"Nothing Abnormal": Disambiguating Medical Reports via Contrastive Knowledge Infusion

Zexue He[1], An Yan[1], Amilcare Gentili[1,2], Julian McAuley[1], Chun-Nan Hsu[1,2,3]

[1] University of California, San Diego, La Jolla, CA
[2] VA San Diego Healthcare System, San Diego, CA
[3]VA National AI Institute, Washington, DC
zehe@eng.ucsd.edu, ayan@eng.ucsd.edu, amilcare.gentili2@va.gov,
jmcauley@eng.ucsd.edu, chunnan@ucsd.edu

Abstract

Sharing medical reports is essential for patient-centered care. A recent line of work has focused on automatically generating reports with NLP methods. However, different audiences have different purposes when writing/reading medical reports – for example, healthcare professionals care more about pathology, whereas patients are more concerned with the diagnosis ("*Is there any abnormality?*"). The expectation gap results in a common situation where patients find their medical reports to be ambiguous and therefore unsure about the next steps. In this work, we explore the *audience expectation gap* in healthcare and summarize common ambiguities that lead patients to be confused about their diagnosis into three categories: *medical jargon*, *contradictory findings*, and *misleading grammatical errors*. Based on our analysis, we define a disambiguation rewriting task to regenerate an input to be unambiguous while preserving information about the original content. We further propose a rewriting algorithm based on contrastive pretraining and perturbation-based rewriting. In addition, we create two datasets, OpenI-Annotated based on chest reports and VA-Annotated based on general medical reports, with available binary labels for ambiguity and abnormality presence annotated by radiology specialists. Experimental results on these datasets show that our proposed algorithm effectively rewrites input sentences in a less ambiguous way with high content fidelity. Our code and annotated data will be released to facilitate future research.

	Report Sentence	Diagnosis
Medical Jargon	Am: *Unremarkable bony structure.* Re: ***Normal** bony structure.*	Normal
Contradictory Findings	Am: *The lung volumes are low normal.* Re: *The lung volumes are in the **lower half of the normal limit**.*	Normal
Misleading Grammatical Errors	Am: *Cardiomegaly and hiatal hernia without an acute abnormality identified.* Re: *Cardiomegaly and hiatal hernia **.** **Without** an acute ab-normality identified.*	Abnormal

Table 1: Ambiguous sentences (Am) from three categories with the unambiguous rewritten (Re). We highlight the parts causing ambiguity in gray, and show comparisons in bold.

Introduction

Effective communication between healthcare providers and patients plays a critical role in patient outcome. *Patient-centered care* (Catalyst 2017; Stewart et al. 2013) is reforming traditional healthcare to shift a patient's role from an "order taker" to an active "team member" in their own healthcare process, to improve individual health outcomes and satisfaction (Stewart et al. 2000), and advocates sharing medical information fully and in a timely manner with patients. It is also required by legal obligation (e.g., HIPAA[1] in the US) that patients have a legal right to access their personal health information. Failure of healthcare providers to communicate with patients efficiently and effectively about the results of medical examinations may lead to delays in proper treatment or malpractice lawsuits against providers (Mityul et al. 2018; Srinivasa Babu and Brooks 2015).

As a carrier of medical information, medical reports are shared with their patients by healthcare providers nowadays. Medical reports serve many communication purposes with different audiences including ordering physicians, other care team staff members, patients and their families, and researchers (Hartung et al. 2020; Gunn et al. 2013). Each group has different needs and expectations when reading the reports: peer medical professionals pay more attention to actionable findings, while patients usually care more about the diagnostic outcome[2] (i.e., *Is there anything abnormal?*). How to address various communication needs for different audiences, and to bridge the *expectation gap between audiences* without increasing workload of report writers is critical.

To build such a bridge, it is important for medical reports to 1) be understandable with little specialized terms and 2) to have no ambiguity about the significance of findings when communicating with patients (Hartung et al. 2020; Mityul et al. 2018). Previous works mainly focus on the first point where they change terminology to lay-person terms with replacement-based or deep learning methods (Qenam et al. 2017; Oh, Cook, and Kahn 2016; Xu et al. 2022). However,

[1]Shorten for Health Insurance Portability and Accountability Act

[2]We use exam result, diagnostic decision, abnormality existence interchangeably, to express "if there is anything abnormal".

how to mitigate the ambiguity in a comprehensible report is crucial but rarely investigated.

In our work, we consider medical reports written in free text and analyze the ambiguity where patients are unsure about their exam results. We first collect medical report data and ask domain specialists to label the binary abnormality presence associated with each sentence, and non-experts to label sentences that they deem ambiguous. Our medical team analyze the results and categorize the major causes behind ambiguity primarily into three categories: the report sentence is ambiguous due to containing (1) **medical jargon** with meanings different from everyday general usage, such as *unremarkable*; (2) **contradictory findings** in the same sentence; (3) **misleading grammatical errors** such as no period between full sentences. Examples are shown in Table 1.

