# MedicalSum: A Guided Clinical Abstractive Summarization Model for Generating Medical Reports from Patient-Doctor Conversations

**George Michalopoulos**[1], **Kyle Williams***, **Gagandeep Singh**[2]
**Thomas Lin**[1]
Microsoft Corporation, One Microsoft Way, Redmond, WA 98052[1],
Nuance Communications, 1 Wayside Road, Burlington, MA 01803[2]
{georgemi,tlin}@microsoft.com, kyle@kmw.ai, gagandeep.singh1@nuance.com

## Abstract

We introduce *MedicalSum*, a transformer-based sequence-to-sequence architecture for summarizing medical conversations by integrating medical domain knowledge from the Unified Medical Language System (UMLS). The novel knowledge augmentation is performed in three ways: (i) introducing a guidance signal that consists of the medical words in the input sequence, (ii) leveraging semantic type knowledge in UMLS to create clinically meaningful input embeddings, and (iii) making use of a novel weighted loss function that provides a stronger incentive for the model to correctly predict words with a medical meaning. By applying these three strategies, *MedicalSum* takes clinical knowledge into consideration during the summarization process and achieves state-of-the-art ROUGE score improvements of 0.8-2.1 points (including 6.2% ROUGE-1 error reduction in the PE section) when producing medical summaries of patient-doctor conversations.

## 1 Introduction

The volume of data created in healthcare has grown considerably as a result of record keeping and regulatory requirements policies (Kudyba, 2010). The documentation requirements for electronic health records (EHR) have been shown to be a significant factor contributing to physician burnout (van Buchem et al., 2021; Tran et al., 2020). As a result, the automatic creation of medical documentation has been proposed as one way to address this issue.

To date, there have been several attempts at automatically generating summaries of clinical encounters. Enarvi et al. (2020) employed a transformer model for summarizing doctor-patient conversations. Joshi et al. (2020) developed models to summarize dialogue snippets between two to ten physician-patient turns long. Finally, Jeblee et al. (2019) and Lacson et al. (2006) utilized extractive methods to identify the most important utterances which are combined to form the final summary.

The summaries generated by current summarization models are not straightforwardly controllable (Li et al., 2018). Dialogue summarization is also challenging because casual conversation can include interruptions, repetitions, and sudden topic transitions (Khalifa et al., 2021), and generally does not follow the structure of a written document (Zhu and Penn, 2006). These challenges can lead to problems, such as the omission of key information, or the hallucination of unsupported information. Summaries for medical documentation must also use the correct medical terminology expected by physicians (Knoll et al., 2022).

**Target Summary**
… I will prescribe olopatadine 0.2 percent **ophthalmic drops**. If the symptoms do not improve, the patient will return and consider bacterial cause.---**Dyslipidemia**.

**Baseline Model Output**
… Patient Education and Counseling : The patient was advised to keep wiping green and thick mucus from the eye.

**MedicalSum Model Output**
…Medical Treatment : The patient will continue daily loratadine and **eye drops**.---**Dyslipidemia**.

Figure 1: Distinct output from the baseline model and the MedicalSum model, with formatting tokens removed. MedicalSum generates a clinical summary that contains relevant medical facts.

To help address this problem, we propose a novel knowledge-augmented transformer model that uses medical knowledge to guide the summarization process in various ways to increase the likelihood of relevant medical facts being included in the summarized output (An example of such output is in Figure 1). Key paper contributions include: (i) We are the first, to the best of our knowledge, to pro-

*This work was conducted at Microsoft.



*Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749
December 7-11, 2022 

pose the usage of medical knowledge from a clinical Metathesaurus (UMLS (Bodenreider, 2004)) in the summarization process of a transformer-based model in order to generate 'medically focused' clinical note summaries. (ii) We answer the question of how to incorporate structured medical knowledge in medical documentation generation by designing 3 specific signals over medical entities. (iii) By leveraging these methods the MedicalSum model achieves ROUGE-1 and ROUGE-L improvement between 0.8% and 2.1% in all experiments on medical note summarization.

## 2 Related Work

There are two main approaches for summarization. *Extractive* methods (Kupiec et al., 1995) where the summary is created from passages that are copied from the source text and *abstractive* (Chopra et al., 2016) methods where phrases and words not in the source text can be used to create the summary.

