to generate comprehensive full-note summaries. We hypothesize that better subsection classification can lead to better full-note summaries. As a future direction, we believe that annotating more files could further improve performance.

References

[1] C. A. Sinsky, R. Willard-Grace, A. M. Schutzbank, T. A. Sinsky, D. Margolius, T. Bodenheimer, In search of joy in practice: a report of 23 high-functioning primary care practices, Annals of Family Medicine 11 (2013) 272–278. doi:10.1370/afm.1531.

[2] B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, W.-J. Tuan, C. A. Sinsky, V. J. Gilchrist, Tethered to the EHR: Primary care physician workload assessment using EHR event log data and time-motion observations, Annals of Family Medicine 15 (2017) 419–426. doi:10.1370/afm.2121.

[3] M. Tai-Seale, C. W. Olson, J. Li, A. S. Chan, C. Morikawa, M. Durbin, W. Wang, H. S. Luft, Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine, Health Affairs 36 (2017) 655–662. doi:10.1377/hlthaff.2016.0811.

[4] M. W. Friedberg, P. G. Chen, K. R. Van Busum, F. Aunon, C. Pham, J. Caloyeras, S. Mattke, E. Pitchforth, D. D. Quigley, R. H. Brook, et al., Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy, Rand Health Quarterly 3 (2014). PMID: 28083306.

[5] T. D. Shanafelt, L. N. Dyrbye, C. Sinsky, O. Hasan, D. Satele, J. Sloan, C. P. West, Relationship between clerical burden and characteristics of the electronic environment with physician burnout and professional satisfaction, Mayo Clinic Proceedings 91 (2016) 836–848. doi:10.1016/j.mayocp.2016.05.007.

[6] S. Yadav, N. Kazanji, N. KC, S. Paudel, J. Falatko, S. Shoichet, M. Maddens, M. A. Barnes, Comparison of accuracy of physical examination findings in initial progress notes between paper charts and a newly implemented electronic health record, Journal of the American Medical Informatics Association 24 (2017) 140–144. doi:10.1093/jamia/ocw067.

[7] S. K. Bell, T. Delbanco, J. G. Elmore, P. S. Fitzgerald, A. Fossa, K. Harcourt, S. G. Leveille, T. H. Payne, R. A. Stametz, J. Walker, C. M. DesRoches, Frequency and Types of Patient-Reported Errors in Electronic Health Record Ambulatory Care Notes, JAMA Network Open 3 (2020) e205867–e205867. doi:10.1001/jamanetworkopen.2020.5867.

[8] K. Walker, M. Ben-Meir, W. Dunlop, R. Rosler, A. West, G. O'Connor, T. Chan, D. Badcock, M. Putland, K. Hansen, et al., Impact of scribes on emergency medicine doctors' productivity and patient throughput: multicentre randomised trial, British Medical Journal Publishing Group 364 (2019). doi:10.1136/bmj.l121.

[9] 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 27 (2020) 808–817. doi:10.1093/jamia/ocaa020.

[10] G. Finley, E. Edwards, A. Robinson, M. Brenndoerfer, N. Sadoughi, J. Fone, N. Axtmann, M. Miller, D. Suendermann-Oeft, An automated medical scribe for documenting clinical encounters, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 11–15. URL: https://aclanthology.org/N18-5003. doi:10.18653/v1/N18-5003.

[11] 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.

[12] N. H. Crampton, Ambient virtual scribes: Mutuo health's autoscribe as a case study of artificial intelligence-based technology, Healthcare Management Forum 33 (2020) 34–38. doi:10.1177/0840470419872775.

[13] W. Yim, A. Ben Abacha, N. Snider, G. Adams, M. Yetisgen, Overview of the mediqa-sum task at imageclef 2023: Summarization and classification of doctor-patient conversations, in: CLEF 2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023, p. .

[14] B. Ionescu, H. Müller, A.-M. Drăgulinescu, W. wai Yim, A. B. Abacha, N. Snider, G. Adams, M. Yetisgen, J. Rückert, A. G. S. de Herrera, C. M. Friedrich, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, N. Papachrysos, J. Schöler, D. Jha, A.-G. Andrei, A. Radzhabov, I. Coman, V. Kovalev, A. Stan, G. Ioannidis, H. Manguinhas, L.-D. Ştefan, M. G. Constantin, M. Dogariu, J. Deshayes, A. Popescu, Overview of ImageCLEF 2023: Multimedia retrieval in medical, social media, 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.