To alleviate patient confusion, we propose a new task called *medical report disambiguation*, which is defined as: given an ambiguous sentence from a medical report that patients find hard to understand the exam result ("*Is there any abnormality?*"), rewrite it with minimal edits in a way that the diagnostic decision is expressed to be more explicit, while retaining the precision of the original findings, namely preserving the detected pathologies and preventing new interpretations from being implied.

Paraphrasing models may offer a solution, but they are limited by the need of parallel corpus, which requires significant workload from radiologists. To alleviate the annotation burden, we propose a rewriting framework without parallel corpora for disambiguation (see Figure 1). We first pretrain a Seq2Seq model in the medical domain with contrastive learning. Then, an ambiguous input is rewritten using the model by perturbing its hidden states and pushing the generation towards a direction that is more explicit about its exam results. The pretraining step not only enables a model to capture the underlying language distribution for writing a human-readable medical report, but also enforces a property that sentences sharing similar pathology patterns will reside closely in the latent embedding space. The two steps work together to preserve content fidelity with original input.

In summary, our work makes the following contributions: (1) We explore a novel and important problem in the healthcare domain regarding ambiguous medical reports. We empirically analyze the common reasons for ambiguous reports that make patients confused, and formally define the disambiguation rewriting task. (2) Based on our analysis, we propose an effective rewriting framework. Our model does not require parallel ambiguous and rewritten "golden" sentences for training, which alleviates the workload of medical specialists. (3) In addition, we provide two new datasets, OpenI-Annotated in chest radiology imaging and VA-Annotate data in general medical domain[3] , each annotated with high-quality labels for ambiguity and abnormality presence from radiology specialists. (4) Using these datasets, we perform experiments and evaluate rewritten results based on disambiguation and fidelity preservation. The results of both automatic and human evaluations indicate the effectiveness of our proposed method.

To the best of our knowledge, our work is the first attempt to build an AI system to deal with patient confusion caused by ambiguous reports, which can potentially help promote patient-centered healthcare.

Disambiguating Rewriting Framework

Our task is to disambiguate an input medical sentence when a patient finds it hard to understand the diagnostic decision. For an ambiguous sentence x whose abnormality label is y (abnormality presents or not), we will output a disambiguated sentence $\tilde{x}$ that is more explicit about y.

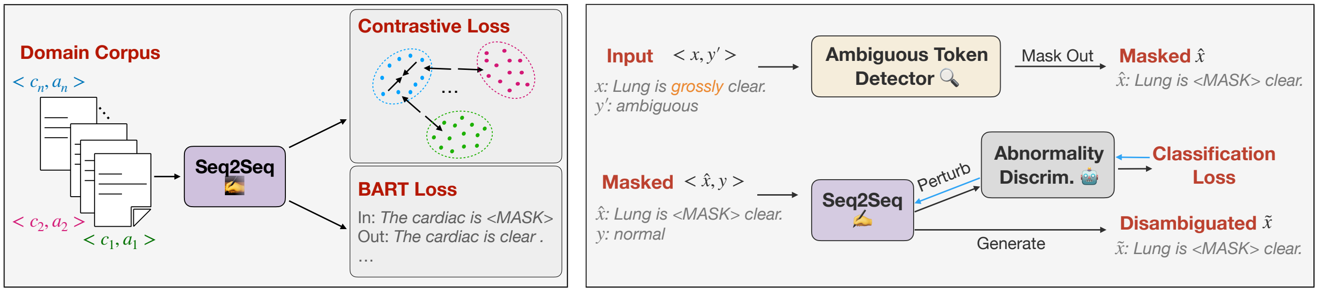
We propose a contrastive knowledge infused rewriting framework to achieve this goal, which comprises a pretraining step and a rewriting step, as shown in Figure 1 (a) and (b). We first obtain a medical-domain Seq2Seq model $\mathcal{G}$ that can effectively capture language patterns in different health situations in the pretraining step, and we generate a less ambiguous sentence using $\mathcal{G}$ in the rewriting step. We introduce each step in following sections.

Contrastive Pretraining

First, we pretrain a domain-specific Seq2Seq model to generate medical language on top of a general domain BART (Lewis et al. 2020). For our task, a pretrained model that only captures the distribution of medical language is not precise enough – there are several ways to rewrite an abnormal diagnostic in order to make it "more abnormal". Consider a patient with a diagnosis of having excessive lung fluid. This abnormal diagnosis can be rewritten to be "more abnormal" by combining it with other abnormalities (such as *unusual liquid and unusual air*) or by changing the disease (from a lung disease to a heart disease). This is undesirable. Therefore, we require the rewritting to preserve the original diagnosis. To achieve this goal, more domain knowledge about different pathologies is required. We capture such domain knowledge by learning from external corpora (such as MIMIC-CXR(Johnson et al. 2019)) that are on a large scale in the same medical domain with fine-grained disease labels. We pretrain the language model by infusing the domain knowledge with supervised contrastive learning, which pushes sentences closer if they express similar pathological findings and pull away sentences if they are not. As a result, we can reduce the probability of rewriting a sentence medically different from the original input. This is crucial for patient safety and a unique issue in our task different from other text rewriting problems.