**Neural Abstractive Summarization:** For the task of abstractive summarization, sequence-to-sequence (seq-to-seq) summarization models have achieved state-of-the-art results (Sutskever et al., 2014). Furthermore, different architectures have been proposed to improve the performance of a seq-to-seq model. In Enarvi et al. (2020), the authors incorporated a transformer-based (Vaswani et al., 2017) encoder-decoder architecture in order to produce highly-accurate summaries. In addition, in See et al. (2017), a pointing mechanism was used for copying words from the source document.

**Guided Summarization:** Several studies have focused on including guidance signals in the standard seq-to-seq architecture. Zhu et al. (2020) proposed the usage of relational triples (subject, relation, object). Narayan et al. (2021) and He et al. (2020) included a set of keywords that are incorporated into the generation process. Finally, Dou et al. (2021) created a guided summarization framework that can support different external guidance signals.

**Medical Summarization:** Pivovarov and Elhadad (2015) introduced a summarization model which was focused on creating accurate summaries for clinical data. Enarvi et al. (2020) used a pointer-generator transformer model to accurately generate notes from doctor-patient conversations. Finally, Joshi et al. (2020) used a variation of the pointer-generator model that leveraged shared medical terminology between source and target to distinguish important words from unimportant words.

## 3 Dataset

For the training of the *MedicalSum* model, we have to select a large enough dataset that would provide the necessary data for the medical signals to meaningfully affect the performance of the model. However, there are no publicly available large-scale datasets for medical summarization and thus we have to use a proprietary one. We use English data consisting of recently recorded Family Medicine patient-doctor visits. The speaker-diarized conversation transcripts corresponding to the audio files were obtained using an automatic speech recognizer system and medical professionals created the associated clinical notes.

The reports for family medicine are organized under three sections that correspond to three broad areas of a medical note: (i) History of Present Illness (HPI) which captures the reason for the visit. (ii) Physical Examination (PE) which captures findings from a physical examination. (iii) Assessment and Plan (AP) which captures the assessment by the doctor and the treatment plan. Table 1 shows detailed statistics of our dataset.

| | Train | Valid | Test | A.W |
|---|---|---|---|---|
| AP | 42106 | 648 | 2525 | 2586 |
| HPI | 43092 | 657 | 2551 | 2584 |
| PE | 39815 | 635 | 2442 | 2633 |
| RAD | 91544 | 2000 | 600 | 49 |

Table 1: Number of reports/encounters for the train/validation/test set of each section of the family medicine reports and the MEDIQA third task; A.W is average word count in those encounters.

As the above-mentioned dataset contains patients' private medical information it cannot be made publicly available, and that is the reason that we decided to experiment with a public dataset as well to allow for a more open comparison. We tackle the third task of the MEDIQA 2021 challenge (Ben Abacha et al., 2019) on summarization of radiology reports (RAD) (Johnson et al., 2019). From Table 1, it can be observed that the input documents in the MEDIQA dataset are much smaller than the documents of the family medicine dataset. However, we include it in order to have an evaluation of the models and the baseline on an external dataset. Our experiments are consistent with the datasets' intended use, as they were created for research purposes and we did not notice any indication of offensive content in the datasets.

## 4 Method

### 4.1 MedicalSum: Medical Guided Transformer Pointer Generator Model

We adopt the transformer self-attention model from (Vaswani et al., 2017) in the encoder and in the decoder to create context-dependent representations of the inputs. Both encoder and decoder consist of six layers of self-attention with 8 attention heads and each decoder layer attends to the top of the encoder stack after the self-attention. We use the base model size of 8 attention heads with a total of 512 token outputs and a 2048-dimensional feed-forward network. Furthermore, each encoder and decoder layer contains a position-wise feed-forward network that consists of two transformations and a ReLU activation in between. A simplified image of the MedicalSum model can be found in Figure 2. The details of each added component are discussed in the following sections.

