Towards Accurate and Clinically Meaningful Summarization of Electronic Health Record Notes: A Guided Approach

Abstract— **Clinicians are often under time pressure when they review patients' electronic health records (EHR), therefore, there are great benefits to providing clinicians high quality summarizations of patients' EHR. However, existing summarization algorithms cannot satisfy their needs. In this paper, we present a novel approach to summarize EHR notes using a guided summarization model. Our model integrates a structured template developed with a clinical domain expert, a Named Entity Recognition (NER) model and sentence classification model for guidance extraction, and a fact-checking metric for evaluating the generated summaries. We trained our model on a large de-identified EHR dataset. The results demonstrate that our guidance, which includes Chief Complaint (CC), NER, guidance from the History of Present Illness (HPI) section, and guidance from the Medical Decision Making (MDM) section, can significantly improve the performance of the models in generating accurate and clinically meaningful summaries. The Gsum (CNN) model with all the guidance aforementioned achieved the highest F1 score of 46.4, demonstrating the effectiveness of introducing precise and informative guidance to models from the general domain when the training data on the clinical domain is prohibitively sensitive and expensive. This work contributes to the ongoing efforts to automate the summarization of EHR notes, with the ultimate goal of improving healthcare delivery and patient outcomes.**

Keywords— **Electronic health records, named entity recognition, abstractive summarization, natural language processing, deep learning**

I. Introduction

The rapid growth of medical data, particularly Electronic Health Record (EHR) notes, presents both opportunities and challenges for healthcare delivery. While this wealth of data can provide valuable insights into patient health, the sheer volume and complexity of EHR notes can make it difficult for healthcare providers to extract the most relevant details. This is further complicated by the use of specific medical terms and pharmaceutical information, as well as the potential for interruptions, repetitions, and abrupt shifts in the topic in clinician-patient conversations.

Existing solutions for summarizing EHR notes, such as DecSum [1], have made significant strides in this area. DecSum, for instance, uses the decision of EHR notes to generate formal medical decision-making summaries. However, these models often struggle to retain key information and avoid repetition of irrelevant details [2]. For example, Moreover, the lack of a standardized structure in casual conversations further complicates the task of dialogue summarization [3], [4].

In this paper, we propose a novel approach to summarize EHR notes that addresses these limitations. Our approach leverages guidance extraction and a structured template to generate concise, clinically meaningful summaries. We also introduce a fact-checking metric to evaluate the accuracy and completeness of the generated summaries from a clinical perspective. This model aims to allow clinicians to rapidly identify key factors from previously documented clinical visits.

The contributions of our study are threefold. First, we collaborated with a clinical domain expert to develop a detailed template for the summary. This template, which includes key aspects such as patient demographics, chief complaint (CC), OPQRST (a common mnemonic for patient assessment), diagnostics, treatments, and disposition, serves as a guideline for generating summaries. Second, we utilized a Named Entity Recognition (NER) model to extract medical entities related to the aspects outlined in our proposed template. The extracted entities, serve as a "guidance signal" for the summarization model. This approach ensures that the generated summaries accurately reflect the most important aspects of the patient's medical condition and treatment. Third, we designed a fact-checking metric based on the aspects outlined in our proposed template. This metric evaluates the accuracy and completeness of the generated summaries by checking for the presence and accuracy of clinically meaningful phrases. It provides a quantitative measure of the quality of the summaries, ensuring that they accurately reflect the key clinical information found in the original EHR notes. Collectively, these contributions aim to improve the quality of medical decision-making and enhance patient care in healthcare settings.

II. Related Work

One of the most effective abstractive summarization methods is the Sequence to Sequence (Seq2seq) [5] approach, which has achieved state-of-the-art results. The BART model [6] and its follow-up methods have been successful in shortening long text to a concise summary. Several studies have focused on including guidance signals in the standard seq-to-seq architecture. Zhu et al. [7] introduced relational triples (subject, relation, object). Narayan et al. [8] and He et al. [9] incorporated keywords into the generation process. Then, Gsum [10] is introduced as a guided summarization framework that can support different external guidance signals.

