Automatic Speech Recognition Error Correction on ICU Clinical Narration Dataset

Stanford CS224N Custom Project

Zhuoyi Huang, Adam Sun, Han Bai
zhuoyih@stanford.edu, adsun@stanford.edu, hanbai@stanford.edu

Abstract

Automatic Speech Recognition (ASR) Yu and Deng (2016) in clinical settings is gaining popularity as clear communication is critical in healthcare delivery. Scenarios in intensive care units (ICUs) are more complex, involving obscure medical terminology and noisy environments. In this paper, we aimed to create an ASR error corrector using a small dataset of nurse-corrected ICU-clinical narration transcribed by WhisperRadford et al. (2022). Given the limited data, we augmented the Mtsamples dataset Boyle (2019) and pretrained a ConstDecoder model Yang et al. (2022) on our augmented dataset, finetuning the model on our own nurse-annotated ICU narration correction dataset. Our findings show that our model is able to outperform baselines and reduce the WER by up to 16%, proving the superiority of our approach, and confirming the model's ability to be a reliable and effective error corrector in the ICU.

1 Introduction

Automatic speech recognition (ASR) models can be used to convert an audio recording of speech to a textual transcription. A survey conducted by Blackley et al. (2019) showed that more than 90% of surveyed hospitals planned to use ASR to assist in the clinical documentation of electronic health records due to its potential to reduce cost, increase efficiency, and reduce the burden of documentation on nurses. However, ASR systems struggle with the intricacies and unpredictability of conversational human language and the quality of voice audio, leading to errors to occur in the outputted transcripts Zhang et al. (2023). Additionally, the adoption of ASR systems in medical settings can be significantly more challenging due to the presence of complex environments with noises from multiple speakers and medical equipment, as well as the use of highly specialized and complicated medical terminology.

In response to the limitations of ASR systems, error correction techniques have been proposed to further refine the outputs of ASR models for certain contexts (Yang et al., 2022)(Leng et al., 2021b). The primary objective of the error corrector is to correct errors present in the outputted sentence generated by the ASR model by utilizing the ground-truth sentence as the target sequence. Previous error correction techniques, such as ConstDecoder and FastCorrect, were designed to address errors inherent in ASR systems, including misspellings, incorrect grammar, and misunderstandings of audio input. For our project, we opted to utilize Whisper Radford et al. (2022) as the transcription tool for narrating intensive care unit (ICU) activities. OpenAI's Whisper stands out for its remarkable robustness and unparalleled accuracy, which greatly reduces the occurrence of errors in ASR systems. However, our objective is not limited to minimizing ASR errors and producing an accurate transcription of the spoken content; instead, our goal, which sets our task apart from previous ASR error correction tasks, is to create a concise and medically precise record that effectively communicates the activities related to patient care in the ICU.

As a result, our mission extends beyond addressing ASR errors to include the detection and correction of disfluencies and irrelevancies, rephrasing, and note continuation, all of which may affect the

quality of the clinical notesLybarger et al. (2017). The main edit types in our task are listed at 1. Notice that most of these errors are not the fault of the ASR model itself, but rather indicative of the noisy nature of our data.

Table 1: Correction Types and examples in Nurse Corrected Dataset

Correction Types	ASR results	Corrected by Nurse
ASR Error	"The patient 6."	"The patient is asleep."
Filler Words Removal	"Alright." "I'll just start."	DELETE
Repetition Removal	"Nurse is straightening out the Foley tubing, Foley catheter tubing."	"Nurse is straightening out the Foley catheter tubing."
Medical Words Correction	"The nurse is replacing the patient's EKG leaves."	"The nurse is replacing the patient's EKG leads."
Remove Irrelevant, Noisy Sentences	"Oh, it's recording."	DELETE
Reduce Ambiguity	"The patient is sitting in bed at a 30 to 45 degree angle."	"The patient is sitting in bed at 45 degree angle."