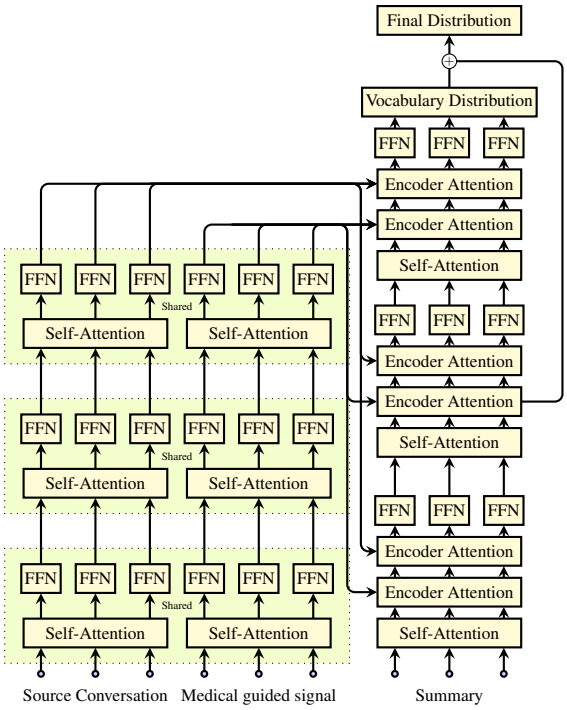


Figure 2: Illustration of **MedicalSum** a transformer sequence-to-sequence model with a pointer-generator and guidance mechanism.

### 4.2 Pointer-Generator

We implement the pointer generator mechanism as described in (Enarvi et al., 2020; See et al., 2017). We choose to use a single attention head to attend to the tokens that are good candidates for copying. In (Garg et al., 2019) it was stated that the penultimate layer seems to naturally learn alignments, so we use its first attention head for pointing.

### 4.3 Medical Guidance Signal

We include a medical guidance signal in the summarization process, that consists of all the medical terms in the input sequence that could be identified in UMLS using the MedCAT toolkit (Kraljevic et al., 2021), by introducing two encoders (that share weights) that encode the input text and the guidance signal respectively (Dou et al., 2021).

Each encoder layer for the input and the guidance signal consists of a self-attention block and a feed-forward block. Each decoder layer consists of a self-attention block, a cross-attention block with the medical guidance signal, in order to inform the decoder which sections of the source document are important, a cross-attention block with the encoded input where the decoder attends to the whole source document based on the guidance-aware representations and a feed-forward block.

As *MedicalSum* focuses on the creation of summaries on medical data, we create a medical guidance signal with all the words with a medical meaning (as they are written in the input text). We believe that this signal will be beneficial to the performance of the model as a guidance signal which is created as a set of individual keywords $\{w_1, ..., w_n\}$, can help the model to focus on specific desired aspects of the input (Dou et al., 2021). We chose to identify medical entities with UMLS as it is a compendium of many biomedical vocabularies (e.g. MeSH (Dhammi and Kumar, 2014), ICD-10 (WHO, 2004)) and thus it contains all the major standardized clinical terminologies.

### 4.4 Semantic Type Embeddings

We introduce a new embedding matrix called $S \in \mathbb{R}^{D_s \times d}$ into the input layer where $d$ is the transformer hidden dimension and $D_s = 50$ is the number of UMLS semantic types used by our model. It should be noted that in the $S$ matrix, each row represents the unique semantic type in UMLS that a word can be identified with.

To incorporate the $S$ embedding matrix into the input embedding layer, all the words with a clinical meaning defined in UMLS are identified and their corresponding semantic type is extracted. By introducing the semantic type embedding, the input vector for each word $w_j$ is updated to:

$$u_{input}^{(j)\prime} = p^{(j)} + Ew_j + S^\top s_{wj} \quad (1)$$

where $s_{wj} \in \mathbb{R}^{D_s}$ is a 1-hot vector corresponding to the semantic type of the medical word $w_j$ and $p^{(j)} \in \mathbb{R}^d$ is the position embedding of the $j^{th}$ token in the sentence. Finally, $E \in \mathbb{R}^{d \times D}$ is the token embedding matrix where $D$ is the size of the model's vocabulary and $w_j \in \mathbb{R}^D$ is a 1-hot vector corresponding to the $j^{th}$ input token. It should be noted that the semantic type vector is set to a zero-filled vector for words that do not have a clinical meaning

### 4.5 Medical Weighted Loss Function

We update the loss function of the summarization task to provide a stronger incentive to correctly predict medical words. In our summarization model we use the cross-entropy loss of the Fairseq library (Ott et al., 2019) for the target word $x_t$ for each timestep $t$. We modify the loss function to a weighted loss function where the weight for all the medical words is higher in order to provide a stronger incentive to the model to correctly predict the words with a medical meaning. Specifically, the summarization loss is updated to :

$$loss = -logP(x_t) * w_t \quad (2)$$

where $w_t = 1$ for all the non-medical words and $w_t = 1 + \alpha$ for all the medical words, where $\alpha$ is an additional weight value for these words.

