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

Anonymous ACL submission

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 domain knowledge into consideration during the summarization process and achieves state-of-the-art ROUGE score improvements of 0.8-2 points (including 6.2% error reduction in PE section ROUGE-1) when producing medical summaries of patient-doctor conversations. Furthermore, a qualitative analysis shows that medical summaries produced by the knowledge augmented model contain more relevant clinical facts from the patient-doctor conversation.

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., 2021a; Tran et al., 2020). As a result, the automatic creation of medical documentation has been proposed as one way to address this issue. For instance, automatic speech recognition (ASR) for dictating medical documents has contributed significantly to the efficiency of physicians in creating narrative reports (Payne et al., 2018).

Medical Note generation by abstractive summarization is another approach to automating clinical documentation and aims to decrease the workload associated with creating summaries of clinical encounters. It does this by taking a transcript of a patient-doctor conversation as input and automatically providing a summary of relevant clinical discussion (Finley et al., 2018). The extracted information can then be passed on to other healthcare providers, who may use it for the creation of clinical notes or billing codes (van Buchem et al., 2021b). The expected impact is reduced physician burnout, as well as enabling physicians to devote more quality time and attention to their patients.

To date there have been several attempts at automatically generating summaries of clinical encounters. Enarvi et al. (2020) created a transformer model for summarizing doctor-patient conversations. Joshi et al. (2020) demonstrated the effect of considering medical knowledge in the summarization of dialogue snippets. Finally, Jeblee et al. (2019) and Lacson et al. (2006) used extractive methods to identify the most important utterances which are combined to form the final summary.

However, 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 in generated notes, such as the omission of key information, or the hallucination of 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 a medical dialogue summarization model should capture specific parts of the conversation (Joshi et al., 2020) that are needed for a medical decision. 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.

The Unified Medical Language System (UMLS) Metathesaurus (Bodenreider, 2004) is a compendium of many biomedical terminologies with associated information, such as synonyms and categorical groupings. It allows the grouping of concepts according to their semantic type (McCray et al., 2001). For example, 'migraine' and 'epistaxis' are in the 'Disease or Syndrome' semantic type. Our model uses UMLS to identify the words that have a medical meaning and to create semantically enriched representation for each of these words.

Specifically, in this paper, we designed 3 specific strategies to leverage medical information: (i) We propose a novel guided summarization signal which consists of all the words in the input sentence with a medical meaning in order to demonstrate to the model the importance of these words in the creation of an accurate summary. (ii) In addition, we introduce a semantic type embedding that enriches the input embedding process of the model by forcing the model to take into consideration the associations between words that have the same semantic type. (iii) Finally, we update the standard loss function of a Transformer summarization model into a novel weighted loss function which provides a stronger incentive to the model to correctly predict 'medical' words by passing a higher weight to these words.

We present a novel architecture for integrating medical knowledge during the summarization process via a novel knowledge augmentation strategy. Key paper contributions include:

1. We are the first, to the best of our knowledge, to propose the usage of medical knowledge from a clinical Metathesaurus (UMLS) in the summarization process of a Transformer-based model (MedicalSum) in order to generate 'medically focused' clinical note summaries from full encounter transcripts.

2. We answer the question of how to incorporate structured medical knowledge in medical documentation generation by designing 3 specific signals over medical entities and their connections (details in the earlier paragraph), implementing them in MedicalSum, and evaluating them.

3. By leveraging these methods the MedicalSum model achieved a ROUGE-1 and ROUGE-L improvement between 0.8% and 2.0% in all experiments on medical note summarization. In addition, qualitative analysis verified MedicalSum's ability to better determine which key information (medical terms) should pass the model's decision process and appear in the generated summaries.

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. Li et al. (2018) included a set of keywords that are incorporated into the generation process. Zhu et al. (2020) proposed the usage of relational triples (subject, relation, object). Finally, in Dou et al. (2021) the authors created a guided summarization framework which 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. Furthermore, 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 had to choose a large enough dataset that would provide the necessary data for the medical signals to meaningfully affect the model's performance. However, there are no publicly available large scale datasets for medical summarization and thus we had to use a proprietary one. We use English data consisting of encounters in a family medicine setting. These encounters were recorded at the time of the encounter and then transcribed using ASR. Data also includes the associated clinical note summaries (which are stored in a HIPAA compliant environment and the required steps have been taken to protect them).