The external medical corpora consist of medical report pairs including both sentence c_i and its associated fine-grained pathological label a_i (such as disease labels like *atelectasis, edema, no finding, etc*). We pretrain an encoder-decoder transformer $\mathcal{G}$ with supervised contrastive learning, following Khosla et al. (2020). For each sentence c_i in a mini-batch B, we first obtain its representation H_{c_i} by taking the last hidden states from the decoder in $\mathcal{G}$. Then a τ-temperature similarity $s_{i,j}$ between c_i and another sentence

[3]Our annotated data and source codes will be released at https://github.com/ZexueHe/Med-DEPEN



(a) Contrastive Pretraining

(b) Rewriting Framework

Figure 1: Model illustration. Our model contains two steps: first do (a) constrastive pretraining and then (b) rewriting.

c_j in B is calculated:

$$s_{i,j} = \text{sim}(H_{c_i}, H_{c_j}) = H_{c_i} \cdot H_{c_j} / \tau \tag{1}$$

We use $S(i) = \{c_j : a_j = a_i\}$ to denote the set of sentences sharing the same disease label a_i, then the contrastive learning loss $\mathcal{L}_{\text{CL}}$ for the mini-batch B is defined as

$$\mathcal{L}_{\text{CL}} = \sum_{c_i \in B} \mathcal{L}_{\text{CL},i} \tag{2}$$

where

$$\mathcal{L}_{\text{CL},i} = -\frac{1}{|S(i)|} \log \frac{\sum_{c_j \in S(i)} \exp(s_{i,j})}{\sum_{c_j \in B} \exp(s_{i,j})} \tag{3}$$

Our constrative pretraining is also applicable even when the pathological label a_i is not available. A recent empirical study (Oakden-Rayner et al. 2020) shows that the representations from deep neural networks carry information of labels and unlabeled features. Inspired by this, in the case where a_i is not available, we first extract sentence representations from a medical Bert pretrained with a radiology report corpus (Yan et al. 2022b). Then we follow Sohoni et al. (2020) to cluster sentences with a Gaussian Mixture Model. The clustered results carry fine-grained information about different pathological patterns, and work as an approximation of the labels used in optimizing Eq. (3).

Besides the contrastive learning objective, our pretraining also includes a token infilling task (Lewis et al. 2020) in order to obtain an informative representation H_{c_i}, which reconstructs the original sentence c_i from its randomly masked version $\hat{c}_i$:

$$\mathcal{L}_{\text{BART}} = -\sum_{i}^{|B|} \sum_{t}^{|c_i|} \log p(c_i^t | c_i^1, \cdots, c_i^{t-1}; \hat{c}_i) \tag{4}$$

Therefore, our pretraining goal is to learn the medical language distribution (language modeling loss $\mathcal{L}_{\text{BART}}$) and capture language patterns of different medical conditions (contrastive loss $\mathcal{L}_{\text{CL}}$), and is formulated by minimizing their weighted sum in Eq. (5):

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{BART}} + \lambda_2 \mathcal{L}_{\text{CL}} \tag{5}$$

Rewriting Framework

During the rewriting process for an ambiguous input x_i, the following objectives are targeted: 1) the main content is retained, and 2) the diagnostic decision is more explicitly expressed in the rewritten sentence. While contrastive pretraining ensures a reasonable level of content fidelity, the first objective also suggests minimal changes during rewriting, which only touch those portions necessary for disambiguation. The second one requires a controllable generation that pushes the generation closer to the diagnostic decision.

Inspired by recent advance in controlled text generation (He, Majumder, and McAuley 2021), we leverage a plug-and-play method to rewrite the sentences without the need of parallel annotated training data. It includes a *detect* stage to mask potential tokens that are highly predictable for an attribute and a *perturb* stage to do neutralization rewriting w.r.t. that attribute. Since we need to detect tokens in x_i that are highly predictable in their ambiguity, we first train an ambiguity classifier during detect stage. The tokens with the top-K highest attention scores will be detected as salient for ambiguity and will be masked. Then in the perturb stage, we require an edit that is more explicit in the direction of its diagnostic decision y_i, rather than making it more neutral for the ambiguity. Therefore, we modify the perturb stage to suggest an explicit edit by maximizing the likelihood of making the right diagnostic decision at each generation step t:

$$\tilde{x}_i^t = \arg\max_{\tilde{x}_i^t} p(y|\tilde{x}_i^t), \tag{6}$$

where the distribution p is output from a classifier f which predicts the diagnostic decision y_i, pretrained by minimizing Cross-Entropy$(f(x_i), y_i)$.

Then, during generation, we add a perturbation to decoder hidden states in $\mathcal{G}$ by taking the gradient w.r.t. $\tilde{x}_i^t$ from the Cross-Entropy loss, and regenerate the token distribution since the hidden states have been updated. Alternatively, adding perturbation and (re-)generation push the rewritten sentence towards the direction of its diagnostic decision, that is to say, being less ambiguous.