### 4.6 Discussion

Previous work (UmlsBERT (Michalopoulos et al., 2021)) introduced a semantic type embedding, for the medical words that could be tokenized into a single token. Our semantic type signal extends the semantic policy for all the medical words (i.e multi-token words). Also, our medical guidance signal is the first attempt to 'guide' a summarization model by combining the dual-encoder architecture with structured medical information. Finally, our loss function, which incorporates a different weight for the medical terms, has not been used in prior work.

## 5 Experiments

### 5.1 Results

We report the results of the comparison of our proposed MedicalSum model with the baseline pointer generator model (Enarvi-PG) (Enarvi et al., 2020). We also experiment with a model which contains only the guidance signal (MedicalSum$_{guidance}$), a model that only includes the semantic type embedding (MedicalSum$_{semantic}$), and a model with the medical weighted loss function (MedicalSum$_{loss}$). These models are trained for a maximum of 20k steps using the Fairseq library (Ott et al., 2019) on PyTorch 1.5.0 on V100 GPU with 32G GB of system RAM on Ubuntu 18.04.3 LTS.

#### 5.1.1 Hyperparameter tuning

We provide the search strategy and the bound for each hyperparameter: the batch size is set between 4 and 8, and the $\alpha$ parameter of the medical weight loss is tested with the values $0.01$, $0.1$, and $0.2$. The best values are chosen based on the validation set micro ROUGE-1 F1 values, using the scoring code with the same setting, that is provided with the family medicine dataset For the Enarvi-PG, MedicalSum, and the models with each individual medical signal, the batch size is set to 4 and the medical weight loss parameter to $0.01$.

We run our model on three different (random) seeds and we provide the average scores and standard deviation. We compare the models on the ROUGE-1 F1 score (the overlap of unigram) and ROUGE-L F1 score (the lengths of the longest common subsequences) between the summary and the output of the model.

#### 5.1.2 Summarization model comparison

The mean and standard deviation of ROUGE-1 F1 and ROUGE-L F1 for all the competing models on the test set of each dataset are reported in Table 2 (we also provide the results on the validation set in Appendix A.2). MedicalSum outperforms the Enarvi-PG baseline on all the datasets. It achieves an improvement between 0.8% (on the publicly available radiology dataset) and 2.1% (on the PE section, where the ROUGE-1 improvement from 66.11 to 68.22 is a 6.2% reduction in error). These results indicate that the combination of all three previously mentioned medical signals can indeed boost the performance of a medical summarization model. We also provide a qualitative review of summaries produced by each model variant in Appendix A.1, where we observe that MedicalSum can generate clinical notes with desirable medical terms missing from the output of the baseline Enarvi-PG model. MedicalSum$_{semantic}$, MedicalSum$_{loss}$ model, and the Enarvi-PG baseline model have similar running times (117K seconds for the family medicine and 64K seconds for the radiology dataset). MedicalSum and the MedicalSum$_{guidance}$ are slower (by 4%) due to the second 'guidance' encoder.