[15] M. Li, L. Zhang, H. Ji, R. J. Radke, Keep meeting summaries on topic: Abstractive multi-modal meeting summarization, in: , Association for Computational Linguistics, Florence, Italy, 2019, pp. 2190–2196. doi:10.18653/v1/P19-1210.

[16] J. Shin, H. Yu, H. Moon, A. Madotto, J. Park, Dialogue summaries as dialogue states (DS2), template-guided summarization for few-shot dialogue state tracking, in: , Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3824–3846. doi:10.18653/v1/2022.findings-acl.302.

[17] K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, P. Blunsom, Teaching machines to read and comprehend, Advances in neural information processing systems 28 (2015).

[18] 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. doi:10.18653/v1/D18-1206.

[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. doi:10.18653/v1/D19-5409.

[20] Y. Chen, Y. Liu, L. Chen, Y. Zhang, DialogSum: A real-life scenario dialogue summarization dataset, in: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, Association for Computational Linguistics, Online, 2021, pp. 5062–5074. doi:10.18653/

v1/2021.findings-acl.449.

[21] J. Liu, Y. Zou, H. Zhang, H. Chen, Z. Ding, C. Yuan, X. Wang, Topic-aware contrastive learning for abstractive dialogue summarization, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 1229–1243. doi:10.18653/v1/2021.findings-emnlp.106.

[22] Y. Zhang, A. Ni, Z. Mao, C. H. Wu, C. Zhu, B. Deb, A. Awadallah, D. Radev, R. Zhang, Summ^n: A multi-stage summarization framework for long input dialogues and documents, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 1592–1604. URL: https://aclanthology.org/2022.acl-long.112. doi:10.18653/v1/2022.acl-long.112.

[23] S. Enarvi, M. Amoia, M. D.-A. Teba, B. Delaney, F. Diehl, S. Hahn, K. Harris, L. McGrath, Y. Pan, J. Pinto, et al., 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, 2020, pp. 22–30.

[24] L. Zhang, R. Negrinho, A. Ghosh, V. Jagannathan, H. R. Hassanzadeh, T. Schaaf, M. R. Gormley, Leveraging pretrained models for automatic summarization of doctor-patient conversations, in: , Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3693–3712. doi:10.18653/v1/2021.findings-emnlp.313.

[25] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 7871–7880. doi:10.18653/v1/2020.acl-main.703.

[26] G. Michalopoulos, K. Williams, G. Singh, T. Lin, MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations, in: , Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 4741–4749. URL: https://aclanthology.org/2022.findings-emnlp.349.

[27] A. Joshi, N. Katariya, X. Amatriain, A. Kannan, Dr. summarize: Global summarization of medical dialogue by exploiting local structures., in: , Association for Computational Linguistics, Online, 2020, pp. 3755–3763. doi:10.18653/v1/2020.findings-emnlp.335.

[28] A. See, P. J. Liu, C. D. Manning, Get to the point: Summarization with pointer-generator networks, arXiv preprint arXiv:1704.04368 (2017).

[29] W.-w. Yim, M. Yetisgen, Towards automating medical scribing : Clinic visit Dialogue2Note sentence alignment and snippet summarization, in: , Association for Computational Linguistics, Online, 2021, pp. 10–20. doi:10.18653/v1/2021.nlpmc-1.2.

[30] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[31] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al., Scaling instruction-finetuned language models, arXiv preprint arXiv:2210.11416 (2022).

[32] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., Llama: Open and efficient foundation language models, arXiv preprint arXiv:2302.13971 (2023).

[33] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, arXiv preprint arXiv:2106.09685 (2021).

[34] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford alpaca: An instruction-following llama model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[35] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, (2020). arXiv:2005.14165.

A. Appendices

A.1. Statistics of the Datasets for the Three Subtasks

Tables 6 and 7 show the statistics of the training and validation sets for Subtask A&B and C.

Table 6
Statistics of subtask A and subtask B training and validation sets. The subsection headers are sorted in descending order with respect to the counts of the subsections in the training set. The Description column provides a short description of the subsection header. Compression ratio is the ratio of clinical subsection note length over its corresponding dialogue length for the validation set (mean ± std.).