A. Summarization on Medical Domain

Seq2seq models have been used in biomedical summarization tasks, including summarization of doctor-patient dialogues [11]. Biomedical language models such as BioBART

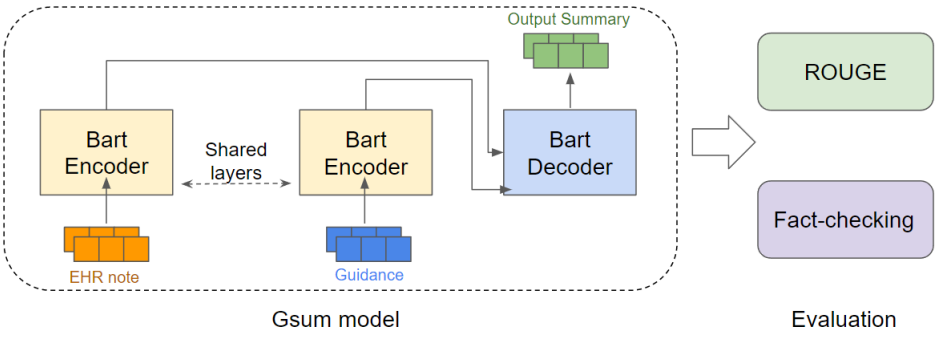

Fig. 1. Overview of our proposed summarization pipeline and evaluation metrics.

[12] have demonstrated enhanced performance compared to BART and set strong baselines on several tasks. Other tasks leverage biomedical information such as UMLS [13] to guide the generation process [4]. Medical dialogue summarization is a common task in the biomedical summarization domain, with significant research in recent years focused on developing methods to enhance communication between healthcare professionals and patients. Medical dialogues typically involve interactions between patients and doctors, and the field of medical dialogue summarization is rapidly evolving [14], [15]. In addition, clinical data summarization includes not only dialogue summarization but also summarization of Electronic Health Record (EHR) notes [4], [16].

III. Methodology

A. Summary Template

To enhance the quality and consistency of the generated summaries, we collaborated with a clinical domain expert to develop a structured template for the output target. This template serves as a guideline for the summarization process, ensuring that all critical aspects of a patient's condition are captured and presented in a standardized format.

Our proposed template is as follows: "Patient is a [age] year old [male/female] who presents with [chief complaints]. [Insert important OPQRST information]. [Insert important diagnostics]. [Insert treatments (medications, procedures)]. The patient was [Disposition (admitted to hospital/discharged to home)]."

Each bracketed field in the template corresponds to a key aspect of the patient's medical condition: 1) Age and gender provide basic demographic information; 2) The chief complaints are the reason for the patient's visit; 3) "OPQRST" [17] is a widely used method for understanding a patient's symptoms in a detailed and systematic way. OPQRST information refers to the *onset, provocation, quality, region/radiation, severity,* and *timing* of the patient's symptoms; 4) Diagnostics include diagnostic studies that were performed; 5) Treatments encompass any medications prescribed or procedures performed; 6) Disposition indicates the patient's status at the end of the visit. By using this template, we aim to generate summaries that are not only accurate and concise but also clinically meaningful and easy to interpret for healthcare providers.

TABLE I
Dataset Statistics compare with the original corpus of Gsum

Dataset	Train	Valid	Test
CNN/DM	287,226	13,368	11,490
EHR-OURS	55,954	18,652	18,652

B. Guidance

Our approach to guidance extraction consists of three main steps, each designed to capture different aspects of the clinical information contained in Electronic Health Record (EHR) notes:

1) Medical Entity Extraction: We utilized the Bio-Epidemiology-NER [18] Python library, built on top of the biomedical-ner-all model, to extract medical entities related to the aspects outlined in our proposed template. The extracted entities, which include the demographic, CC, treatment, diagnosis, and disposition, serve as a "NER guidance signal" for the summarization model.

2) OPQRST Sentence Classification: Bio_ClinicalBERT [19] sequence classification model was fine-tuned on a set of 589 sentences (from 86 notes) for the HPI section of EHR notes, annotated by a domain expert to classify whether a sentence mentions OPQRST information or not. The Bio_ClinicalBERT model, pre-trained on a large corpus of medical literature including PubMed [20] and MIMIC-III [21] notes, effectively captures the medical context and terminology present in the HPI notes. The test accuracy of the OPQRST sentence classification model was 92.05%. Sentences predicted as positive by this model are used as HPI guidance.