Experimental Setup

Our rewriting algorithm is tested in two practical settings. First, disambiguating chest reports in a specialized medi-

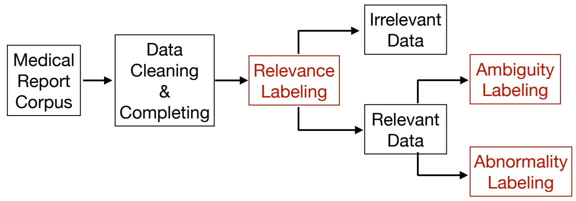


Figure 2: Data Annotation Pipeline. Three different labels are annotated in red steps.

Dataset	Total	Ambiguous	Abnormal
OpenI Annotated	15,023	988	6,111
VA Annotated	5,180	1,461	2,358

Table 2: Statistics of Annotated Datasets

cal domain. Secondly, disambiguating general medical reports that cover many imaging modalities (e.g., x-ray, CT, etc.) and body parts. Our medical team created annotation datasets (OpenI-Annotated and VA-Annotated) for each experiment. During pretraining, an additional large-scale medical corpus is used in each experiment.

Human-Annotated Datasets for Rewriting

The overall pipeline for building our annotated dataset is shown in Figure 2. We elaborate each dataset as follows.

OpenI-Annotated We take the sentence-level subset of OpenI released by Harzig et al. (2019). Our medical team conducts data cleaning by removing identical sentences and completing missing terms (mistakenly masked in de-identification) according to their domain knowledge. We distinguish sentences that are irrelevant to our task and re-label sentences that contain abnormal findings and that are ambiguous according to corresponding criteria. In the end, the OpenI-Annotated dataset consists of sentences with associated binary labels for being irrelevant, ambiguous, or abnormal. Statistics are shown in Table 2. In the experiment, we split the OpenI-Annotated data to train/validation/test sets by 70%, 10%, 20%.

VA Annotated We create the VA-Annotated dataset and use it in general-domain medical report rewriting. We use the VA radiology report corpus, recently introduced in (Yan et al. 2022b). As a general medical report corpus, it covers 8 modalities and 35 body parts for 70 modality-bodypart combinations. We sample a subset and split them into sentences.

Disease	Num.	Disease	Num.
Enlarged Cardiomediastinum	17,944	Atelectasis	32,445
Cardiomegaly	56,099	Pneumothorax	5,539
Lung Opacity	62,865	Pleural Effusion	36,537
Lung Lesion	9,838	Pleural Other	3,350
Edema	14,605	Fracture	10,893
Consolidation	3,905	Support Devices	8,355
Pneumonia	1,365	No Finding	865,738

Table 3: Fine-grained diseases in MIMIC-CXR.

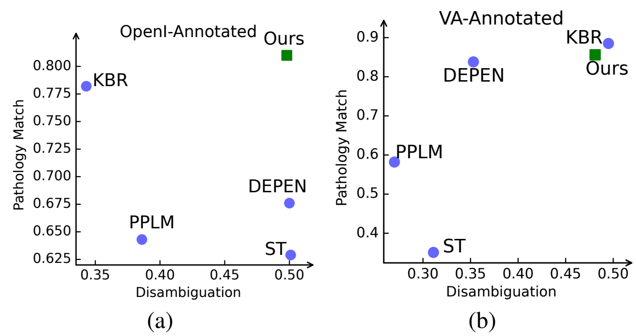


Figure 3: Trade-off between Disambiguation and Fidelity on (a) OpenI-Annotated (b) VA-Annotated. Higher disambiguation and pathology match (more upper-right corner in geometric) indicates a better rewriting.

Then similar data cleaning steps for OpenI-Annotated are used. Each sentence is annotated with binary labels for relevance, ambiguity and abnormality. We call it VA-Annotated and its statistics are listed in Table 2. In the experiment, we split it into train/validation/test sets by 70%, 10%, 20%.

Human Labeling Procedures In each experiment, experts start the labeling procedures after data cleaning.

First, for relevance labeling, the sentences only containing facts (e.g., *CT of the chest are taken*) and body parts (*left knee was not evaluated*) are regarded as irrelevant to our task since there are no abnormal/normal diagnoses mentioned. Our medical team provide their binary labels for relevance. Secondly, for relevant sentences, the medical team annotates a binary *abnormality label*, indicating if there is an abnormal symptom found in the sentence. Sentences containing abnormal symptoms usually imply a diagnostic decision of being sick. Our non-expert team annotates an *ambiguity label*, indicating whether the diagnostic decision looks too ambiguous for patients to understand.

The annotations are performed iteratively until inter-annotator Cohen's Kappa higher than substantial agreement (≥ 0.8). The final discrepant labels were resolved by doctors in our medical team.

Contrastive Pretraining Datasets

Here we introduced the corpus used in contrastive pretraining and elaborate their fine-grained pathological labels about different health conditions.

MIMIC-CXR MIMIC-CXR is the largest public-domain chest x-ray dataset proposed in (Johnson et al. 2019) with 220k reports. We obtain the report sentences after de-duplication. For fine-grained pathological labels, we use CheXbert (Irvin et al. 2019), an automated deep-learning based chest radiology report labeler trained with MIMIC-CXR data (therefore no domain shift occurs), to label 14 fine-grained diseases. We keep sentences that have at most one disease noted. We end up with 1,129,478 sentences. The 14 diseases and statistics are listed in Table 3. This dataset is used in pretraining of the chest rewriting experiment.