Subsection headers	**Subsection Count**		**Description**	**Compression ratio**
	Training	Validation		
FAM/SOCHX	351	22	Family and social history	0.2764 ± 0.1824
GENHX	282	20	General history	0.4786 ± 0.1833
PASTMEDICALHX	118	4	Past medical history	0.2369 ± 0.1770
CC	77	4	Chief complaint	0.1327 ± 0.1219
PASTSURGICAL	63	8	Past surgical history	0.1713 ± 0.1028
ALLERGY	60	4	Allergy condition	0.2164 ± 0.1692
ROS	60	11	Review of systems	0.3778 ± 0.2031
MEDICATIONS	54	7	Medications currently prescribed	0.1963 ± 0.1465
ASSESSMENT	34	4	Assessment of current condition	0.2430 ± 0.1992
EXAM	23	1	Physical exam	0.2944 ± 0.1810
DIAGNOSIS	19	1	Diagnosis of current condition	0.1714 ± 0.1608
DISPOSITION	15	2	Destination of the patient after hospital discharge	0.2168 ± 0.2180
PLAN	11	3	Treatment plan	0.2732 ± 0.2791
EDCOURSE	8	3	Emergency department course	0.5288 ± 0.3608
IMMUNIZATIONS	8	1	Immunization	0.0859 ± 0.0883
IMAGING	6	1	Results for imaging	0.3078 ± 0.1810
GYNHX	5	1	Gynecological history	0.1914 ± 0.1626
PROCEDURES	3	1	Surgical procedures	0.0579 ± 0.0311
LABS	2	1	Lab results	0.4896 ± 0.1771
OTHER_HISTORY	2	1	Other history	0.1778 ± 0.0444
Total	1201	100		

Table 7
Statistics of Subtask C training and validation sets. The subsection headers are extracted from the free-text full notes using regular expressions and are sorted in descending order with respect to the counts of the subsections in the training set.

Subsection headers	**Subsection Count**	
	Training	Validation
CHIEF COMPLAINT	59	17
RESULTS	52	18
REVIEW OF SYSTEMS	50	15
HISTORY OF PRESENT ILLNESS	45	13
PHYSICAL EXAM	44	14
ASSESSMENT AND PLAN	34	8
INSTRUCTIONS	32	11
PLAN	32	12
ASSESSMENT	29	10
SOCIAL HISTORY	28	10
VITALS	23	9
MEDICATIONS	19	6
MEDICAL HISTORY	18	6
PHYSICAL EXAMINATION	16	3
FAMILY HISTORY	10	5
PAST HISTORY	9	4
CURRENT MEDICATIONS	8	3
ALLERGIES	7	-
SURGICAL HISTORY	7	-
EXAM	4	2
IMPRESSION	4	2
CC	4	2
HPI	4	2
VITALS REVIEWED	3	1
PROCEDURE	1	-
PAST MEDICAL HISTORY	-	2
PAST SURGICAL HISTORY	-	2
BIRTH HISTORY	-	1
Total # of subsections	542	178
Total # of notes	67	20

A.2. Mapping between Subsection and Section Headers

The subsection headers used in Subtask A & B and Subtask C are similar but not identical, as shown in the first two columns in Table 8. The third column shows the canonical subsection headers for Subtask C. The evaluation for Subtask C is done at the section level (see the last column). The classifier for Subtask A and Classifier-I for Subtask C both use the subsection headers in the first column. Classifier-II for Subtask C uses the headers in the third column. The dash lines show the mapping from the first two columns to the third column.

Table 8

Mapping from 20 Subtask A&B and 25 Subtask C subsection headers to 12 Subtask C canonical subsection headers and 4 section headers.

Subtask A&B header	Subtask C header	Subtask C canonical header	Section header
CC	CC CHIEF COMPLAINT	CHIEF COMPLAINT	SUBJECTIVE
FAM/SOCHX	FAMILY HISTORY SOCIAL HISTORY BIRTH HISTORY	FAMILY AND SOCIAL HISTORY	
ROS	REVIEW OF SYSTEMS	REVIEW OF SYSTEMS	
GENHX OTHER_HISTORY GYNHX EDCOURSE	HISTORY OF PRESENT ILLNESS PAST HISTORY	HISTORY OF PRESENT ILLNESS	
PASTMEDICALHX IMMUNIZATIONS	MEDICAL HISTORY	MEDICAL HISTORY	
PASTSURGICAL	SURGICAL HISTORY	SURGICAL HISTORY	
MEDICATIONS	MEDICATIONS CURRENT MEDICATIONS	MEDICATIONS	
ALLERGY	ALLERGIES	ALLGERGIES	
DIAGNOSIS LABS IMAGING	RESULTS	RESULTS	OBJECTIVE RESULTS
	VITALS VITALS REVIEWED	VITALS	
EXAM	PHYSICAL EXAM EXAM PHYSICAL EXAMINATION	PHYSICAL EXAM	OBJECTIVE EXAM
ASSESSMENT PLAN PLAN PROCEDURES DISPOSITION	ASSESSMENT PLAN INSTRUCTIONS ASSESSMENT AND PLAN IMPRESSION PROCEDURE	ASSESSMENT AND PLAN	AP