3) Treatment Sentence Classification: We fine-tuned another Bio_ClinicalBERT model on a set of 741 sentences (from 86 notes) in MDM section of EHR notes, annotated to classify whether a sentence belongs to treatment, diagnostic results, summarization, or is unrelated. The test accuracy of this model was 85.23%. Sentences predicted by this model to belong to treatment are used as MDM guidance.

By integrating these guidance extraction processes into our summarization model, we aim to generate summaries that accurately reflect the most important aspects of a patient's medical condition and treatment, as outlined in our proposed

template.

C. Model Structure

We adopt the Gsum model as the foundational architecture for our method, with BART serving as the backbone. This model is partitioned into the encoder and decoder components. The encoder architecture consists of two BART encoders: one for encoding the input source document and the other for encoding the guidance signals. Each BART encoder is composed of N encoding layers, each layer containing a self-attention block and a feed-forward block. The BART decoder attends to the guidance signals first, generating corresponding representations, and then attends to the entire source document based on these guidance-aware representations. The output representation is then processed further in a feed-forward block. This structure allows for the effective integration of guidance signals into the summarization process, enhancing the quality and relevance of the generated summaries.

D. Fact-checking metric

To evaluate the accuracy and completeness of the generated summaries, we designed a fact-checking metric based on the aspects outlined in our proposed template. This metric assesses whether the generated summary accurately reflects the key clinical information found in the original Electronic Health Record (EHR) notes.

The fact-checking metric operates by comparing the generated summary with the ground truth summary at the entity level, checking for the presence and accuracy of the clinically meaningful entities from six key aspects in the template. For each aspect, the metric checks whether the corresponding information in the generated summary matches the information in the original EHR notes. We calculate the precision, recall, and F1 score for each clinical entity as follows:

- True Positive: entities mentioned in both the ground truth and predicted summary
- False Positive: entities mentioned only in the predicted summary
- False Negative: entities mentioned only in the ground truth summary

The scores are then calculated as:

$$Precision = \frac{\#TruePositive}{\#TruePositive + \#FalsePositive} \quad (1)$$

$$Recall = \frac{\#TruePositive}{\#TruePositive + \#FalseNegatives} \quad (2)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

The proposed metric can be automatically performed by NER model, or manually checked by domain experts at the entity level. In this study, the fact-checking metric was evaluated by the biomedical-ner-all model. By using this metric, we can quantitatively assess the quality of the generated summaries, ensuring that they accurately reflect the key clinical information found in the original EHR notes.

IV. Experiment Setup

A. Dataset

We utilized a large de-identified EHR dataset from local healthcare institutes for model training and testing. This dataset includes real de-identified medical text data from emergency department patients, such as chief complaints, histories of illness, medical decision-making, and diagnosis. We excluded the review of systems and lab parts from our dataset to focus on the summarization task. The dataset contains a total of 93,258 EHR samples, with 55,954 for training, 18,652 for validation, and 18,652 for testing, as shown in Table I. Due to patient privacy concerns, the dataset is not publicly available.

B. Experiment settings

We trained Gsum model using Label smoothed cross entropy [22] as the loss function, with dropout and attention dropout set to 0.1 to combat overfitting. The initial learning rate was set to 3e-05, and we used a polynomial decay-based learning rate scheduler and the Adam optimizer. The model was trained on an NVIDIA GeForce RTX 2080 Ti GPU for 12 epochs with a batch size of 1024. GSum (CNN) refers to the original Gsum model trained using the CNN/DM corpus in [10]. GSum (HPI) refers to the Gsum model trained on our EHR notes corpus with the input of HPI section, and uses the Chief complaint (CC) as the guidance. Gsum (HPI + MDM) refers to the Gsum model trained using the HPI and MDM section of the original note.

For testing, we set the output length to 140 words and the minimum word length to 55 words to ensure complete sentence generation. The beam size for text generation was set to 4. The input for all models is the HPI and MDM, with the MDM serving as the ground truth. The testing settings were kept the same across all the experiments including the baseline models. The BART and BioBART models do not support guidance. We experiment with the Gsum-based models under different guidance settings. CC is when the chief complaint is added to the guidance. NER refers to the NER tags that are extracted by the biomedical-ner-all model from the clinical note. NER subset is the subset of NER tags that are marked as clinically important by domain expert from the whole set of extracted NER tags. The process of obtaining the HPI and MDM guidance is described in section III-B. Different combinations of the guidance signal were experimented and the top 3 best-performing combinations for each type of model were reported in Table II.

