Data Augmentation for Radiology Report Simplification

Ziyu Yang
CIS, Temple University
zyyang@temple.edu

Santhosh Cherian
Temple University Hospital
santhosh.cherian@tuhs.temple.edu

Slobodan Vucetic
CIS, Temple University
vucetic@temple.edu

Abstract

This work considers the development of a text simplification model to help patients better understand their radiology reports. This paper proposes a data augmentation approach to address the data scarcity issue caused by the high cost of manual simplification. It prompts a large foundational pre-trained language model to generate simplifications of unlabeled radiology sentences. In addition, it uses paraphrasing of labeled radiology sentences. Experimental results show that the proposed data augmentation approach enables the training of a significantly more accurate simplification model than the baselines.

1 Introduction

Radiology reports are unstructured documents written by radiologists to communicate imaging findings to another physician or a qualified medical professional (Goldberg-Stein and Chernyak, 2019). Radiology reports have been increasingly available to patients through portals (Lourenco and Baird, 2020), which has been generally welcomed by patients (Cooper et al., 2020). However, the health literacy of most patients is insufficient to fully comprehend radiology reports (Lalor et al., 2018) because such reports rely on complex medical jargon and use explanations that imply highly specialized medical knowledge (Delbanco et al., 2012). Several studies even identified adverse effects of sharing radiology reports with patients, including dissatisfaction with care (Rosenkrantz and Flagg, 2015) and undue anxiety and stress (Arora, 2013).

There is an increasing need for patient-friendly radiology reporting that can communicate results clearly and be understandable by a diverse patient population. However, asking a radiologist to supplement a traditional report with a patient-friendly summary would negatively impact their cognitive load and productivity. This problem motivated recent research on the automatic simplification of health records. The proposed approaches include both lexical simplification that paraphrases text (Chen et al., 2018; Biran et al., 2011; Weng et al., 2018) and semantic simplification that seeks to simplify grammatically complex text (Shardlow, 2014; Leroy et al., 2016) which recently included deep learning approaches (Lewis et al., 2019; Zhang et al., 2020). However, training deep learning models for medical text simplification requires the collection of costly labeled data.

To alleviate the data scarcity issue in simplifying health reports, particularly radiology reports, this paper proposes a novel approach for data augmentation. It augments manually-created labeled data with simplifications generated by a large pre-trained language model such as GPT-3 (Brown et al., 2020). To improve the quality of data augmentation, the approach develops a separate deep learning model that evaluates the quality of generated simplifications. Furthermore, the approach also provides data augmentation through paraphrasing the originally labeled radiology sentences.

The proposed data augmentation approach is experimentally evaluated on a unique corpus of manually generated labeled data for radiology report simplification. The evaluation includes both automatic measures and human evaluation.

Our research claims are: 1) Our augmentation methods enable training of a more accurate model than baselines in solving low-resource radiology sentence simplification problems. 2) We address the challenge of selecting qualified augmentations for radiology sentence simplification. 3) We create unique real data containing expert-annotated simplifications for radiology reports' sentences regarding liver conditions.

2 Related Work

Text Simplification. In text simplification, the output text is a linguistically simplified version of the input text (Adduru et al., 2018). Previous work on

Findings of the Association for Computational Linguistics: EACL 2023, pages 1922–1932

simplification includes lexical and semantic simplification (Alva-Manchego et al., 2020).

Lexical simplification by lexical substitution refers to replacing complex words or phrases with simpler synonyms (Oh et al., 2016; Zeng and Tse, 2006) and has found some practical success (Cook et al., 2017). In the health domain, lexical text simplification often relies on medical dictionaries (UMLS (Bodenreider, 2004) , MeSH (Lipscomb, 2000), etc.). Lexical simplification approaches also include rule-based methods (Chen et al., 2018; Biran et al., 2011) and deep learning (Weng et al., 2018, 2019).

Semantic simplifications seek to simplify grammatically complex text by splitting long sentences into shorter ones, changing passive voice to active, resolving ambiguities and anaphora (Shardlow, 2014), splitting complex noun phrases (Leroy et al., 2016), or reducing morphological negations (Mukherjee et al., 2017). Recently, transformer encoder-decoder based pre-trained seq-to-seq models (Lewis et al., 2019; Zhang et al., 2020) were proved to be robust in solving text simplification problems. However, fine-tuning pre-trained models require large quantities of labeled data, which are costly and difficult to obtain in the health domain.

Previous research has explored different methods for text simplification in low-resource domains. To address data scarcity recent studies include unsupervised methods (Surya et al., 2018; Sakakini et al., 2020; Enayati et al., 2021) and reinforcement learning (Laban et al., 2021).

Data Augmentation is a method that automatically generates labeled data to enhance manually labeled data (Liu et al., 2020). One approach is to use paraphrasing to create different variants of the original or simplified sentences (Wei and Zou, 2019). Another approach is to use pre-trained language models to generate labeled data (Bayer et al., 2021). LAMBADA (Anaby-Tavor et al., 2020) augments data for text classification tasks by encoding labels in the input. Similarly, PromptDA (Wang et al., 2022) use language models to augment data for NLU tasks. Back-translation (Edunov et al., 2018) is used to generate different variants of the input text.