A.3. GPT-3.5 Prompt for Subtask B

We used OpenAI's Chat Completion API call, which consists of three roles — *System*, *User*, and *Assistant*. *System* message defines the overall task instructions for GPT-3.5. *User* message provides an example dialogue, and *Assistant* message provides the gold summary for the corresponding dialogue. We constructed our prompt as follows:

- *System*: "You are a helpful medical knowledge assistant. Provide comprehensive and accurate summaries of the conversations between doctors and patients."
- *User*: "Summarize the following conversation: [example dialogue one]"
- *Assistant*: [gold standard for dialogue one]
- *User*: "Summarize the following conversation: [example dialogue two]"
- *Assistant*: [gold standard for dialogue two]
- *User*: "Summarize the following conversation: [test dialogue]"

For Subtask B, [example dialogue] and [gold standard for dialogue] are the input and the output of a training instance in the Subtask B training data. [test dialogue] is the input in Subtask B validation or test set.

A.4. The LLMs used in Subtask B

For Subtask B, we experimented with seven large pre-trained LMs. Table 9 shows the urls of the LLMs. Their performance on Subtask B validation set is in Table 10.

Table 9
Links to the pre-trained language models that we used in Subtask A & B.

Model	Weight URL
BART-Large (Samsum)	https://huggingface.co/lidiya/bart-large-xsum-samsum
T5-Base	https://huggingface.co/t5-base
T5-Large	https://huggingface.co/t5-large
Flan-T5-Base (Samsum)	https://huggingface.co/philschmid/flan-t5-base-samsum
Flan-T5-Large	https://huggingface.co/google/flan-t5-large
Alpaca-LoRA	https://huggingface.co/tloen/alpaca-lora-7b
GPT-3.5	https://platform.openai.com/docs/api-reference/chat

Table 10
Performances of pre-trained and fine-tuned language models on the validation set for Subtask B. Fine-tuned T5-Large model has the best performance overall.

Method	ROUGE-1	ROUGE-2	ROUGE-L
Pre-trained			
BART-Large	0.288	0.112	0.231
Flan-T5-Base	0.255	0.091	0.198
Alpaca-LoRA	0.274	0.097	0.204
GPT-3.5	0.315	0.127	0.237
Fine-tuned			
BART-Large	0.394	0.188	0.319
Flan-T5-Large	0.321	0.120	0.279
T5-Large	**0.423**	**0.203**	**0.355**

A.5. Input Template for Subtask C classifiers

We built two classifiers for Subtask C, both using the following template:
"previous section: [LAST TWO EXCHANGES] current section: [CURRENT EXCHANGE TO CLASSIFY] next section: [NEXT TWO EXCHANGES] Question: what is the current section topic? and Is the current section a different topic from the previous section? [ONTOLOGY]".

Classifier I uses the 20 subsection headers for subtask A (see the 1st column in Table 8). If a header is an acronym (e.g., CC), its full name is used as the class label; that is, the [ONTOLOGY] in the template is replaced by "Topic categories: CHIEF COMPLAINTS | FAMILY AND SOCIAL HISTORY | REVIEW OF SYSTEMS | GENERAL HISTORY | OTHER_HISTORY | GYNECOLOGICAL HISTORY | ED COURSE | PAST MEDICAL HISTORY | IMMUNIZATIONS | PAST SURGICAL | MEDICATIONS | ALLERGY | DIAGNOSIS | LABS | IMAGING | EXAM | ASSESSMENT | PLAN | PROCEDURES | DISPOSITION".

Classifier II uses the twelve Subtask C canonical headers (see the 3rd column in Table 8) as class labels; that is, the [ONTOLOGY] in the template is replaced by "Topic categories:

CHIEF COMPLAINT | FAMILY AND SOCIAL HISTORY | REVIEW OF SYSTEMS | HISTORY OF PRESENT ILLNESS | MEDICAL HISTORY | SURGICAL HISTORY | MEDICATIONS | ALLERGIES | RESULTS | VITALS | PHYSICAL EXAM | ASSESSMENT AND PLAN".