V. Results and Discussion

Table II provides a comprehensive overview of the performance of various models under different guidance settings. Several key observations can be made from these results. Firstly, the addition of guidance to the clinical note for summarization significantly enhances the performance. Both the Gsum (CNN) model and the Gsum (HPI + MDM) model outperform the BioBART and BART models, with an F1 score improvement of 19.1%. Secondly, the type and quantity of guidance also influence the model's performance. For

TABLE II

SUMMARIZATION PERFORMANCE OF MODELS UNDER DIFFERENT EXPERIMENT SETTINGS. THE **BEST**/2ND-BEST SCORES IN EACH COLUMN ARE IN BOLD/UNDERLINED.

Models	Guidance Settings	ROUGE	Fact-checking metrics		
			Precision	Recall	F1
BART	-	63.8	51.7	27.2	33.5
BioBART	-	42.6	38.3	24.0	27.3
Gsum (CNN)	CC	59.7	51.7	23.4	30.0
Gsum (CNN)	CC + NER + HPI guide + MDM guide	**75.7**	**58.6**	**42.1**	**46.4**
Gsum (CNN)	CC + NER subset + HPI guide + MDM guide	73.8	56.9	40.3	44.8
GSum (HPI)	CC	42.1	31.3	26.1	26.3
GSum (HPI)	CC + NER + HPI guide + MDM guide	43.6	32.5	29.2	28.8
GSum (HPI)	CC + NER subset + HPI guide + MDM guide	43.7	32.5	29.3	28.8
GSum (HPI + MDM)	CC	49.9	44.3	27.7	31.8
GSum (HPI + MDM)	CC + NER + HPI guide + MDM guide	53.1	46.1	31.8	35.3
GSum (HPI + MDM)	CC + NER subset + HPI guide + MDM guide	53.2	46.1	30.8	34.6

TABLE III

ASPECT LEVEL PERFORMANCE OF BART

Aspects	Precision	Recall	F1
Total	51.7	27.2	33.5
Disposition	12.4	12.3	12.3
Treatment	15.2	11.2	12.2
Diagnostics	28.7	11.9	15.1
Demographics	44.2	44.5	44.2
OPQRST	54.1	50.4	47.1
CC	23.6	21.3	21.6

instance, incorporating NER tags, HPI, and MDM guidance improves the F1 scores by 14% for the Gsum (CNN) model and by 3.5% for the Gsum (HPI + MDM) model. Furthermore, the scores increase when all NER tags are included as guidance compared to when only a subset of NER tags is used. The Gsum (CNN) model, along with the CC + NER + HPI and MDM guidance, performs the best with an F1 score of 46.4. This indicates that even though the Gsum (CNN) model was not trained on Clinical corpus, it can still achieve good performance by introducing precise and informative guidance.

However, it is important to note that the ROUGE score [23] may not be the only appropriate metric for evaluating the efficacy of our model. Our primary objective is not merely to summarize the clinical note but also to populate a summary paragraph following the template. According to clinical experts, a high-quality clinical summary should cover all these aspects. The ROUGE score, which measures sentence similarity, does not provide information about the 'informative' nature of the summary. Therefore, a fact-checking metric was proposed to perform an entity-level comparison to determine whether the entities of interest in the ground truth also appear in the predictions from aspects of the template. The fact-checking metric, unlike the ROUGE score, can also perform a fine-grained analysis at the aspect level, as shown in Table III.

By examining the aspect-level performance, we can identify the entities for which the model does not perform well. Even though BART and BioBART outperform the GSum (HPI) and Gsum (HPI + MDM) models in terms of ROUGE scores, they fail to extract all the entities of interest, as indicated by their low recall scores, especially for the aspects of disposition, treatment, and diagnostics. By looking at the output generated by BART and BioBART, most output sentences are directly copied from the original text, whereas Gsum (HPI) and Gsum (HPI + MDM) tend to paraphrase the original text into more concise and coherent sentences. The challenge, therefore, is to paraphrase the information in a way that generates a summary containing 'all' the important information.