There are several public benchmark data sets that are related to our paper. There are paragraph level medical text simplifications (Devaraj et al., 2021) focusing on medical paper abstracts. There is a corpus parsed aligned sentences from Wikipedia and Simple English Wikipedia [1] (Pattisapu et al., 2020; Van den Bercken et al., 2019) that has been a popular text simplification benchmark. However, none of these data sets have properties similar to the radiology text simplification task.

3 Problem Definition

Let us assume we are given a labeled corpus for text simplification $\mathbf{D}_{Lab} = \{(\mathbf{X}_1, \mathbf{Y}_1), (\mathbf{X}_2, \mathbf{Y}_2), ..., (\mathbf{X}_n, \mathbf{Y}_n)\}$, where $\mathbf{X}_i$ is the i^{th} original document, $\mathbf{Y}_i$ is its simplification provided by a human expert, and n is the number of labeled documents. Let us also assume we are given an unlabeled corpus of documents $\mathbf{D}_{Unl} = \{\mathbf{X}_1, \mathbf{X}_2..., \mathbf{X}_m\}$, where m is the number of unlabeled documents. The objective of data augmentation is to automatically create a synthetic set $\mathbf{D}_{Syn} = \{(\mathbf{X}_1^*, \mathbf{Y}_1^*), (\mathbf{X}_2^*, \mathbf{Y}_2^*), ..., (\mathbf{X}_K^*, \mathbf{Y}_K^*)\}$, where $\mathbf{X}_i^*$ is one of the original documents from $\mathbf{D}_{Lab}$ or $\mathbf{D}_{Unl}$ or their derivative, and $\mathbf{Y}_i^*$ is its corresponding simplification from $\mathbf{D}_{Lab}$, its derivative, or an automatically generated simplification. $\mathbf{D}_{Syn}$ is appended to $\mathbf{D}_{Lab}$ and the resulting set $\mathbf{D}_{Aug} = \mathbf{D}_{Lab} \cup \mathbf{D}_{Syn}$ is called the augmented training corpus. The assumption is that a seq2seq model for text simplification trained on the augmented corpus will have higher simplification quality than the one trained on the original training corpus. This paper focuses on the radiology report simplification problem where $\mathbf{X}_i$ is a sentence and $\mathbf{Y}_i$ is its simplification. As a result, our augmentation approach explained in the next two sections is specifically tailored for this application.

4 Methodology

4.1 GPT-based Semantic Augmentation

We propose two types of augmentation. The first, referred to as semantic augmentation, relies on using the large-scale language model GPT-3 (Brown et al., 2020) to generate simplifications automatically. The second, referred to as lexical augmentation, relies on modifying original documents from the labeled corpus. The first augmentation type is described in this section, while the second type is described in the next.

4.1.1 Generating simplifications with GPT-3

It has been demonstrated that large-scale language models such as GPT-3 are capable of doing a wide

[1] simple.wikipedia.org

Table 1: Good, Not Simple Enough, and Incorrect examples of radiology liver sentences (**ORI**), expert-written simplifications (**SIM**), and generated simplifications from GPT-Curie (**GPT**).

ORI:	The liver demonstrates diffusely low attenuation, consistent with fatty infiltration
SIM:	There is increased amount of fat in the liver.
GPT:	There is a lot of fat in the liver. (Good)
ORI:	The liver is stable in appearance with again note made of a hypodense lesion in the medial left lobe adjacent to the gallbladder fossa
SIM:	The liver may have one or more tumors or masses.
GPT:	The liver has a small lesion on one side, close to the gallbladder. (Good)
ORI:	The liver shows diffuse low attenuation consistent with fatty infiltration
SIM:	There is increased amount of fat in the liver.
GPT:	The liver has diffuse low attenuation, indicative of fatty infiltration. (Not Simple Enough)
ORI:	The liver is unremarkable without intra- or extra-hepatic biliary dilatation
SIM:	The liver appears normal.
GPT:	The liver is not enlarged or tender without intra- or extra-hepatic biliary dilatation. (Not Simple Enough)
ORI:	Complex hepatic abscess predominantly occupying segments VI and VII of the liver is overall little changed in size compared to date, measuring approximately 7.9 (AP) x 8.7 (TRV) x 11.4 (CC) cm
SIM:	There is infection and pus seen within the liver.
GPT:	This means that the abscess is smaller than it was before, by about 3 inches on all sides. (Incorrect)
ORI:	The liver demonstrates prominent intrahepatic biliary ducts
SIM:	The liver ducts that carry digestive enzymes may be blocked.
GPT:	The liver has many bile ducts. (Incorrect)

variety of NLP tasks through prompting (Brown et al., 2020). Our preliminary experiments showed that the two largest GPT-3 models, Curie and Davinci, are surprisingly good at generating simplifications of radiology report sentences, despite never being trained explicitly for that task. Table 1 shows several representative examples of simplifications by the Curie model prompted by *My patient asks me to simplify this radiology sentence "<original sentence>":*. Each example shown contains an original radiology sentence, its simplification by GPT-3, and its simplification provided by collaboration between a radiologist and a layperson. The first two examples show that GPT-3 can provide factually correct and easy to understand simplifications comparable to the manually created simplifications. Next two examples show that GPT-3 might provide factually correct text that is not sufficiently simple. The final two examples show that GPT-3 might provide factually incorrect simplifications.