A.6. Manual Alignment between Dialogue and Clinical Notes for Subtask C

To better understand the alignment between dialogue exchanges and clinical note subsections and to provide training data for Classifier-II, we manually align fourteen training instances randomly drawn from the training data for Subtask C. Each training instance is a (dialogue, clinical note) pair.

For preprocessing, we split the dialogues into exchanges as explained in §4.3.1, and split the clinical notes into sentences by using the spaCy sentence segmentizer[3]. Then we manually align each sentence in the clinical note with the dialogue exchanges that are deemed associated with the current sentence. It turns out that the mapping between exchanges and sentences is many-to-many. Table 11 presents the inter-annotator agreement, measured with five metrics as defined below.

Table 11
Statistics for inter-annotation agreement for alignment annotation. Exact match is the percentage of sentences which are aligned to the same dialogue exchanges by the two annotators; relaxed match is the percentage of sentences for which the two annotators' alignments overlap.

	Exact match	Relaxed match	Precision	Recall	F-score
Mean	0.7518	0.9461	0.8576	0.8388	0.8400
Std.	0.1616	0.0719	0.0900	0.1393	0.0989

Let $S = \{s_1, \cdots, s_n\}$ be the sentences in a clinical note and $D = \{d_1, \cdots, d_m\}$ be the dialogue exchanges in the corresponding dialogue. Each pair (s_i, d_j) takes either 1 or 0 as its value indicating whether or not the annotator aligns s_i to d_j. The alignment can be represented by an $n \times m$ matrix $X = (s_i, d_j)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, m$. Let X_1 and X_2 be the two matrices of the two annotators. Let $X_{1,i} \in 1 \times m$ and $X_{2,i} \in 1 \times m$ denote the i^{th} row in X_1 and X_2 respectively. The indicator function $\mathbf{1}\{\cdot\}$ returns 1 if the condition is met and 0 otherwise.

Exact Match is the percentage of sentences that are aligned to the same set of dialogue exchanges by the two annotators, and is computed as follows:

$$\mathrm{ExactMatch}(X_1, X_2) = \frac{\sum_{i=1}^{n} \mathbf{1}\{X_{1,i} = X_{2,i}\}}{n} \tag{1}$$

Relaxed Match metric measures the percentage of sentences whose corresponding exchange sets assigned by the two annotators overlap.

$$\mathrm{RelaxedMatch}(X_1, X_2) = \frac{\sum_{i=1}^{n} \mathbf{1}\{X_{1,i} \cdot X_{2,i}^T > 0\}}{n} \tag{2}$$

[3] https://spacy.io/api/sentencizer

Let $C(X) \in \mathbb{Z}^+$ be the function that takes the annotation matrix X as the argument and counts the number of 1's in X. For Precision, Recall and F-score, we treat X_1 as the gold standard and X_2 as the predicted label matrix.

$$\text{Precision}(X_1, X_2) = \frac{C(X_1 \odot X_2)}{C(X_1)} \tag{3}$$

$$\text{Recall}(X_1, X_2) = \frac{C(X_1 \odot X_2)}{C(X_2)} \tag{4}$$

$$\text{Fscore}(X_1, X_2) = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \tag{5}$$

If a dialogue exchange is aligned to one or more sentences, we can label the exchange with the subsection headers of those sentences. Table 12 shows the number of exchanges that have zero or more subsection headers. It can be seen that approximately a third of exchanges are not aligned to any sentences and thus have no subsection labels; these can be things such as greetings which appear in the dialogues but not in the clinical notes. About a quarter of exchanges have more than one subsection label, indicating a single-label classifier such as Classifier I would not perform well for this task.

Table 12
The number (and the percentage) of dialogue exchanges that are associated with zero or more subsection headers for all fourteen annotated examples. The **None** row shows the number of exchanges that are not aligned to any sentence and thus have no subsection labels. The **One** row shows the number of exchanges that have exactly one subsection label. The other two rows are defined similarly.

# of subsection labels	Annotator 1	Annotator 2
None	124 (37.8%)	118 (35.9%)
One	123 (37.5%)	135 (41.2%)
Two	53 (16.2%)	52 (15.9%)
Three or more	28 (8.5%)	23 (7.0%)
Total	328	328

As an exchange can have zero or more subsection headers, we can also look at the distributions of those headers, as shown in Table 13. The "NONE" row shows the number of exchanges that are not aligned to any subsection. The next 11 rows show the number of exchanges aligned to that subsection. Both Table 12 and 13 demonstrate that a good classifier for Subtask C needs to handle cases when an exchange has zero or more than one subsection label.