		Ambiguity Acc.	Decision Acc.	Pathology Match	PLL
OpenI-Annotated	**Raw Text**	0.855	0.950	1.000	-6.062
		Disambigutation	**Content Fidelity**		**Language Fluency**
		$\Delta\text{Acc}_{\text{Am}} \uparrow$	$\Delta\text{Acc}_{\text{Dis}} \downarrow$	Pathology Match $\uparrow$	PLL $\uparrow$
	KBR	0.343	0.001	0.782	-6.862
	ST	0.501	0.051	0.629	-6.454
	PPLM	0.386	0.115	0.643	-6.890
	DEPEN	0.500	0.052	0.676	-6.529
	Ours	0.496	0.032	0.809	-6.232
		Ambiguity Acc.	**Decision Acc.**	**Pathology Match**	**PLL**
VA-Annotated	**Raw Text**	0.955	0.946	1.000	-5.652
		Disambigutation	**Content Fidelity**		**Language Fluency**
		$\Delta\text{Acc}_{\text{Am}} \uparrow$	$\Delta\text{Acc}_{\text{Dis}} \downarrow$	Pathology Match $\uparrow$	PLL $\uparrow$
	KBR	0.495	0.007	0.885	6.109
	ST	0.311	0.235	0.351	-7.284
	PPLM	0.270	0.146	0.582	-6.147
	DEPEN	0.353	0.047	0.838	-6.102
	Ours	0.481	0.009	0.856	-5.821

Table 4: Automatic Evaluation Results on OpenI- and VA-Annotated. Statistics about the original data is provided separately.

	OpenI-Annotated		**VA-Annotated**	
Models	Disam $\uparrow$	Fidelity $\uparrow$	Disam $\uparrow$	Fidelity $\uparrow$
KBR	0.317	0.908	**0.488**	**0.988**
ST	0.609	0.526	0.225	0.350
PPLM	0.376	0.624	0.333	0.575
DEPEN	0.571	0.795	0.282	0.941
Ours	**0.792**	**0.921**	0.383	0.808

Table 5: Human Evaluations. *Disam*: Disambiguation.

VA-Rest The remaining unannotated sentences of the VA corpus (Yan et al. 2022b) are used as a contrastive pretraining corpus. VA contains general medical reports covering different body parts, therefore, CheXbert is not applicable. Instead, we use clustering results as pseudo-labels for different fine-grained pathological patterns. We first obtain the sentence representations by feeding them into a RadBERT model (Yan et al. 2022b), which is finetuned with the VA corpus by language modeling, and extracting the last hidden states. Then we reduce the dimension to D with Uniform Manifold Approximation and Projection (McInnes et al. 2018). Based on the reduced embeddings, sentences are clustered with a Gaussian Mixture Model into K clusters. After experimenting with different parameters, we notice $K = 14$ and $D = 256$ achieves a good Silhouette score.

Baselines and Ablations

We follow the experiment design of Xu et al. (2022), and choose baseline models that are commonly used and have publicly available code:

- **Knowledge-Based Replacement (KBR)** regenerates a sentence by replacing ambiguous terms with unambiguous alternatives. Following the previous work (Qenam et al. 2017), we build a dictionary for replacement by looking up the Consumer Health Vocabulary.[4] We notice that difficult special terms are also replaced with their layman language.
- **Style Transformer (ST)** A strong style-transfer model (Dai et al. 2019) with adversarial training and a transformer architecture to transfer style while preserving content by reconstruction.
- **Controllable Generation** We include two perturbation-based controllable generation models – *PPLM* (Dathathri et al. 2019) and *DEPEN* (He, Majumder, and McAuley 2021). PPLM is a decoder-based language model but not capable of regeneration. In order to use it in our task, we modify it into a Seq2Seq model. We also adapt DEPEN so that it generates a less ambiguous sentence, as it is originally proposed for bias neutralization rewriting.

After adapting PPLM and DEPEN, they can be regarded as ablations – PPLM can be considered as our algorithm without contrastive pretraining and 'detect' steps, while DEPEN is ours without contrastive pretraining.

Evaluation Metrics

Following the evaluation of Xu et al. (2022), we compare rewritten results from the following aspects.

- **Disambiguation:** We measure the level of ambiguity using the accuracy of a Bert classifier, which is finetuned to predict ambiguity labels in OpenI-Annotated or VA-Annotated. The accuracy deduction $\Delta\text{Acc}_{\text{Am}}$ is regraded as disambiguation performance.