Our semantic augmentation approach uses GPT-3 to generate simplifications of unlabeled documents from $\mathbf{D}_{Unl}$ and add them to the augmented corpus $\mathbf{D}_{Aug}$. As noted in previous research (Liu et al., 2021) the choice of prompting can have a significant impact on the quality of the generated text and accuracy on a particular task.

Our prompting approach relies on the in-context learning that has been used with success with GPT-3 models. Instead of relying on costly fine-tuning of a language model, it pastes a few labeled examples into the prompt and asks the language model to generate label of an unlabeled example. In our specific application, we select K labeled examples $(\mathbf{X}, \mathbf{Y})$ from $\mathbf{D}_{Lab}$ and insert each of them into template '***Sentence:*** $< \mathbf{X} >$***; Simplification:*** $< \mathbf{Y} >$'. A triple pound sign, *###*, is used to separate templates for the K labeled examples. The prompt ends with '***Sentence:*** $< \mathbf{X} >$***; Simplification:***', where $\mathbf{X}$ is an unlabeled document from $\mathbf{D}_{Unl}$. GPT-3 model is expected to write a simplification by mimicking the style of the labeled examples from the prompt.

As noted in previous work (Brown et al., 2020) the success of prompting that uses in-context learning depends on the particular choice of K examples. Therefore, we select most related sentence simplification pairs from the training set $\mathbf{D}_{Lab}$ given any unlabeled document from $\mathbf{D}_{Unl}$. In detail, we use BERTScore (Zhang et al., 2019), which leverages the pre-trained contextual embeddings from BERT (Devlin et al., 2018) and matches words in unlabeled and labeled radiology sentences by cosine similarity. Thus, each prompt consists of K most related examples rated by BERTScore for an unlabeled sentence that is appended to the end. Moreover, we evaluate more example selection scenarios in our ablation study.

4.1.2 BERT-Checker

Language models such as GPT-3 provide token probabilities as their output. When generating text, one option is to use brute force and generate the most likely token. However, in the context of text simplification, the most likely tokens are not guaranteed to produce the best simplification. An alter-

native is to generate tokens by selecting among the most likely choices, which the temperature hyperparameter in GPT-3 can control. In our approach, we invoke a GPT-3 model N times for each prompt using a temperature higher than zero, which results in N different simplifications. Then, we automatically select the best one of the N generated simplifications and add it to the augmented corpus.

As seen in Table 1, some of the generated simplifications are good while others are not. Separating good from inadequate simplifications is a non-trivial challenge. Related work on automatic evaluation of the generated text includes GPT-3-ENS (Chintagunta et al., 2021), which measures the complexity of terms in simplifications, and GPT3Mix (Yoo et al., 2021), which treats the likelihood scores of generated labels as confidence scores. However, we found that the existing approaches are inappropriate for our application. Thus, we developed a novel approach called BERT-Checker.

BERT-Checker is a fine-tuned BERT model (Devlin et al., 2018) to a task similar to entailment. In particular, we convert our labeled corpus into training data matching the format of the entailment task. We add label 1 to each example from $\mathbf{D}_{Lab}$ to create positive examples in new training data set, $\mathbf{D'}_{Lab} = \{[(\mathbf{X}_i, \mathbf{Y}_i), 1]\}$. To create negative examples in $\mathbf{D'}_{Lab}$, we use four different strategies as outlined next:

- **Precision:** To ensure that simplification is closely related to the original text, we corrupt the original text $\mathbf{X}$ by replacing the medical terms with randomly selected medical terms, and generate negative example from labeled example $(\mathbf{X}, \mathbf{Y})$ as $[(corrupt(\mathbf{X}), \mathbf{Y}), 0]$.
- **Simplicity:** To penalize simplifications that are too similar to the original sentence, we create negative examples by using the original text as simplification, $[(\mathbf{X}, \mathbf{X}), 0]$.
- **Correctness:** To penalize incorrect simplifications, we randomly select two labeled examples $(\mathbf{X}_1, \mathbf{Y}_1)$ and $(\mathbf{X}_2, \mathbf{Y}_2)$ and create a negative example by mixing the original and simplified text, $[(\mathbf{X}_1, \mathbf{Y}_2), 0]$.
- **Robustness:** For labeled example $(\mathbf{X}, \mathbf{Y})$ we replace the simplification with an empty string or a sentence generated by a GPT-3 given the prompt 'Generate a radiology report sentence about liver' and high temperature of 0.8 to create negative example $[(\mathbf{X}, GPT()), 0]$.

Thus, for each positive example, we generate four negative examples. As a result, we can obtain a negative dataset $\mathbf{D'}_{Neg}$. We fine-tune Clinical BERT (Alsentzer et al., 2019) on the text entailment task using the generated data set.