VI. CONCLUSION

In this paper, we presented a novel approach to summarizing EHRs that leverages a structured template, guidance extraction, and a fact-checking metric. The addition of informative guidance has been shown to significantly enhance the performance of the summarization models. The Gsum (CNN) model, despite not being trained on a clinical corpus, demonstrated superior performance. This underscores the effectiveness of introducing precise and informative guidance to models from the general domain when the training data on the clinical domain is prohibitively sensitive and expensive. The fact-checking metric provides a quantitative measure of the quality of the summaries, ensuring their accuracy and completeness. Our approach addresses the challenges of existing summarization methods, providing a more accurate and clinically meaningful summary of EHRs. However, there is always room for further improvement and exploration. Future work could focus on refining the guidance extraction process and improving the fact-checking metric with expert interactions.

REFERENCES

[1] C.-C. Hsu and C. Tan, "Decision-focused summarization," 2021.

[2] C. Li, W. Xu, S. Li, and S. Gao, "Guiding generation for abstractive text summarization based on key information guide network," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 55–60.

[3] M. Khalifa, M. Ballesteros, and K. McKeown, "A bag of tricks for dialogue summarization," *arXiv preprint arXiv:2109.08232*, 2021.

[4] G. Michalopoulos, K. Williams, G. Singh, and T. Lin, "Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations," in *Findings of the Association for Computational Linguistics: EMNLP 2022*, 2022, pp. 4741–4749.

[5] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," 2014.

[6] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[7] C. Zhu, W. Hinthorn, R. Xu, Q. Zeng, M. Zeng, X. Huang, and M. Jiang, "Boosting factual correctness of abstractive summarization with knowledge graph," *arXiv preprint arXiv:2003.08612*, 2020.

[8] S. Narayan, Y. Zhao, J. Maynez, G. Simões, V. Nikolaev, and R. McDonald, "Planning with learned entity prompts for abstractive summarization," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1475–1492, 2021.

[9] J. He, W. Kryściński, B. McCann, N. Rajani, and C. Xiong, "{CTRL}sum: Towards generic controllable text summarization," *arXiv*, 2020.

[10] Z.-Y. Dou, P. Liu, H. Hayashi, Z. Jiang, and G. Neubig, "GSum: A general framework for guided neural abstractive summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, Jun. 2021, pp. 4830–4842. [Online]. Available: https://aclanthology.org/2021.naacl-main.384

[11] S. Jeblee, F. K. Khattak, N. Crampton, M. Mamdani, and F. Rudzicz, "Extracting relevant information from physician-patient dialogues for automated clinical note taking," in *Proceedings of the tenth international workshop on health text mining and information analysis (LOUHI 2019)*, 2019, pp. 65–74.

[12] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and evaluation of a biomedical generative language model," in *Proceedings of the 21st Workshop on Biomedical Language Processing*. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 97–109. [Online]. Available: https://aclanthology.org/2022.bionlp-1.9

[13] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32 Database issue, pp. D267–70, 2004.

[14] M. Khalifa, M. Ballesteros, and K. R. McKeown, "A bag of tricks for dialogue summarization," *CoRR*, vol. abs/2109.08232, 2021. [Online]. Available: https://arxiv.org/abs/2109.08232

[15] K. Krishna, S. Khosla, J. P. Bigham, and Z. C. Lipton, "Generating SOAP notes from doctor-patient conversations," *CoRR*, vol. abs/2005.01795, 2020. [Online]. Available: https://arxiv.org/abs/2005.01795

[16] N. Kanwal and G. Rizzo, "Attention-based clinical note summarization," in *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 2022, pp. 813–820.

[17] J. Eloge, T. C. Napier, and B. Dantz, "Opqrst (u): Integrating substance use disorders or "use" into the medical history," *Substance Abuse*, vol. 39, no. 4, pp. 505–508, 2018.

[18] S. Raza, D. J. Reji, F. Shajan, and S. R. Bashir, "Large-scale application of named entity recognition to biomedicine and epidemiology," *PLOS Digital Health*, vol. 1, no. 12, p. e0000152, 2022.

[19] E. Alsentzer, J. R. Murphy, W. Boag, W.-H. Weng, D. Jin, T. Naumann, and M. B. A. McDermott, "Publicly available clinical bert embeddings," 2019.

[20] K. Canese and S. Weis, "Pubmed: the bibliographic database," *The NCBI handbook*, vol. 2, no. 1, 2013.

[21] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, no. 1, pp. 1–9, 2016.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[23] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, Jul. 2004, pp. 74–81. [Online]. Available: https://aclanthology.org/W04-1013