Despite considerable research on the use of ASR technology for clinical documentation, the majority of studies have focused on radiology departments (39.3%), with emergency medicine accounting for 8.2% from 1990 to 2018Blackley et al. (2019) However, research on the application of ASR in other specialties, especially in the ICU, is lacking Blackley et al. (2019)Lybarger et al. (2018). In our project, we aim to implement an ASR error correction and cleaning system that is both efficient and effective in the clinical setting. Our goal is to ensure that the model can accurately interpret transcriptions produced by ASR systems, in line with the corrections that would actually be made by nurses. To this end, we applied an existing error correction method, ConstDecoderYang et al. (2022), and trained it with our limited clinical narration dataset, which only contains 473 sentences. To compensate for the small size of our dataset, we developed a data augmentation pipeline that allowed us to incorporate a larger medical transcription dataset sourced from mtsamples Boyle (2019) for pre-training. We then fine-tuned the resulting model on our own dataset. Finally, we conducted ablative studies to further assess the role of each aspect of our approach. We replaced the base BERT model with BioBERT - a BERT model that has been pre-trained on a vast corpus of biomedical documents –, tested different combinations of augmentations, and finally tested the model's performance after pretraining on an augmented version of a non-medical narration dataset, QuerYD Oncescu et al. (2020).

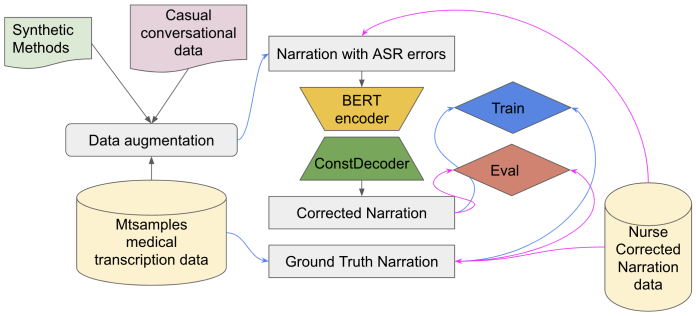

Figure 1: Schema Diagram illustrating the workflow of our method. The Mtsamples transcription data is augmented and combined with casual conversational data and passed through the encoder-decoder model, and is compared with the ground truth uncorrupted samples from the medical transcription data during training. During the fine tuning and evaluation process, we use our own nurse-corrected dataset. We pass in the raw ASR-outputted sentences through the model and compare the outputs to the ground truth corrected sentences to assess model performance.

2 Related Work

Several systematic reviews, such as those conducted by Johnson et al. (2014), Hammana et al. (2015), and Hodgson and Coiera (2016), have investigated the impact of ASR in the clinical setting. These reviews have consistently found that ASR can bring significant benefits to clinical documentation, but have also identified concerns regarding the accuracy of ASR,speed of transcription, and the time required for editing. Lybarger et al. (2017) and Lybarger et al. (2018) analyzed the errors in ASR-assisted clinical documents by detecting differences between ASR-transcribed notes and nurses' intended documentation, which in turn identified opportunities for improvement in the clinical note creation process. Several studies have investigated the frequency and clinical implications of ASR errors in dictated notes. Common types of edits made during the clinical note creation process include correcting speech recognition errors, correcting disfluencies, standard phrasing and formatting, as well as rephrasingLybarger et al. (2017) Lybarger et al. (2018)Zhou et al. (2018). These paper share similar errors in our ICU narration datset.

ASR error correction techniques provide an effective and efficient solution to reduce the time required to edit ASR transcripts while improving the quality of clinical notes by detecting and correcting errors of outputs from ASR systems. This method can be structured as a sequence-to-sequence generation task, where the ASR-transcribed text acts as the input source sequence, and the ground-truth speech-to-text transcription serves as the target. There are typically two main methods to perform this task: autoregressive (AR) models and non-autoregressive (NAR) models. Autoregressive methods decode the target sequence by conditioning each output on previously generated outputs. D'Haro and Banchs (2016) proposed an ASR correction method that involves utilizing a phrase-based machine translation system. Liao et al. (2020) integrated the MASS Song et al. (2019)) pre-training into ASR correction, aiming to enhance the human-readability of the ASR. The MASS approach adopts the encoder-decoder framework to reconstruct sentence fragments by predicting the masked tokens. Weng et al. (2020) proposed a word confusion pointer network (WCN-Ptr) model with multi-heads self attention to jointly address ASR error correction and language understanding (LU) model. Additionally, Mani et al. (2020) utilized a Transformer-based sequence-to-sequence model to train an ASR correction model in an autoregressive manner. Li et al. (2021) presented a model that utilized the pre-trained BERT in the encoder and a copying mechanism in the decoder for ASR error correction.