4.2 Dictionary-based Lexical Augmentation

We propose lexical augmentation to supplement semantic augmentation described in the previous section. Lexical simplification refers to replacing complex terms in original documents $\mathbf{X}$ with their synonyms, which might also be complex. In the related work on text simplification of general-purpose text, EDA approach (Wei and Zou, 2019) paraphrases original documents by replacing randomly selected words or phrases with their synonyms in WordNet (Miller, 1995). We modify EDA by replacing only specialized medical terms.

Inspired by (Pattisapu et al., 2020; Hasan et al., 2016), we use medical dictionaries Medical Subject Headings (MeSH) (Lipscomb, 2000) and Unified Medical Language System (UMLS) (Bodenreider, 2004) to find the synonyms. We use pre-trained named entity recognition model (Honnibal and Montani, 2017) to extract medical terms from the original documents in labeled corpus $\mathbf{D}_{Lab}$. The medical terms are linked to Concept Unique Identifier (CUI) in UMLS and the concept_id in MeSH. Each medical code in UMLS and MeSH is mapped to a list of synonyms. We iteratively select a synonym to replace the medical term from the original document.

We illustrate the lexical simplification process in Fig 1, where *hepatic steatosis* in the sentence '*Probable diffuse hepatic steatosis*' is recognized as a medical term and replaced with its synonyms. In particular, CUI codes 'C0015695' and 'C2711227' are found to match *hepatic steatosis*, where the canonical names are *Fatty Liver* and *Steatohepatitis*. Similarly, 'D005234' from MeSH also provides several synonyms. This process identifies five synonyms used to create five different versions of the original document.

Once the synonyms for a medical term in original document $\mathbf{X}$ of labeled example $(\mathbf{X}, \mathbf{Y})$ are identified, we paraphrase the original document as $lexical(\mathbf{X})$ and generate an augmented example $(lexical(\mathbf{X}), \mathbf{Y})$. The new example is added to the augmented corpus $\mathbf{D}_{Aug}$.

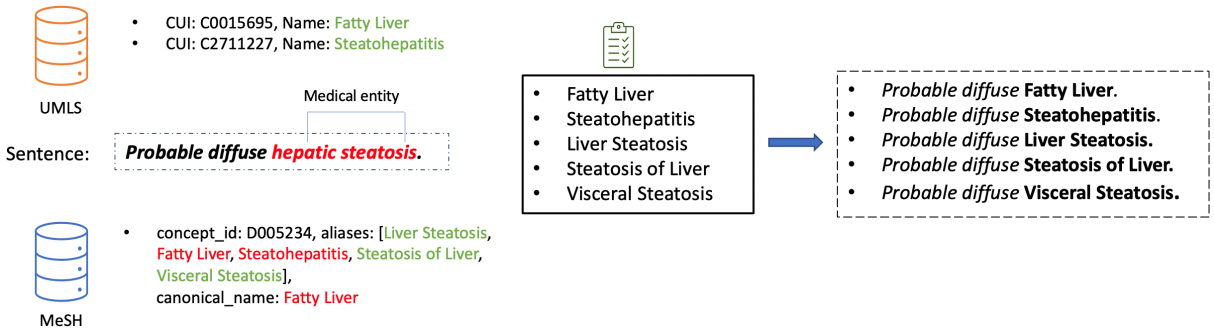

Figure 1: Workflow for lexical augmentation. It shows the linked synonyms of the entity "hepatic steatosis" from UMLS/MeSH and five synthetic sentences.

5 Experiments

5.1 Data

To the best of our knowledge, there is no readily available corpus for simplifying radiology sentences. To experimentally evaluate our data augmentation approach, we created a new corpus for this purpose. In particular, we collected 540 sentences from radiology reports describing the liver condition and manually created their simplifications: 170 sentences were obtained from CT-Abdomen radiology reports from a university hospital (UH), and the remaining 370 were extracted from CT-Abdomen radiology reports from publicly available MIMIC-III (Johnson et al., 2016) data. All sentences were de-identified with Health Insurance Portability and Accountability Act (HIPAA) standards in order to facilitate public accesses and human annotations.

We asked a radiologist to provide a simplification for each selected sentence. A layperson joined the radiologist to provide feedback about the generated simplifications. If the layperson thought the simplification was too complicated, this was communicated to the radiologist, who proceeded to improve the simplification. The process was repeated until the layperson could understand all the simplification and could correctly guess the severity of the described conditions.

During this sequence simplification process, the radiologist and the layperson agreed that it is sufficient to use simplification '*The liver looks normal*' for sentences explaining that nothing concerning was observed about the liver. 39% of the the university hospital sentences and 21% of the MIMIC-III sentences were simplified as '*The liver looks normal*'. For simplification of sentences that described concerning findings was to ignore technical details that might be confusing to patients. Any relevant medical terms were stated in simple terms familiar to laypeople. If possible, grammar was kept simple, and the sentences were kept short. Table 1 shows several examples of the original sentences (ORI) and their manual simplifications (SIM).

For our experiments, we randomly selected 100 sentences and their simplifications for training and the remaining 70 for testing for both the university hospital and MIMIC-III labeled data. Thus, we had 200 labeled examples for training denoted as $\mathbf{D}_{Lab}$, and 140 for testing. We used the remaining 200 MIMIC-III sentences as the unlabeled corpus $\mathbf{D}_{Unl}$ and used their simplifications to better evaluate the data augmentation approaches.