| TEST | | | | | |
|---|---|---|---|---|---|
| Model | Micro F1 | HPI | PE | AP | RAD |
| *Enarvi-PG* | Rouge-1 | 48.04 ± 0.4 | 66.11 ± 0.3 | 43.02 ± 0.4 | 27.01 ± 0.2 |
| | Rouge-L | 34.21 ± 0.3 | 63.15 ± 0.2 | 36.19 ± 0.3 | 25.01 ± 0.3 |
| $MedicalSum_{loss}$ | Rouge-1 | 48.64 ± 0.2 | 67.37 ± 0.2 | 43.85 ± 0.4 | 27.34 ± 0.2 |
| | Rouge-L | 34.32 ± 0.3 | 63.77 ± 0.3 | 36.67 ± 0.5 | 25.37 ± 0.2 |
| $MedicalSum_{guidance}$ | Rouge-1 | 48.79 ± 0.3 | 68.02 ± 0.2 | 43.72 ± 0.5 | 27.57 ± 0.2 |
| | Rouge-L | 35.14 ± 0.3 | 64.17 ± 0.2 | 36.65 ± 0.3 | 25.66 ± 0.2 |
| $MedicalSum_{semantic}$ | Rouge-1 | 48.90 ± 0.2 | 67.80 ± 0.3 | 43.64 ± 0.4 | 27.56 ± 0.3 |
| | Rouge-L | 34.79 ± 0.2 | 63.93 ± 0.2 | 36.42 ± 0.2 | 25.39 ± 0.3 |
| $MedicalSum$ | Rouge-1 | **48.98 ± 0.3** | **68.22 ± 0.2** | **44.54 ± 0.3** | **27.77 ± 0.3** |
| | Rouge-L | **35.22 ± 0.3** | **64.48 ± 0.3** | **37.34 ± 0.2** | **26.06 ± 0.2** |

Table 2: Results of mean ± standard deviation for each model on the test set; best values are **bolded**

We chose to compare our model with the Enarvi-PG model (Enarvi et al., 2020), as it has achieved state-of-the-art results in a similar medical summarization dataset. In addition, in their experimentation setup, they actually compared their model with other summarization models like the model of (See et al., 2017) and showcased that their model outperformed it in the task of medical summarization.

We did not re-do the experiments multiple times with different splits in order to be consistent with the literature in terms of testing. For both datasets the splits were provided by the team who created them and creating new splits will not provide a fair comparison with other (current and future) research models that will be tested on these datasets. However, we run each model multiple times (with different random seeds) and we provide the average scores and standard deviation for the testing and the validation set in order to be sure that the improvement was not due to the random seed.

#### 5.1.3 Ablation Study

In order to understand the effect that each medical signal has on the model performance, we conduct an ablation test where the performance of three variations of the MedicalSum model are compared, where each model is allowed access to only one of the medical signals. The results of this comparison are listed in Table 2.

We observe that for every dataset, MedicalSum achieves its best performance when all the medical signals are available, and each model that has access to any of the medical signals outperforms the baseline model. The guidance signal (MedicalSum$_{guidance}$) appears to have the most positive effect as it can guide the model to the most important sections of each input. Also, enriching the input embedding with semantic information (MedicalSum$_{semantic}$) appears to boost the performance of the model as it forces the embeddings of words that are associated with the same semantic type to become more similar in the embedding space. The medical weight loss model (MedicalSum$_{loss}$) appears to have the least improvement but it still outperformed the baseline.

## 6 Conclusion and Future Work

In this paper, we present MedicalSum, a novel approach for medical summarization. MedicalSum can provide external medical guidance that helps key information pass the model's decision process and appear in the summary. Furthermore, its novel weighted loss function provides a stronger incentive to the model to correctly predict words with a medical meaning. MedicalSum can also create more meaningful input embeddings by forcing the embeddings of the words that are associated with the same semantic type to become more similar. Our analysis shows that these features allowed MedicalSum to produce more accurate AI-generated medical documentation. Future work includes examining additional guidance signals (e.g., relational triples), and exploring UMLS hierarchical associations.

This work is the first to show how external medical domain (UMLS) knowledge can effectively improve the performance of a medical note-generation model. Leveraging external knowledge may become an important component of scaling and improving future medical AI systems that automatically generate medical documentation to combat physician burnout and improve patient care.

## Limitations

In this paper, we present MedicalSum, a novel medical conversation summarization model which achieves state-of-the-art ROUGE score improvements by integrating structured medical knowledge into the summarization process of a contextual word embedding model. However, one of the obstacles for adopting such a model in any system lies in the computing cost of training. For example, our MedicalSum model was trained on V100 GPU with 32G GB of system RAM on Ubuntu 18.04.3 LTS, and we acknowledge that investing in these types of computational resources is not a viable option for many research groups, let alone regular healthcare providers. In addition, another limitation of our work is that relies on the existence of an external medical metathesaurus (UMLS) and thus our model may not be easily adapted to other languages for which a detailed medical database (such as the UMLS for the English language) may not exist.