A.7. Performance on the Subtask C

We also conducted "cheating" experiments for Subtask C using our human annotated data to investigate 1) the performance difference between human annotation and automatic alignment and 2) the degree to which the summarization model can utilize the alignment relationships.

Table 13
The number of occurrences of canonical subsection headers in the human annotation examples. The fourteen files do not include the "SURGICAL HISTORY"subsection. "NONE" means that the dialogue exchange has no corresponding subsection header in the clinical note that is relevant as determined by the annotators.

Subsection header	**Annotator 1**	**Annotator 2**
NONE	124	118
HISTORY OF PRESENT ILLNESS	94	92
ASSESSMENT AND PLAN	90	94
REVIEW OF SYSTEMS	46	40
PHYSICAL EXAM	27	23
CHIEF COMPLAINT	23	20
RESULTS	14	16
FAMILY AND SOCIAL HISTORY	12	17
VITALS	8	8
MEDICATIONS	5	5
MEDICAL HISTORY	5	4
ALLERGIES	3	4
Total	451	441

A.7.1. Human Annotation

The first set of experiments uses the human annotated data as input for the summarizer. We separately feed the annotated labels from annotator 1, annotator 2 and the union of the two annotators as input for the summarizer. The mean evaluation scores for the union of the two annotators are the highest among the three settings, indicating an ensemble of human alignment appears to be more effective.

A.7.2. Automatic Alignment

The second set of experiments employs the two classifiers: Classifier-I, which is trained on Subtask A data, and Classifier-II, which is trained on the human annotation data. We also consider the union of the two classifiers. The evaluation scores show that Classifier-II ties with the union of the two classifiers simply because Classifier-I trained on Subtask A data is a single-label classifier, where the predicted label is highly likely to be included in the set of multi-labels from Classifier-II. Compared with the experiment using human annotations, the automatic alignment approach outperforms the one using human annotation with respect to only ROUGE-2. Note that due to the small sample size ($n = 14$), all the mean differences are not statistically significantly different from 0 (see the coverage of the 95% bootstrap confidence intervals).

A.7.3. Subtask C Test Results — section evaluation

Table 15 presents the test results for Subtask C at four section level — Subjective, Objective Exam, Objective Results, Assessment & Plan.

Table 14

Evaluation scores (mean and 95% bootstrap confidence interval) for the five experiments based on ROUGE-1, -2, -L and -Lsum. Underlined scores are the highest for each column. **Human annotation** experiment uses the human annotated labels as the input to the summarizer. **Automatic alignment** uses the union of system outputs from Classifier-I and Classifier-II on the 14 examples.

	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Annotator 1	0.3096 (0.29, 0.33)	0.0623 (0.06, 0.07)	0.1385 (0.13, 0.15)	0.2619 (0.25, 0.28)
Annotator 2	0.3144 (0.29, 0.34)	0.0686 (0.06, 0.08)	0.1416 (0.13, 0.15)	0.2720 (0.25, 0.30)
Human annotation	0.3318 (0.30, 0.37)	0.0808 (0.06, 0.11)	0.1531 (0.14, 0.17)	0.2807 (0.26, 0.31)
Classifier1	0.3167 (0.29, 0.36)	0.0841 (0.07, 0.10)	0.1495 (0.13, 0.17)	0.2720 (0.25, 0.30)
Classifier2	0.3156 (0.29, 0.35)	0.0846 (0.07, 0.11)	0.1459 (0.13, 0.16)	0.2748 (0.25, 0.30)
Automatic alignment	0.3156 (0.30, 0.36)	0.0846 (0.07, 0.10)	0.1459 (0.14, 0.17)	0.2748 (0.27, 0.31)