The corpus is available to the research community to support further research on medical text simplification.[2]

5.2 Data Augmentation

To implement the proposed semantic augmentation approach, we used GPT-3 Curie model (6.7B parameters) with the few-shot learning prompt described in Section 4.1 with $K = 5$ to automatically generate simplifications for each unlabeled sentence in $\mathbf{D}_{Unl}$. We used the API provided by OpenAI [3]. We generated $N = 5$ simplifications for each liver sentence with $temperature = 0.5$, which was selected to provide a good balance between factual correctness and diversity.

We trained BERT-Checker to select the best among the $N = 5$ generated simplifications for each liver sentence. BERT-checker was fine-tuned using 80% of the training data as positives and four copies of negatives for each positive, as explained in Section 4.2. BERT-Checker was a fine-tuned BERT base model (110M parameters) consisting of

[2] https://github.com/Ziyu-Yang/Radiology-Text-Simplification-Liver

[3] https://openai.com/api/

12 transformer encoder layers. A fully connected linear layer was added to BERT on its [CLS] output to score the simplification quality. The binary cross-entropy loss was used. 20% of the training data was used for validation and early stopping. We fine-tuned for up to 20 epochs with the patience for early stopping of 3, batch size 16, and learning rate 1e-4. All experiments were implemented with a single GTX 1080Ti.

The accuracy of trained BERT-Checker on validation data was 0.924. Its precision (the fraction of true positives among positive predictions) was 0.899 and its recall (the fraction of positives that were predicted correctly) was 0.958. We consider it to be high enough accuracy for BERT-Checker to be used to determine the quality of simplifications produced by GPT-3.

In the lexical augmentation, we annotated the recognized entities in the liver sentences from the labeled corpus with Type Unique Identifier (TUI) [4]. TUI is the code to represent hierarchical semantic types of all medical concepts in UMLS and MeSH. Specifically, we only paraphrased terms that belong to "*T023 | Body Part, Organ, or Organ Component*" or "*T033 | Finding*" groups. Because many medical concepts have only one synonym, many sentences mentioned only a single body part other than the liver, and a single finding, we finally obtained 242 unique lexical augmentations from $\mathbf{D}_{Unl}$. In order to control the effect of augmentation size, we randomly selected 200 of them for further experiments.

5.3 BART model

BART (Lewis et al., 2019) is a pre-trained model that uses a seq2seq architecture with a bidirectional encoder and a left-to-right decoder. It achieves state-of-the-art performance on many seq2seq benchmarks. We fine-tuned a BART base model (406M parameters) on different mixes of 450 labeled and augmented data to create different radiology simplification models. The fine-tuning was implemented using PyTorch-lightning[5]. 20% of the training data was used for validation and early stopping. We used the cross entropy loss. We used the same training setting as for BERT-Checker.

[4]https://lhncbc.nlm.nih.gov/semanticnetwork/index.html

[5]https://www.pytorchlightning.ai/

5.4 Baselines

We first introduce two model baselines that do not use augmentations. Then we introduce two baseline augmentation methods that are appropriate to our task.

5.4.1 Model Baselines

The first baseline is BART base model fine-tuned with the labeled data (**BART-base**). As the second baseline, we used simplifications by the same implementation of GPT-Curie model that is used to augment the labeled data. Specifically, we selected the most related $K = 5$ sentences from the labeled set to a test sentence as the few-shot prompt, generated $N = 5$ simplifications and used BERT-Checker to select the best one. We name this baseline **GPT-FS**.

5.4.2 Augmentation Baselines

We implemented and evaluated two widely used baseline data augmentation methods: 1) Easy Data Augmentation (**EDA**) (Wei and Zou, 2019),a rule-based augmentation that includes synonym replacement, random insertion, random swap, and random deletion. We reproduced this baseline with its source code [6]. 2) Back translation (**BT**), that uses a pre-trained machine translation model to translate sentences into another language and then translate them back to English. The back-translated English sentences are fused with the corresponding simplifications to provide augmented data. Following previous work (Brown et al., 2020), we used GPT-3 Curie to back translate the original sentences to French and back to English. French was selected because it provided a good balance between factual correctness and diversity of generated back-translations.

We generated 200 augmented examples for each baseline approach.

6 Evaluation Methods

6.1 Automated Evaluation

We used multiple automated metrics to evaluate text simplification accuracy. **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) is a set of metrics used for seq2seq tasks. It calculates the overlapping of unigrams, bigrams, and the longest common subsequences between the expert-provided and machine-generated simplifications. Similarly, **BLEU** (bilingual evalua-

[6]https://github.com/jasonwei20/eda_nlp

Table 2: Comparison of augmentation methods.