## Ethical Consideration

Medical Note generation by abstractive summarization is crucial for reducing physician burnout due to the vast amount of documentation requirements for electronic health records (EHR). Traditionally, clinical professionals review clinical documents and manually create the appropriate summaries by following specific guidelines. Models such as our MedicalSum model could help to reduce physician burnout, as well as enable physicians to devote more quality time and attention to their patients.

However, we need to be aware of the risks of over-relying on any automatic abstractive summarization model. No matter how efficient a summarization model is, it is still possible to omit key information or to hallucinate unsupported information. This is especially of concern in the medical domain, as inaccuracies could have a significant adverse effect on future patient health outcomes. Thus we believe that any automatic summarization model should only be used to assist, not replace trained clinical professionals.

## References


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

O. Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–270.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California. Association for Computational Linguistics.

Ish Kumar Dhammi and Sudhir Kumar. 2014. Medical subject headings (mesh) terms. *Indian journal of orthopaedics vol.*, 48,5.

Zi-Yi Dou, Pengfei Liu, Hiroaki Hayashi, Zhengbao Jiang, and Graham Neubig. 2021. 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*, pages 4830–4842, Online. Association for Computational Linguistics.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, B. Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam R. McGrath, Yue Pan, Joel Pinto, Luca Rubini, Miguel Ruiz, Gagandeep Singh, Fabian Stemmer, Weiyi Sun, Paul Vozila, Thomas Lin, and Ranjani Ramamurthy. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *NLPMC*.

Sarthak Garg, Stephan Peitz, Udhyakumar Nallasamy, and Matthias Paulik. 2019. Jointly learning to align and translate with transformer models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4453–4462, Hong Kong, China. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. Ctrlsum: Towards generic controllable text summarization. *ArXiv*, abs/2012.04281.

Serena Jeblee, Faiza Khattak, Noah Crampton, Muhammad Mamdani, and Frank Rudzicz. 2019. Extracting relevant information from physician-patient dialogues for automated clinical note taking. pages 65–74.

Alistair Johnson, Tom Pollard, Seth Berkowitz, Nathaniel Greenbaum, Matthew Lungren, Chih-ying Deng, Roger Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317.

Anirudh Joshi, Namit Katariya, X. Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Lingustics: EMNLP 2020*.

Muhammad Khalifa, Miguel Ballesteros, and Kathleen McKeown. 2021. A bag of tricks for dialogue summarization. *ArXiv*, abs/2109.08232.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel D. Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *ArXiv*, abs/2205.02549.

Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A Folarin, Angus Roberts, Rebecca Bendayan, Mark P Richardson, Robert Stewart, Anoop D Shah, Wai Keong Wong, Zina Ibrahim, James T Teo, and Richard J B Dobson. 2021. Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. *Artif. Intell. Med.*, 117:102083.

Stephan Kudyba. 2010. *Healthcare informatics: Improving efficiency and productivity*. CRC Press. Publisher Copyright: © 2010 by Taylor & Francis Group, LLC.

J. Kupiec, Jan O. Pedersen, and Francine R. Chen. 1995. A trainable document summarizer. In *SIGIR '95*.

Ronilda C Lacson, Regina Barzilay, and William J Long. 2006. Automatic analysis of medical dialogue in the home hemodialysis domain: structure induction and summarization. *Journal of biomedical informatics*, 39(5):541–555.

Chenliang Li, Weiran Xu, Si Li, and Sheng Gao. 2018. 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)*, pages 55–60, New Orleans, Louisiana. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Shashi Narayan, Yao Zhao, Joshua Maynez, Gonçalo Simões, Vitaly Nikolaev, and Ryan McDonald. 2021. Planning with learned entity prompts for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 9:1475–1492.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Rimma Pivovarov and Noémie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 22.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

Brian D Tran, Yunan Chen, Songzi Liu, and Kai Zheng. 2020. How does medical scribes' work inform development of speech-based clinical documentation technologies? A systematic review. *Journal of the American Medical Informatics Association*, 27(5):808–817.

M. M. van Buchem, H. Boosman, M. P. Bauer, I. Kant, S. Cammel, and E. Steyerberg. 2021. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digital Medicine*, 4.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

World Health Organization WHO. 2004. Icd-10 : international statistical classification of diseases and related health problems : tenth revision.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 2020. Enhancing factual consistency of abstractive summarization. *arXiv preprint arXiv:2003.08612*.