The reports are organized under three sections that corresponds to three broad areas of a medical note: (i) History of Present Illness (HPI), which captures the reason for visit, and the relevant clinical and social history. (ii) Physical Examination (PE), which captures both normal and abnormal findings from a physical examination. (iii) Assessment and Plan (AP), which captures the assessment by the doctor and the treatment plan. We report experimental results on a dataset that consists of around 40,000 encounters for each section. Table 1 shows detailed statistics of our dataset.

	Train	Valid	Test	A.W	P.D (%)
AP	42106	648	2525	2586	99.2
HPI	43092	657	2551	2584	96.9
PE	39815	635	2442	2633	91.7
RAD	91544	2000	600	49	100

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 number of word for each section; P.D is the percentage of distribution per section (percentage of encounters which have the section in their report).

In order to allow for a more open comparison, we also experimented with a public dataset. We tackle the third task of the MEDIQA 2021 challenge (Ben Abacha et al., 2019) of automatic summarization of English radiology reports (RAD) of the MIMIC-CXR dataset (Johnson et al., 2019) (license: https://tinyurl.com/mimic-licence). From the Table 1, it can be observed that the input documents in the MEDIQA dataset are much smaller than the documents of the other real-world datasets that we experimented on and thus they contain less medical information. However, we have included 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. We manually investigated the existence of information that names individual people or offensive content, but we did not find any indication of either of them.

4 Method

4.1 MedicalSum: Medical Guided Transformer Pointer Generator Model

We adopt the transformer self-attention model (Vaswani et al., 2017) 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.

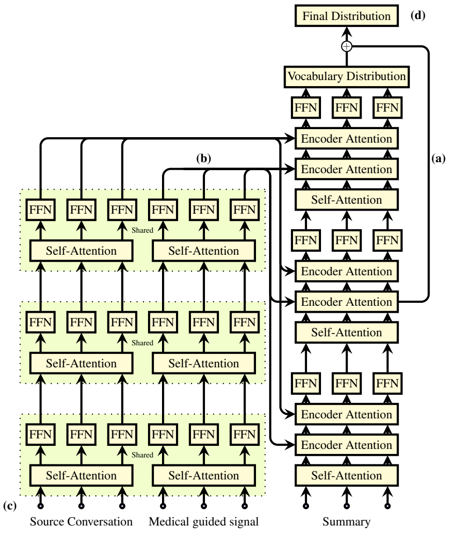

Figure 1: Illustration of **MedicalSum** a transformer sequence-to-sequence model with a pointer-generator and guidance mechanism.

A simplified image of the MedicalSum model can be found in Figure 1. We improve the performance of our model by introducing to the standard transformer encoder-decoder model for summarization (i) a pointing mechanism for copying out-of-vocabulary (OOV) words from the source document (part (a) in Figure 1), (ii) a novel guided summarization signal which consists of all the medical words in the input sentence in UMLS (part (b)), (iii) a new semantic type embedding that enriches the input embeddings process (part (c)) (iv) a novel weighted loss function which provides a stronger incentive to the model to correctly predict medical words (part (d)). The details of each added component are discussed in the following sections.

4.2 Pointer-Generator

First, we implement the pointer generator network as described in (Enarvi et al., 2020; See et al., 2017). Because the Transformer model creates several encoder-decoder attention distributions, we can choose any distribution over the source tokens for the copying mechanism. Following (Enarvi et al., 2020) we chose to use a single attention head in order to only train the parameters of a single head to attend to the tokens that are good candidates for copying. Finally, in (Garg et al., 2019) it was stated that the penultimate layer seems to naturally learn alignments, so we chose to use its first attention head for pointing (Enarvi et al., 2020).

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 both the input and the guidance signal consists of a self-attention block and feed-forward block. In addition, each decoder layer consist of a self-attention block, a cross-attention block with the medical guidance signal in order to inform the decoder which section 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. 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.

There are other research works that used external knowledge from knowledge bases to enhance the performance of deep learning models such as the model in (Soares et al., 2019). However, we believe the approach of marking the entities is not fitted for the task of abstractive summarization as the markers will change the format of the sentence and they will affect the performance of summarization models especially in the case when they are pre-trained on large corpus (which is mostly the case for transformer-based models). Finally, there is no significant overhead by the additional encoder as the two encoder share their weights.

4.4 Semantic Type Embeddings

We also introduce a new embedding matrix called $S \in \mathbb{R}^{D_s \times d}$ into the input layer where d is transformer hidden dimension (512) and $D_s = 50$ is the number of unique UMLS semantic types that are relevant to the domain of the dataset.