	# Aug	ROUGE 1/2/L	BLEU	SARI	BERTScore	FKGL↓
Baseline Models						
GPT-FS	0	56.00/42.20/54.64	0.2363	0.5455	0.9457	5.392
BART-base	0	59.81/50.34/58.87	0.4240	0.5324	0.9411	5.560
Augmentation Methods						
EDA	200	60.90/51.90/60.06	0.4461	0.5460	0.9429	5.315
BT	200	63.47/53.50/62.33	0.4504	0.5740	0.9470	5.133
HUMAN	200	**71.06/62.89/70.20**	**0.5322**	**0.6047**	**0.9566**	**4.870**
LEX	200	68.58/60.84/68.24	**0.5391**	0.5769	**0.9559**	5.353
SEM	200	66.11/56.55/64.76	0.4709	0.5875	0.9510	5.629
AUG-SUB	200	67.92/58.81/67.04	0.5020	0.5960	0.9524	5.314
AUG	400	**69.03/60.37/68.51**	0.5036	**0.6029**	0.9550	**5.021**

tion understudy) (Papineni et al., 2002) also evaluates overlap of n-grams between the simplifications. Unlike ROUGE and BLEU, **BERTScore** (Zhang et al., 2019) computes a contextual similarity score between tokens in the simplifications. **SARI** (Xu et al., 2016) is a gold standard edit-based metric for text simplification evaluation. Unlike other metrics, it compares the machine-generated simplification with respect to both the original sentence and the human-provided simplification. To evaluate simplicity, we used **FKGL** (Flesch Kincaid Grade Level) (Kincaid et al., 1975), which is a widely used readability formula that assesses the approximate reading grade level of a text. The lower score indicates simpler texts.

6.2 Human Evaluation

Applying automatic evaluation metrics is insufficient to compare quality of simplifications by different methods. Therefore, we also used human evaluation. We asked a medical doctor (family physician) that was distinct from the radiologist who provided the simplifications to evaluate the machine-generated simplifications. We asked the evaluator to use 1-5 Likert scale to evaluate the following four aspects of each simplification, the first three being consistent with.

Factuality refers to medical correctness of the simplification. Score one means that the simplification is factually incorrect and five that it is correct. Scores between one and five mean that some information is imprecise, missing, or hallucinated. Lower scores mean there are more serious factual errors. **Fluency** measures the quality of grammar and readability, regardless of factual correctness. If a simplification is both easy to read and grammatically correct it gets a score of five. This measure is consistent with the fluency measure explained in (Nisioi et al., 2017). **Simplicity** evaluates whether the evaluator thought the laypeople would be able to understand the simplification, regardless of factual correctness. Score of five means that the evaluator thought that any patient would be able to completely understand the simplification.

During the initial stages of human evaluation of factuality and simplicity, we observed that the evaluator occasionally preferred machine-generated simplifications to the radiologist-provided ones. That is why we introduced **Consistency**, which measures how closely the simplification matches the radiologist-provided simplification. Score of five means that the simplification is almost identical to the radiologist-provided simplification. We note that Consistency is related to SARI automatic measure (Xu et al., 2016).

7 Results

7.1 Quantitative Results

We fine-tuned BART model including augmented data from baseline methods (EDA, BT), and our lexical and semantic augmented data (LEX, SEM). BART-base and GPT-FS were created according the description in Section 5.4.1. First two rows of Table 2 refer to fine-tuned BART and few-shot prompted GPT-3 Curie using only the radiologist-provided labeled data. The remaining rows refer to inclusion of augmented data to BART tuning. Rows EDA and BT refer to the baseline augmentation methods. Row HUMAN refers to the augmentation provided by the radiologist, and serves to establish the upper bound on accuracy improve-

ment due to augmentation. LEX and SEM rows represent our lexical and semantic augmentation methods. AUG-SUB and AUG use 200 and 400 combined semantic and lexical augmentations, respectively. Aug column shows the number of augmented examples.

We observe that our proposed augmentations are superior to baselines LEX and SEM on almost all metrics. SEM is better than LEX on SARI measure. AUG is better than LEX and SEM on ROGUE, SARI and FKGL. AUG is the best overall augmentation method coming very close to the HUMAN upper bound, after noting that SARI and FKGL are the most useful measures for evaluation of simplicity. We note that GPT-FS has lower overall scores than any of the BART models.

Table 3: Human evaluation results (Factuality, Fluency, Simplicity, and Consistency) on 60 selected testing data.

Method	Factual	Fluency	Simp	Cons
BART-base	3.38	4.85	4.67	3.18
BT	3.22	4.88	4.58	3.13
GPT-FS	4.18	5.00	4.55	3.91
AUG	4.22	4.95	4.62	4.10

7.2 Human Evaluation Results

For Table 3, we asked a medical doctor to evaluate 60 randomly selected simplifications from the test data (30 from each source). We evaluated the most relevant four models from Table 2: BART-base, BT, GPT-FS,and AUG. The results show that all methods have comparable Simplicity and Fluency. AUG and GPT-FS have better Factuality and Consistency than BART-base and BT. AUG is slightly better than GPT-FS on those two important measures, indicating that fine-tuning BART with augmentation produced by few-shot prompted GPT-3 Curie is better than directly using few-shot prompted GPT-3 Curie for simplification.

Table 4: Comparison between different versions of semantic augmentations. **# Aug** is the number of augmented examples. ROUGE refers to ROUGE-L.