[4]Availalbe as a part of the UMLS https://www.nlm.nih.gov/research/umls/index.html

Contradictory Findings	
Original Input	*normal* cardiac contour with *atherosclerotic changes* throughout the aorta.
KBR	normal heart contour with atherosclerotic changes throughout the aorta.
ST	normal cardiac contour with atherosclerotic changes throughout the aorta.
PPLM	unchknown tortuous cardiac contour unchanged tortuous atherosclerotic changes throughout the aorta.
DEPEN	diaphragmclerotic changes throughout the thoracic aorta.
Ours	The cardiac contour shows atherosclerotic changes throughout the aorta.
Medical Jargon	
Original Input	maybe *secondary* to prominent mediastinal fat or tortuous.
KBR	maybe secondary to prominent mediastinum palmitic acid or tortuous.
ST	secondary to prior mediastinal or tortuous.
PPLM	optional secondary to the calcifiedsecondsmediastinal fat or tortuous. Include.
DEPEN	ouching compared the to the mediastinal fat or tortuous.
Ours	maybe due to the mediastinal fat or tortuous.

Table 6: Examples of rewriting by different models for ambiguous sentences from OpenI-Annotated.

- **Fidelity:** We evaluate fidelity at two granularities: (1) a coarse-grained one which evaluates the persistence of the original abnormality label, measured by the accuracy gap $\Delta\text{Acc}_{\text{Dis}}$ from a Bert classifier finetuned to predict abnormality. (2) a fine-grained one which evaluates the match of pathology, measured by the match rate of CheXbert labeled results or pseudo-labels.
- **Language Quality:** Following He, Majumder, and McAuley (2021), we use Pseudo-Log-Likelihood (PLL) (Salazar et al. 2020) score to measure language fluency.

Human Evaluation

Rewritten results generated with different models are reviewed by radiology experts. For an ambiguous sentence, the rewritten result and its associated abnormality labels are shown to reviewers simultaneously. Reviewers decide (1) if the rewriting is successful in disambiguation; and (2) if the original content has been preserved by rewriting. As for the second one, our reviewers have a rigorous objective that includes language quality evaluation – a rewrite will be considered as a failure if there are any significant changes from the original findings or if proper English is not used. We collect the results of human evaluations and calculate the disambiguation and fidelity success rates.

Results and Analysis

Performance Comparison

The automatic evaluation results are shown in Table 4. Notably, it is sub-optimal to achieve the lowest ambiguity while generating a destroyed sentence. Therefore, we believe a good model is the one with an optimal balance between disambiguation rewriting and content preservation. We illustrate the trade-off between disambiguation and fidelity in Figure 3, where the upper-right corner indicates a good model. Our rewriting model resides at the upper-right corner in the two experiments, indicating a superior balance between disambiguation rewriting and content fidelity. This also agrees with human evaluation results shown in Table 5.

We discuss more about the results in the following. First, we compare our model with ST. Though it has a reasonable disambiguation (0.501 on OpenI-Annotated and 0.311 on VA-Annotated), ST has bad fidelity scores in both coarse-grained and fine-grained evaluations (the worst one on VA-Annotated). The generation quality is also worse compared with other models. We notice a rewriting from ST usually changes the original sentences significantly on both OpenI-Annotated and VA-Annotated, which explains why ST is able to disambiguate while fails in preserve fidelity. We provide our conjecture about the underlying reason: as an end-to-end model trained with multiple objectives at the same time, ST is more fragile when balancing objectives, making it difficult to find the sweet point between rewriting for disambiguation and content preservation.

Then, we compare our model with controllable generation baselines – PPLM and DEPEN. PPLM is a variation of our model without the detect step and constrastive pretraining step. Without the detect step, unnecessary edits can be applied, as the model knows little about which parts are ambiguous. And without contrastive pretraining to inject distinguishable domain knowledge, the model will fail to preserve the main pathological content, having bad content fidelity in the end. Therefore, PPLM is not effective at both disambiguation and fidelity on both OpenI-Annotated and VA-Annotated. DEPEN shows improvements on disambiguation and maintains the original abnormality compared with PPLM, as the detect stage is added. But it fails to preserve fine-grained pathology match due to the lack of contrastive pretraining. Our model has the best overall performance in disambiguation and fidelity at different granularities. The improvement between PPLM, DEPEN, and ours indicates the effectiveness of each component in our model.

We notice a clear difference in performance of KBR – it fails to disambiguate in OpenI-Annotated while it becomes a strong baseline in VA-Annotated by achieving the best in both disambiguation and fidelity. We conjecture the reasons to be domain difference. We discuss the divergence below.

Specific Domain vs. General Domain

As one can observe in Table 4, our neural rewriting model is able to substantially outperform other baselines on OpenI-Annotated (specific domain). This indicates that given a reasonable amount of training data, our framework can perform well for a particular domain. On the VA dataset (general domain), KBR becomes a strong baseline. We notice that when creating the dictionary, human medical experts are good at proposing jargons across broadly different diseases and organs in general healthcare domains. However proposing terms that are specific to a domain requires deeper knowledge in that particular discipline. Therefore, VA-Annotated is more well-covered by the dictionary than OpenI-Annotated which is specific to the chest domain. We found the dictionary coverage rate is 17.3% on OpenI-Annotated while 21.5% on VA-Annotated, which explains why replacing works better in VA-Annotated.

However, since knowledge-based models come with a price of dictionary compiling and human (especially expert) effort, it may be difficult to extend them to solve domain-specific problems as each domain requires significant workload and expert experience. Instead, our rewriting framework is potentially a more promising direction to explore for this task, as it alleviates human effort while achieving competitive or even better performance.