To incorporate the S embedding matrix into the input embedding layer, all the words with a clinical meaning defined in UMLS are identified (using the MedCAT toolkit (Kraljevic et al., 2021)) and their corresponding semantic type is extracted. By introducing the semantic type embedding, the input vector for each word w_j is updated to:

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

where $s_{wj} \in \mathbb{R}^{D_s}$ is a 1-hot vector corresponding to the semantic type of the medical word w_j (the semantic type vector $S^\top s_{wj}$ is set to a zero-filled vector for words that are not identified in UMLS) 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 where $D = 48128$ 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.

Previous research work (UmlsBERT (Michalopoulos et al., 2021)) demonstrated that the inclusion of semantic type vectors had a positive effect on the performance of a contextual model in various downstream task as the semantic type embeddings can provide more accurate input vectors for the medical words that are rare in the training corpus. Our model differs from previous work as it extends the semantic policy for all of the medical words that could be identified instead of only including semantic embeddings for the words that could be tokenized in a single token (e.g. our model included the semantic type of the word 'x-ray' but UmlsBERT did not).

4.5 Medical Weighted Loss Function

We update the loss function of the summarization task to provide a stronger incentive to correctly pre-

TEST					
Model	Micro F1	HPI	PE	AP	RAD
$Enarvi\text{-}PG$	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**
VALID					
$Enarvi\text{-}PG$	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 2: Results of mean ± standard deviation for each model on the test/validation set; best values are **bolded**

dict medical words. In our summarization model we are using 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 of 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 :

$$loss = -logP(x_t) * w_t \tag{2}$$

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

4.6 Discussion

Previous work (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) in order to capture all of the relevant medical information. Also, our novel medical guidance signal is the first attempt to 'guide' a summarization model by combining the dual-encoder architecture with structured medical information. Finally, our new loss function, which incorporates a higher weight for all 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 (Enarvi-PG) pointer generator model (Enarvi et al., 2020). We also experiment with variations of our model that only contain, a) the guidance signal (MedicalSum$_{guidance}$) where the guidance signal is composed by all the medical words; b) the semantic type embedding (MedicalSum$_{semantic}$); and c) the medical weighted loss function (MedicalSum$_{loss}$), in order to measure how each signal individually affects the model's performance. These models are implemented using the Fairseq library (Ott et al., 2019) on PyTorch 1.5.0. All experiments are executed on V100 32 GB GPU with 32G GB of system RAM on Ubuntu 18.04.3 LTS.

We use a vocabulary consisting of the 45k most frequent words. The same vocabulary is shared between the source and the target tokens. We train the models a maximum of 20k steps. It should be noted that Enarvi-PG, the MedicalSum$_{guidance}$ and MedicalSum$_{loss}$ model have the exact same number of (74,724,353) parameters (the input and the 'guidance' encoder share their weights). However, MedicalSum and the MedicalSum$_{semantic}$ model have an additional 25,600 parameters due to the inclusion of the semantic type embeddings.

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 α parameter of the medical weight loss was 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. In the interest of providing a fair comparison, we tune the hyperparameters of each model. For the Enarvi-PG, MedicalSum and the models with each individual medical signal, the batch size was set to 4 and the medical weight loss parameter was set to be 0.01.

In order to achieve more robust results, we run our model on three (random) seeds and we provide the average scores and standard deviation for the testing and the validation set. We compare the models on the ROUGE-1 F1 score that is based on the overlap of unigram and ROUGE-L F1 score that is based on the lengths of the longest common subsequences between the actual summary and the output of the model. The ROUGE scores are calculated by using the scoring code, that was provided with the family medicine dataset[1].

5.1.2 Summarization model comparison

The mean and standard deviation of ROUGE-1 F1 and ROUGE-L F1 for all the competing models are reported in Table 2. MedicalSum outperforms the pointer generator (Enarvi-PG) baseline on all the datasets due to the fact that all the (three) previous mentioned medical signals have a positive contribution on its performance (subsection 5.1.3) as they encourage MedicalSum to take into consideration different medical information (section 5.2). It achieved an improvement between 0.8% (on the radiology dataset) and 2% (on the PE section, where the ROUGE-1 improvement from 66.11 to 68.22 is a 6.2% reduction in error). The MedicalSum$_{semantic}$, the MedicalSum$_{loss}$ and the Enarvi-PG model have similar running times (117K seconds for the HPI, AP and PE sections and 64K seconds for the radiology dataset). MedicalSum and the MedicalSum$_{guidance}$ are always slower (by 4%) due to the introduction of the second encoder.