Method	# Aug	ROUGE	BLEU	SARI
First-run	200	60.21	0.4053	0.5590
Similarity	200	55.31	0.3547	0.5387
Five-runs	781	55.77	0.3755	0.5391
SEM	200	**64.76**	**0.4709**	**0.5875**

7.3 Ablation Study

We first evaluated the ability of BERT-Checker to recognize high-quality simplifications. We compared the version we implemented in our experiments (SEM row in Table 4) with three different variants: 'First-run' always selects the first generated simplification, 'Similarity' selects the best simplification based on BERTScore, 'Five-runs' uses all simplifications generated by GPT-3 Curie as augmentations. After removing duplicates, there are 781 augmentations produced by 'Five-runs'. Table 4 shows all three variants are inferior to SEM, showing that any of the ablations would significantly deteriorate the results. The results confirm that the quality of augmentations is critical for success of data augmentation approaches.

Next, we evaluated the importance of GPT-3 prompting. As noted in previous research (Liu et al., 2020), the choice of prompting can significantly impact the quality of the generated text. Thus, we designed an ablation study to compare different prompting approaches for data augmentation.

Table 5: Comparison of different prompting on data augmentation.

Prompts	ROUGE	BLEU	SARI
BART-grader	46.62	0.2862	0.4917
BART-patient	55.81	0.3516	0.5255
BART-top1	58.65	0.3955	0.5511
BART-rd5	53.94	0.3170	0.5360
SEM	**64.76**	**0.4709**	**0.5875**

In our prompt design that has the following form: ***Sentence:*** $< \mathbf{X} >$***; Simplification:*** $< \mathbf{Y} >$', we included $K = 5$ most related labeled examples to the original test sentence in the prompt. We first explored whether the number of few-shot examples matters. We repeated the data augmentation process with $K = 1$ (BART-top1 in the table). Table 5 shows that $K = 5$ resulted in better performance than $K = 1$. Next, we evaluated whether the way we select examples matters. Instead of $K = 5$ closest labeled examples, we selected $K = 5$ random labeled examples (BART-random in the table). From Table 5, we can see that random labeled examples resulted in lower accuracy.

We also explored prompting that does not rely on few-shot learning. One design was explained in section 4.1.1, '*My patient asks me to simplify this radiology sentence* <**X**>', we refer to as BART-

patient in the table. Similarly, inspired by a GPT-3 prompt for the summarization task, we used prompt: *My second grader student asks me to simplify the following sentence:* **<X>**, we refer to as BART-grader in the table. These two prompts are the so-called 'zero-shot' prompts. As shown in Table 5, the 'grader' and 'patient' prompts result in inferior accuracy compared to the few-shot prompting.

8 Conclusion

This paper proposes two novel augmentation methods to enhance the limited labeled data for the radiology sentence simplification problem. Our evaluation using automatic measures and human evaluation shows that data augmentation can substantially improve the quality of simplification models. The ablation results show that the proposed innovations in automatic creation of simplifications for data augmentation are very effective.

9 Limitations

The main limitation of our study is that we only considered simplification of radiology sentences. In future work, it will be important to expand the approach to simplify whole paragraphs, because very often radiologists use multiple sentences to discuss a single observation. Simplifying single sentences can thus be suboptimal because important context from the previous and subsequent sentences might be lost. The second limitation of the study is that our corpus only included sentences related to liver. It will be important in the future work to evaluate the proposed approach on a wider variety of radiology sentences. The third limitation is that we obtained simplifications from a single radiologist. It will be important for future study to include simplifications from multiple radiologists to ensure generalizability of the proposed approach. The fourth limitation is that we used a single medical doctor to evaluate the quality of the simplifications. It would be important in future studies to ask multiple medical doctors to evaluate the quality, which would allow estimating the inter-rater variability. The fifth limitation is that we did not use laypeople to evaluate the quality of simplification. This would require some innovation in the human evaluation process because laypeople are not able to evaluate factual correctness and because it would be important to understand how simplifications improve the overall understanding of the radiology reports. The final limitation is a relatively small size of the labeled data set created for this study. Obtaining high-quality simplifications is very costly because it requires collaboration between radiologists and laypeople.

References

Viraj Adduru, Sadid A Hasan, Joey Liu, Yuan Ling, Vivek V Datla, Ashequl Qadir, and Oladimeji Farri. 2018. Towards dataset creation and establishing baselines for sentence-level neural clinical paraphrase generation and simplification. In *KHD@ IJCAI*.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. 2020. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7383–7390.

Neeraj K Arora. 2013. Patient engagement in a rapidly changing communication environment: reflections of a cancer survivor. *Journal of the National Cancer Institute Monographs*, 2013(47):231–232.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2021. A survey on data augmentation for text classification. *ACM Computing Surveys*.

Or Biran, Samuel Brody, and Noémie Elhadad. 2011. Putting it simply: a context-aware approach to lexical simplification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 496–501.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jinying Chen, Emily Druhl, Balaji Polepalli Ramesh, Thomas K Houston, Cynthia A Brandt, Donna M Zulman, Varsha G Vimalananda, Samir Malkani, and Hong Yu. 2018. A natural language processing system that links medical terms in electronic health

record notes to lay definitions: system development using physician reviews. *Journal of medical Internet research*, 20(1):e26.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76.