Table 15

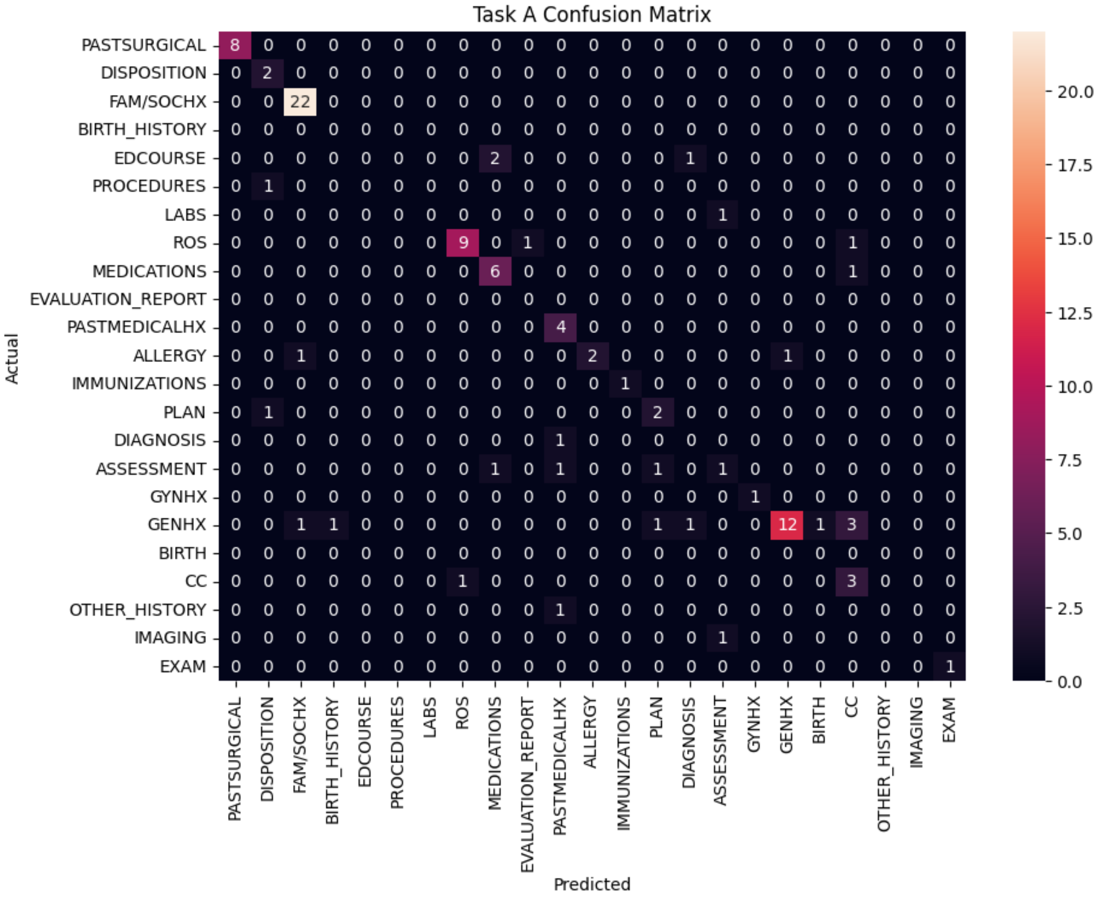
Subtask C, Team Ranking: Performances of final challenge submissions on the test set. The first four columns show the performance on the four sections; the 5th column lists the "Aggregated Score". The rank of the systems are in the last column.

Team	Subjective	Obj. Exam	Obj. Results	Assesment & Plan	Agg. Score	Rank
Tredence	0.5141	0.4045	0.4746	0.4285	0.4554	1
uetcorn	0.4843	0.4384	0.3575	0.4970	0.4443	2
HuskyScribe	0.4758	0.4177	0.3668	0.3932	0.4133	3
PULSAR	0.4125	0.1892	0.4393	0.1807	0.3054	4

A.8. Confusion Matrix for Subtask A

To better understand the errors made by our system for Subtask A, we run the system on the validation set, and create the confusion matrix in Figure 1. The X-axis and Y-axis show the predicted subsection header and gold standard, respectively.

If we map Subtask A subsection headers to Subtask C canonical subsection headers (see Table 8), the new confusion matrix is in Fig 2.