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 time 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. MedicalSum achieves its best performance when all the medical signals are available. However, each model that has access to any of the medical signals outperforms the baseline model. The guidance signal (MedicalSum$_{guidance}$) seems to have the most positive effect as it can more clearly guide the model to the most important sections of the input. On the other hand, the medical weight loss (MedicalSum$_{loss}$) seems to have the least influence over the model's performance but as we show in the qualitive analysis (Section 5.2), it can aid the model to focus on medical information. Furthermore, enriching the input embedding with semantic information (MedicalSum$_{semantic}$) seems 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. These experiments provide further evidence of the positive effect of each of the medical signals on the performance of the MedicalSum model. Finally, in

[1] In order to conceal the identity of the authors we did not include the link of the scoring code in this version of the paper.

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.
$Enarvi\text{-}PG$	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–
$Medical\text{-}Sum_{semantic}$	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 .
$Medical\text{-}Sum_{guidance}$	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 .
$Medical\text{-}Sum_{loss}$	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: Distinct output from models of different medical signals

section 5.2, we provide a qualitative analysis where we demonstrate that each signal can 'guide' a summarization model to different medical information that the baseline model ignored.

5.2 Qualitative Model Output Comparison

We have shown how the MedicalSum model improves ROUGE scores compared to a state-of-the-art baseline. In this section, we qualitatively evaluate some of the differences in summaries produced by each model variant and show how each contributes to more accurate and 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 was that they should contain medical entities. Some of the shorter input text contained no medical entities, and we did 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 were identified as medical entities and were 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 included all the previous mentioned medical words as it take advantage of all the medical signals.

Another example which 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

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** .
$Enarvi\text{-}PG$	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: Distinct output from models of different medical signals

MedicalSum$_{guidance}$ and the MedicalSum$_{loss}$ models have included the word 'Dyslipidemia' as it was identified as a medical word. The effect of this is that it was part of the medical guidance signal for the MedicalSum$_{guidance}$ model and provided a higher weight during loss computation for the MedicalSum$_{loss}$ model. The MedicalSum$_{semantic}$ model has also included the medical concept 'eye drops' as an replacement of 'ophthalmic drops'. 'Eye' and 'ophthalmic' have the same semantic type in UMLS and thus the model had 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 included all of the previous 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 showcased 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.

6 Conclusion and Future Work

In this paper, we present MedicalSum, a novel approach for medical conversation summarization that integrates medical knowledge into the summarization process of a contextual word embeddings model. MedicalSum can provide external medical guidance that helps key information pass the model's decision process and appear in the summary. Its novel weighted loss function provides a stronger incentive to the model to correctly predict words with a medical meaning. Lastly, it creates more meaningful input embeddings by forcing the embeddings of the words that are associated with the same semantic type to become more similar by incorporating information from the semantic type of each biomedical word.

Our analysis showed that these features allowed MedicalSum to produce more accurate AI-generated medical documentation. MedicalSum achieves ROUGE score gains of 0.8 to 2 points (and up to 6.2% error reduction on the family medicine dataset), which is a respectable amount of gain for this task, and does a more complete job including medical entities that contain crucial information.

As for future work, we plan to address the limitations of this study including: (i) Investigating the generality of MedicalSum to additional datasets, (ii) Exploring UMLS hierarchical associations between words that extend the concept connection we investigated and (iii) Examining different guidance signals such as the inclusion of relational triples.

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.

Ethical Consideration

Medical Note generation by abstractive summarization has the potential to reduce physician burnout, which occurs, in part, as a result of 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 enabling 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 an 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 and 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*.

Greg P. Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *NAACL*.

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.

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.

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*.

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 *NAACL*.

Alexa McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84:216–20.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: 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.

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*.

Thomas H. Payne, W. David Alonso, J. Andrew Markiel, Kevin Lybarger, and Andrew A. White. 2018. Using voice to create hospital progress notes: Description of a mobile application and supporting system integrated with a commercial electronic health record. *Journal of Biomedical Informatics*, 77:91–96.

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

A. See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*.

Livio Baldini Soares, Nicholas Arthur FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905.

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. 2021a. The digital scribe in clinical practice: a scoping review and research agenda. *NPJ Digital Medicine*, 4.

Marieke van Buchem, Hileen Boosman, Martijn Bauer, Ilse Kant, Simone Cammel, and Ewout Steyerberg. 2021b. 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, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

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.