Tessa S Cook, Seong Cheol Oh, and Charles E Kahn Jr. 2017. Patients' use and evaluation of an online system to annotate radiology reports with lay language definitions. *Academic radiology*, 24(9):1169–1174.

Kendall Cooper, Marta E Heilbrun, Shenise Gilyard, Brianna L Vey, and Nadja Kadom. 2020. Shared decision making: Radiology's role and opportunities. *American Journal of Roentgenology*, 214(1):W62–W66.

Tom Delbanco, Jan Walker, Sigall K Bell, Jonathan D Darer, Joann G Elmore, Nadine Farag, Henry J Feldman, Roanne Mejilla, Long Ngo, James D Ralston, et al. 2012. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. *Annals of internal medicine*, 157(7):461–470.

Ashwin Devaraj, Byron C Wallace, Iain J Marshall, and Junyi Jessy Li. 2021. Paragraph-level simplification of medical texts. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2021, page 4972. NIH Public Access.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. 2018. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*.

Saman Enayati, Ziyu Yang, Benjamin Lu, and Slobodan Vucetic. 2021. A visualization approach for rapid labeling of clinical notes for smoking status extraction. In *Proceedings of the Second Workshop on Data Science with Human in the Loop: Language Advances*, pages 24–30, Online. Association for Computational Linguistics.

Shlomit Goldberg-Stein and Victoria Chernyak. 2019. Adding value in radiology reporting. *Journal of the American College of Radiology*, 16(9):1292–1298.

Sadid A Hasan, Bo Liu, Joey Liu, Ashequl Qadir, Kathy Lee, Vivek Datla, Aaditya Prakash, and Oladimeji Farri. 2016. Neural clinical paraphrase generation with attention. In *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, pages 42–53.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Philippe Laban, Tobias Schnabel, Paul Bennett, and Marti A Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. *arXiv preprint arXiv:2107.03444*.

John P Lalor, Hao Wu, Li Chen, Kathleen M Mazor, and Hong Yu. 2018. Comprehenotes, an instrument to assess patient reading comprehension of electronic health record notes: development and validation. *Journal of medical Internet research*, 20(4):e9380.

Gondy Leroy, David Kauchak, and Alan Hogue. 2016. Effects on text simplification: Evaluation of splitting up noun phrases. *Journal of health communication*, 21(sup1):18–26.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Carolyn E Lipscomb. 2000. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265.

Pei Liu, Xuemin Wang, Chao Xiang, and Weiye Meng. 2020. A survey of text data augmentation. In *2020 International Conference on Computer Communication and Network Security (CCNS)*, pages 191–195. IEEE.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Ana P Lourenco and Grayson L Baird. 2020. Optimizing radiology reports for patients and referring physicians: mitigating the curse of knowledge. *Academic radiology*, 27(3):436–439.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Partha Mukherjee, Gondy Leroy, David Kauchak, Srinidhi Rajanarayanan, Damian Y Romero Diaz, Nicole P Yuan, T Gail Pritchard, and Sonia Colina. 2017. Negait: A new parser for medical text simplification using morphological, sentential and double negation. *Journal of biomedical informatics*, 69:55–62.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.

Seong Cheol Oh, Tessa S Cook, and Charles E Kahn. 2016. Porter: a prototype system for patient-oriented radiology reporting. *Journal of digital imaging*, 29(4):450–454.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Nikhil Pattisapu, Nishant Prabhu, Smriti Bhati, and Vasudeva Varma. 2020. Leveraging social media for medical text simplification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 851–860.

Andrew B Rosenkrantz and Eric R Flagg. 2015. Survey-based assessment of patients' understanding of their own imaging examinations. *Journal of the American College of Radiology*, 12(6):549–555.

Tarek Sakakini, Jong Yoon Lee, Aditya Duri, Renato FL Azevedo, Victor Sadauskas, Kuangxiao Gu, Suma Bhat, Dan Morrow, James Graumlich, Saqib Walayat, et al. 2020. Context-aware automatic text simplification of health materials in low-resource domains. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 115–126.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2018. Unsupervised neural text simplification. *arXiv preprint arXiv:1810.07931*.

Laurens Van den Bercken, Robert-Jan Sips, and Christoph Lofi. 2019. Evaluating neural text simplification in the medical domain. In *The World Wide Web Conference*, pages 3286–3292.

Yufei Wang, Can Xu, Qingfeng Sun, Huang Hu, Chongyang Tao, Xiubo Geng, and Daxin Jiang. 2022. Promda: Prompt-based data augmentation for low-resource nlu tasks. *arXiv preprint arXiv:2202.12499*.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6383–6389, Hong Kong, China. Association for Computational Linguistics.

Jun-Cheng Weng, Yu-Syuan Chou, Guo-Joe Huang, Yeu-Sheng Tyan, and Ming-Chou Ho. 2018. Mapping brain functional alterations in betel-quid chewers using resting-state fmri and network analysis. *Psychopharmacology*, 235(4):1257–1271.

Wei-Hung Weng, Yu-An Chung, and Peter Szolovits. 2019. Unsupervised clinical language translation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3121–3131.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyeong Park. 2021. Gpt3mix: Leveraging large-scale language models for text augmentation. *arXiv preprint arXiv:2104.08826*.

Qing T Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.