Case Study

We show some examples in Table 6. Findings in the first example are labeled as abnormal. The contradictory usage of "*normal*" and "*atherosclerotic changes*" in the sentence makes patients confused about the abnormality. As shown in this example, KBR replaces the special term with layman language (*cardiac* → *heart*), but this does not help disambiguation since there is still a contradiction. These limitations suggest that replacement-based models cannot handle patterns outside the dictionary or patterns at the sentence level. ST fails to rewrite a sentence. PPLM suffers from repetition issues and generates output that is not comprehensible. DEPEN can target the editing area with its detect step. However it fails to maintain fidelity without contrastive pretraining, and involves new findings that are inaccurate and change the original content drastically. However, our model achieves successful disambiguation by rewriting a contradiction-free sentence with minimal editing (*normal* → *The*) while maintaining fidelity by preserving the original abnormality (*atherosclerotic changes*).

Ambiguity in the second example is caused by medical jargon "*secondary to*", which implies "*mediastinal or tortuous* " is the reason for an abnormal finding. However, in regular usage, it means "*less important*" which is not the case or "*coming after*" which diminishes the causation. While other baselines either fail to disambiguate or introduce new content, ours is able to find a rewriting that mostly matches the context to describe the pathology causation.

Related Work

AI for Medical Reporting Recent advances in AI have enabled novel applications involving medical reports. Medical report generation (Li et al. 2018; Chen et al. 2020; Yan et al. 2021) aims to automatically generate descriptions for clinical radiographs, which may alleviate the development workload of radiologists. Some recent works notice the communication gaps between medical professionals and patients, and focus on changing terminology in medical reports to lay-person terms, using replacement-based methods based on dictionary or lexical rules (Qenam et al. 2017; Oh, Cook, and Kahn 2016), or leveraging style transfer techniques (Xu et al. 2022). However, in this study, we highlight that the confusion also comes from ambiguity in readable texts where a knowledge base search may be insufficient.

Controlled Text Generation The task of ccontrolling the output of a text generation model has been investigated recently. One line of research mainly focuses on the training process, by finetuning models with desired attributes (Gururangan et al. 2020) or creating a class-conditioned model (Keskar et al. 2019). Another direction is to design decoding approaches which are lightweight and effective. For example, Yang and Klein (2021) train classifiers to predict whether an attribute can be satisfied. PPLM (Dathathri et al. 2019) controlls generation by updating a pretrained model's hidden states, and DEPEN (He, Majumder, and McAuley 2021) adapted PPLM into the seq2seq model to rewrite a neutral output. There is less consideration of content preservation and domain knowledge in previous works, limiting their potential in sensitive or rigorous disciplines like healthcare. In contrast, our approach incorporates a contrastive pretraining step that effectively infuses domain knowledge and enhances content fidelity.

Contrastive Learning Contrastive learning (Gutmann and Hyvärinen 2010; Oord, Li, and Vinyals 2018) learns representations by pulling positives together and negatives far from each other. Recent work has explored contrastive learning for NLP applications, such as contrastive data augmentation (Shen et al. 2020; Qu et al. 2020), sentence embedding learning (Kim, Yoo, and Lee 2021) and text generation generation (Yan et al. 2022a; Zhu et al. 2022). Our work differs in that we apply a label-aware contrastive learning method to the pretraining stage of a language decoder, and mainly focus on its application to the healthcare domain.

Conclusion

Sharing medical information, especially reports with patients is essential to patient-centered care. Due to the communication gap between audiences, there is always ambiguity in reports, leading patients to be confused about their exam results. We collect and annotate two datasets containing radiology reports from healthcare systems in this study. We analyze and summarize three major causes of ambiguous reports: jargon, contradictions, and misleading grammatical errors, and propose a framework for disambiguation rewriting. Experimental results show that our model can achieve effective disambiguation while maintaining content fidelity.

Medical reporting is time-consuming and labor-intensive. Our work aims to inspire more research so that in the future, more AI systems like ours can be created and used to assist real-world medical services under expert auditing.

Ethics Statement

The use of medical report data from VA received ethical approval from the IRB approval ID #H200086. All patient information was de-identified and patient consent was waived. Despite the intention to mitigate ambiguity, an AI system may have unexpected outputs. Also rewriting a sentence may subtly alter its detailed content. A malicious user could adversarially uses the unexpected results. Hence, we emphasize that a system like ours should best be used with expert auditing to ensure safety. Reporting ambiguity is a practical task. Our work aims to establish a benchmark for this new problem and present a possible solution using NLP techniques to alleviate human effort. As an initial attempt, we hope to inspire more future research on this challenging problem.

Acknowledgments

This work was financially supported in part by the Office of the Assistant Secretary of Defense for Health Affairs through the Accelerating Innovation in Military Medicine (AIMM) Research Award program W81XWH-20-1-0693 and in part by the National Science Foundation Award #1750063. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the funding agencies.

References

Catalyst, N. 2017. What is patient-centered care? *NEJM Catalyst*, 3(1).