Xiaodan Zhu and Gerald Penn. 2006. Summarization of spontaneous conversations. In *Ninth International Conference on Spoken Language Processing*.

# A Appendix

## A.1 Qualitative Model Output Comparison

We qualitatively evaluate some of the differences in summaries produced by each model variant and illustrate how each feature contributes to the creation of accurate medical summaries.

Table 3 shows a sample target summary and we compare it with the output of the Enarvi-PG model and the medical signals of MedicalSum. It should be noted that the only criteria for choosing these examples is that they should contain medical entities. Some of the shorter input texts contain no medical entities, and we do not include those here as they would not showcase the differences between the baseline and the guided medical summarization model. We observe that the summary from the pointer generator (Enarvi-PG) baseline model does not include the following medical information: (i) the medicines: ibuprofen, Advil, or Aleve and (ii) x-rays. However, the MedicalSum$_{guidance}$ model has included in its summary the medical entities x-ray and NSAID (Non-steroidal anti-inflammatory drugs) as they are identified as medical entities and are part of the guidance signal. In addition, the MedicalSum$_{loss}$ model has included the x-ray token as a result of providing a stronger incentive to predict 'medical' words by passing a higher weight to the words that have a medical meaning. The MedicalSum model includes all the previously mentioned medical words as it takes advantage of all the medical signals.

Another example that showcases the strength of the medical signals can be found in Table 4. The summary from the Enarvi-PG model lacks two important medical facts: (i) ophthalmic drops, and (ii) Dyslipidemia. However, the MedicalSum$_{guidance}$ and the MedicalSum$_{loss}$ models have included the word 'Dyslipidemia' as it is identified as a medical word. The MedicalSum$_{semantic}$ model has also included the

| model | output |
|---|---|
| target-summary | ASSESSMENT AND PLAN –lf1– Acute left hip pain . –lf1– The patient has been experiencing acute left hip pain since . The pain is in her sacroiliac joint on her left side . I have prescribed a short course of prednisone to take in the morning after breakfast . I have instructed her not to take **ibuprofen , Advil , or Aleve** . I have also instructed her to apply heat to her back . we can get an **x-ray** of her lower back. |
| *Enarvi-PG* | ASSESSMENT AND PLAN –lf2– 1 . Left hip pain . –lf1– . I recommend the patient take a short course of prednisone . I have advised her to apply heat to the affected area . –lf2– 2 . Preventative care .–lf2– Instructions –lf1– |
| *Medical-Sum$_{semantic}$* | Assessment and Plan : –lf2– 1 . Acute left hip pain . –lf1– the left hip pain is sciatic pain. I recommend the patient take a short course of prednisone for 10 days . She can take it in the morning after breakfast . I also recommend she take a shower going to the kitchen daily . I also recommend she can also try applying heat to the affected area . |
| *Medical-Sum$_{guidance}$* | ASSESSMENT AND PLAN –lf2– 1 . Acute left hip pain . –lf1– I have prescribed a short course of prednisone . I advised the patient to take this in the morning after breakfast . I advised the patient to apply heat to the affected area for 10 days .advised the patient to avoid **NSAIDs** . I have ordered an **x-ray** of the lower back . |
| *Medical-Sum$_{loss}$* | ASSESSMENT AND PLAN –lf2– 1 . Acute left hip pain . –lf1– I have prescribed her a short course of prednisone . If her symptoms do not improve , I will refer her to physical therapy to strengthen the muscle of her back . –lf2– 2 . Preventative care . –lf1– The patient is due for a **x-ray** . |
| *MedicalSum* | ASSESSMENT AND PLAN –lf1– 1 . Acute left hip pain . –lf1– I suspect the left hip pain is sciatic pain . I will order an **x-ray** of the left hip . I have instructed the patient to apply heat to the lower back and see unusual daily chores . I advised the patient to avoid **NSAIDs** I recommended that she perform acute exercise when she is acutely inflamed after 10 days. |