	PASTSURGICAL	DISPOSITION	FAM/SOCHX	BIRTH_HISTORY	EDCOURSE	PROCEDURES	LABS	ROS	MEDICATIONS	EVALUATION_REPORT	PASTMEDICALHX	ALLERGY	IMMUNIZATIONS	PLAN	DIAGNOSIS	ASSESSMENT	GYNHX	GENHX	BIRTH	CC	OTHER_HISTORY	IMAGING	EXAM
PASTSURGICAL	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DISPOSITION	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FAM/SOCHX	0	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
BIRTH_HISTORY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
EDCOURSE	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0
PROCEDURES	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
LABS	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
ROS	0	0	0	0	0	0	0	9	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
MEDICATIONS	0	0	0	0	0	0	0	0	6	0	0	0	0	0	0	0	0	0	0	1	0	0	0
EVALUATION_REPORT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
PASTMEDICALHX	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0
ALLERGY	0	0	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1	0	0	0	0	0
IMMUNIZATIONS	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
PLAN	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0
DIAGNOSIS	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
ASSESSMENT	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	0	0	0	0	0	0
GYNHX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
GENHX	0	0	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	12	1	3	0	0	0
BIRTH	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
CC	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3	0	0	0
OTHER_HISTORY	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
IMAGING	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
EXAM	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Figure 1: Subtask A confusion matrix when running our system on Subtask A validation set. The X-axis and Y-axis are predicted and gold standard labels, respectively. The labels are Subtask A subsection headers.

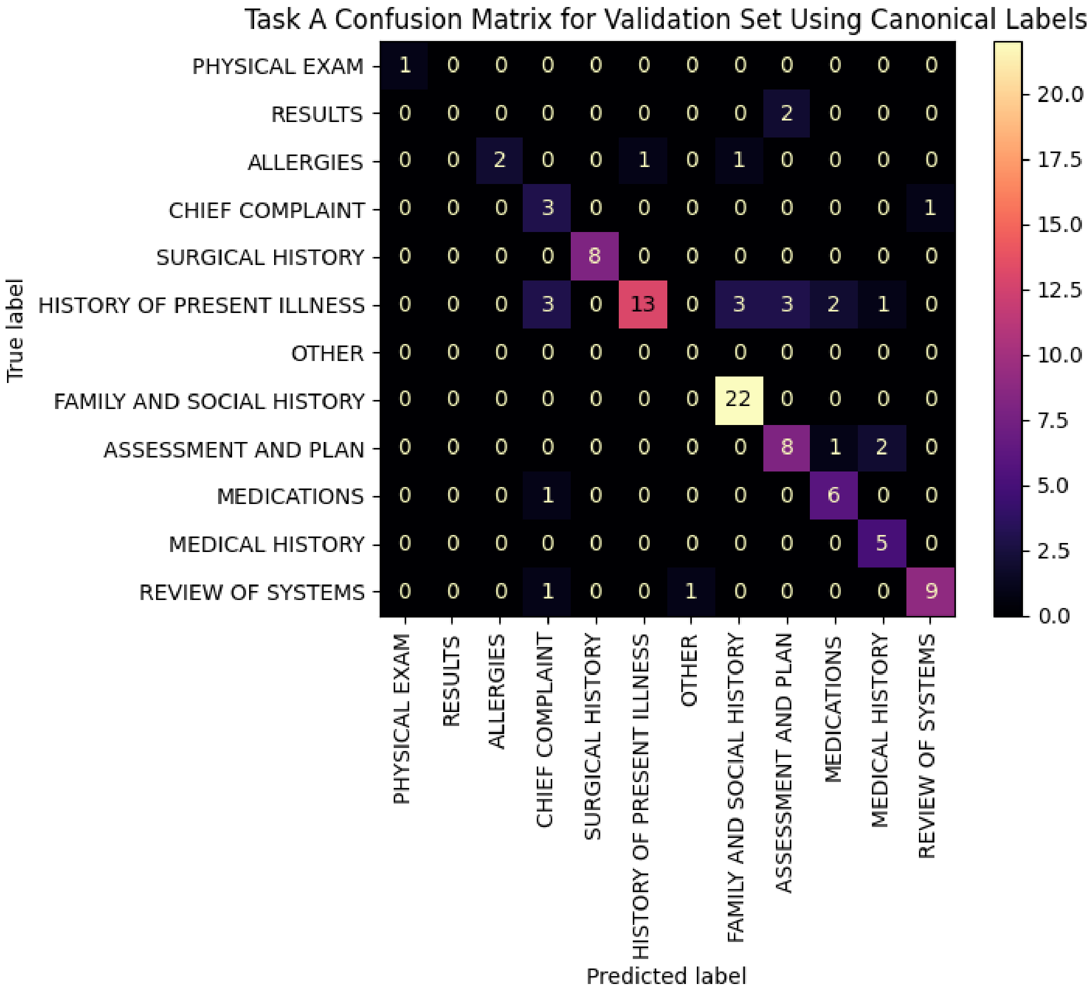


Figure 2: Task A confusion matrix when Subtask A subsection headers are mapped to Subtask C canonical subsection headers.