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449.

Dai, N.; Liang, J.; Qiu, X.; and Huang, X.-J. 2019. Style Transformer: Unpaired Text Style Transfer without Disentangled Latent Representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5997–6007.

Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; and Liu, R. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In *International Conference on Learning Representations*.

Gunn, A. J.; Sahani, D. V.; Bennett, S. E.; and Choy, G. 2013. Recent measures to improve radiology reporting: perspectives from primary care physicians. *Journal of the American College of Radiology*, 10(2): 122–127.

Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8342–8360.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Hartung, M. P.; Bickle, I. C.; Gaillard, F.; and Kanne, J. P. 2020. How to Create a Great Radiology Report. *RadioGraphics*, 40(6): 1658–1670. PMID: 33001790.

Harzig, P.; Chen, Y.-Y.; Chen, F.; and Lienhart, R. 2019. Addressing data bias problems for chest x-ray image report generation. *arXiv preprint arXiv:1908.02123*.

He, Z.; Majumder, B. P.; and McAuley, J. 2021. Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4173–4181.

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 590–597.

Johnson, A. E.; Pollard, T. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Peng, Y.; Lu, Z.; Mark, R. G.; Berkowitz, S. J.; and Horng, S. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

Keskar, N. S.; McCann, B.; Varshney, L. R.; Xiong, C.; and Socher, R. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 18661–18673.

Kim, T.; Yoo, K. M.; and Lee, S.-g. 2021. Self-Guided Contrastive Learning for BERT Sentence Representations. In *ACL/IJCNLP (1)*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880.

Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. *Advances in neural information processing systems*, 31.

McInnes, L.; Healy, J.; Saul, N.; and Großberger, L. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, 3(29): 861.

Mityul, M. I.; Gilcrease-Garcia, B.; Mangano, M. D.; Demertzis, J. L.; and Gunn, A. J. 2018. Radiology reporting: current practices and an introduction to patient-centered opportunities for improvement. *American Journal of Roentgenology*, 210(2): 376–385.

Oakden-Rayner, L.; Dunnmon, J.; Carneiro, G.; and Ré, C. 2020. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, 151–159.

Oh, S. C.; Cook, T. S.; and Kahn, C. E. 2016. PORTER: a prototype system for patient-oriented radiology reporting. *Journal of digital imaging*, 29(4): 450–454.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Qenam, B.; Kim, T. Y.; Carroll, M. J.; Hogarth, M.; et al. 2017. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12): e8536.

Qu, Y.; Shen, D.; Shen, Y.; Sajeev, S.; Chen, W.; and Han, J. 2020. CoDA: Contrast-enhanced and Diversity-promoting Data Augmentation for Natural Language Understanding. In *International Conference on Learning Representations*.

Salazar, J.; Liang, D.; Nguyen, T. Q.; and Kirchhoff, K. 2020. Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2699–2712.

Shen, D.; Zheng, M.; Shen, Y.; Qu, Y.; and Chen, W. 2020. A simple but tough-to-beat data augmentation approach for natural language understanding and generation. *arXiv preprint arXiv:2009.13818*.

Sohoni, N.; Dunnmon, J.; Angus, G.; Gu, A.; and Ré, C. 2020. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33: 19339–19352.

Srinivasa Babu, A.; and Brooks, M. L. 2015. The Malpractice Liability of Radiology Reports: Minimizing the Risk. *RadioGraphics*, 35(2): 547–554. PMID: 25763738.

Stewart, M.; Brown, J.; Donner, A.; McWhinney, I.; Oates, J.; Weston, W.; and Jordan, J. 2000. The impact of patient-centered care on outcomes. *The Journal of family practice*, 49(9): 796–804.

Stewart, M.; Brown, J. B.; Weston, W.; McWhinney, I. R.; McWilliam, C. L.; and Freeman, T. 2013. *Patient-centered medicine: transforming the clinical method*. CRC press.

Xu, W.; Saxon, M.; Sra, M.; and Wang, W. Y. 2022. Self-supervised knowledge assimilation for expert-layman text style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11566–11574.

Yan, A.; He, Z.; Li, J.; Zhang, T.; and McAuley, J. 2022a. Personalized Showcases: Generating Multi-Modal Explanations for Recommendations. *arXiv preprint arXiv:2207.00422*.

Yan, A.; He, Z.; Lu, X.; Du, J.; Chang, E.; Gentili, A.; McAuley, J.; and Hsu, C.-n. 2021. Weakly Supervised Contrastive Learning for Chest X-Ray Report Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 4009–4015.

Yan, A.; McAuley, J.; Lu, X.; Du, J.; Chang, E. Y.; Gentili, A.; and Hsu, C.-N. 2022b. RadBERT: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, e210258.

Yang, K.; and Klein, D. 2021. FUDGE: Controlled Text Generation With Future Discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 3511–3535.

Zhu, W.; Yan, A.; Lu, Y.; Xu, W.; Wang, X. E.; Eckstein, M.; and Wang, W. Y. 2022. Visualize Before You Write: Imagination-Guided Open-Ended Text Generation. *arXiv preprint arXiv:2210.03765*.