Table 3: First example of distinct output from summarization models of different medical signals

| model | output |
|---|---|
| target-summary | The left eye swelling with burning and itching has been present for 2 weeks . I will prescribe olopatadine 0.2 percent **ophthalmic drops** . If the symptoms do not improve , the patient will return and consider bacterial cause .–lf2– **Dyslipidemia** . |
| *Enarvi-PG* | ASSESSMENT AND PLAN : The patient has allergic conjunctivitis of left eye . –lf1– - Patient Education and Counseling : The patient was advised to keep wiping green and thick mucus from the eye |
| $Medical\text{-}Sum_{semantic}$ | ASSESSMENT AND PLAN : –f1– 1 . Allergic conjunctivitis of left eye. –f1– The patient was advised to continue using his **eye drops** . |
| $Medical\text{-}Sum_{guidance}$ | ASSESSMENT/PLAN –f1– Allergic conjunctivitis . I recommended that the patient continue daily loratadine .–lf2– **Dyslipidemia** |
| $Medical\text{-}Sum_{loss}$ | ASSESSMENT/PLAN –lf1– Allergic conjunctivitis –lf1– –lf1– Medical Treatment : The patient will continue to take loratadine –lf2– **Dyslipidemia** |
| $MedicalSum$ | ASSESSMENT/PLAN –f1– Allergic conjunctivitis . –f1– - Medical Treatment : The patient will continue daily loratadine and **eye drops** .–lf2– **Dyslipidemia** |

Table 4: Second example of distinct output from summarization models of different medical signals

medical concept 'eye drops' as a replacement for 'ophthalmic drops'. 'Eye' and 'ophthalmic' have the same semantic type in UMLS and thus the model has the ability to learn their medical meaning even if one of these words (ophthalmic) is not popular in the training set. Finally, the MedicalSum model includes all of the previously mentioned medical words.

These examples demonstrate how, in addition to improving ROUGE scores, the MedicalSum model also generates clinical summaries that contain more relevant medical facts. In particular, they showcase that a guided medical summarization model can help with the omission of key information, which is especially of concern in the medical domain, because if medical key information is missing from the output, future readers may not have the ability to make an accurate diagnosis.

### A.2 Validation Set Comparison

In order to have a complete comparison with the baseline model, we present in Table 5 the mean and standard deviation of ROUGE-1 F1 and ROUGE-L F1 for all the competing models on the validation set of each dataset. MedicalSum outperforms the Enarvi-PG baseline on all the datasets. Also, MedicalSum achieves its best performance when all the medical signals are available, and each model that has access to any of the medical signals outperforms the baseline model. The results in Table 5 (validation set) and in Table 2 (test set) showcase the positive effect of the medical signals on the performance of a medical summarization model.

| Model | Micro F1 | HPI | PE | AP | RAD |
|---|---|---|---|---|---|
| | | **VALID** | | | |
| *Enarvi-PG* | Rouge-1 | 48.17 ± 0.3 | 67.44 ± 0.2 | 43.23 ± 0.4 | 29.91 ± 0.3 |
| | Rouge-L | 34.88 ± 0.3 | 64.68 ± 0.2 | 36.39 ± 0.3 | 29.95 ± 0.3 |
| $MedicalSum_{loss}$ | Rouge-1 | 49.29 ± 0.2 | 67.89 ± 0.2 | 44.02 ± 0.3 | 30.32 ± 0.3 |
| | Rouge-L | 34.94 ± 0.3 | 64.33 ± 0.3 | 36.70 ± 0.2 | 30.14 ± 0.3 |
| $MedicalSum_{guidance}$ | Rouge-1 | 49.55 ± 0.3 | 68.18 ± 0.3 | 44.32 ± 0.4 | 30.35 ± 0.2 |
| | Rouge-L | 35.14 ± 0.3 | 64.66 ± 0.2 | 37.01 ± 0.3 | 30.81 ± 0.2 |
| $MedicalSum_{semantic}$ | Rouge-1 | 49.39 ± 0.3 | 68.02 ± 0.2 | 44.16 ± 0.4 | 30.30 ± 0.2 |
| | Rouge-L | 34.99 ± 0.4 | 64.41 ± 0.3 | 36.90 ± 0.5 | 30.50 ± 0.2 |
| $MedicalSum$ | Rouge-1 | **49.68 ± 0.2** | **68.37 ± 0.3** | **44.98 ± 0.3** | **30.63 ± 0.3** |
| | Rouge-L | **35.43 ± 0.2** | **64.83 ± 0.2** | **37.90 ± 0.2** | **31.45 ± 0.3** |

Table 5: Results of mean ± standard deviation for each model on the validation set; best values are **bolded**