However, autoregressive models face a major bottleneck, as they cannot be trained in parallel, thus greatly increasing their latency. In contrast, non-autoregressive models greatly reduce sequence generation time. In contrast to prior research, Zhang et al. (2023) utilized non-autoregressive (NAR) models in lieu of autoregressive (AR) models. The joint encoders for both text and phoneme inputs were used to provide inputs to the attention mechanism in the NAR decoder. ConstDecoder Yang et al. (2022), similar to Li et al. (2021), uses a pretrained BERT encoder to encode the input token sequence. Then, the decoder utilizes the encoded hidden representation, the past input, and the attended context vector to output the corrected tokens. On the other hand, FastCorrect Leng et al. (2021b) harnesses non-autoregressive generation by aligning the source and target sentence via edit distance, and training a length predictor for parallel generation. FastCorrect trains on AI-SHELL, a small publicly available large dataset, and adds noise to create a synthetic dataset for pretraining. FastCorrect2 Leng et al. (2021a) improves upon this architecture by utilizing a voting-based system with multiple candidate transcriptions to further inform the model.

We decide to use a non-autoregressive model based approach based on Yang et al. (2022), as our annotation pipeline necessitates an efficient yet effective correction. Due to the small size of our dataset, we follow Leng et al. (2021b), augmenting a larger dataset to create a pseudo correction dataset to train our model on during pretraining. However, our method is set apart from previous work due to our use of our own novel dataset annotated by real nurses, and we also use more novel and robust augmentation methods to account for the disfluencies and irrelevant sentences in our data. See section 3.2 for more information and justification on model usage.

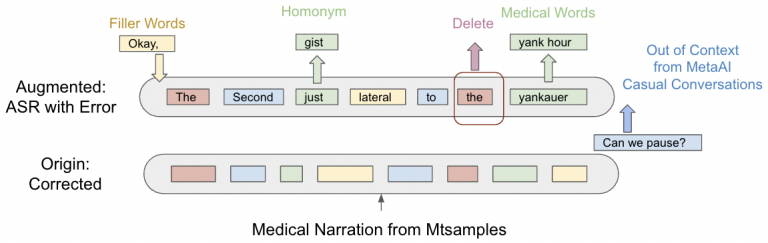

Figure 2: Our data augmentation approach.

3 Approach

3.1 Pre-train Data Synthesis

We use two sources of data as illustrated in Figure 1.

1. ICU Narration Data. The data is collected by Stanford's Partnership for AI-Assisted Care research group and transcribed by Whisper Radford et al. (2022).
2. Mtsamples Boyle (2019) Data. A dataset of big collection of transcribed medical reports.

In order to construct the evaluation dataset on ICU Narration Data, after collecting noisy and stacked narration data from nurses, we aligned it into ground truth-corrected pairs by calculating the sentence embedding similarity using the SpaCyHonnibal et al. (2020) NLP Library. In order to expand the public available data for training a complete model for our unique ICU narration ASR error types, we generated a training set through augmentation techniques. The augmented dataset was primarily sourced from approximately 40,000 medical transcripts that were scraped from `mtsamples.com`. Augmentation methods are shown in 2, and can be classified into:

⋆ **Out-of-Context Sentences.** We incorporated Meta AI's Casual Conversations DatasetHazirbas et al. (2021) into our training set. This data was inserted randomly as out-of-context noise to replicate the background noise nurses experience while recording narrations, which should be removed during correction.
⋆ **Filler Words Insertion.** We randomly inserted 84 frequent filler words, such as "basically" and "okay," between words in a sentence. These words do not contribute to the meaning of a sentence and should be removed.
⋆ **Words Deletion.** We applied a random deletion technique to simulate scenarios in which nurses may inadvertently omit certain words or phrases during their narrations. Specifically, we randomly deleted words from different positions within sentences to simulate the impact of incomplete phrases that can happen when nurses are narrating.
⋆ **Homonym words.** We randomly substituted words with their closest homonym. The pre-generation of a comprehensive homonym dictionary with Python's SoundsLike library for every word in the medical narration dataset significantly reduced the time required to search for homonymous words per word from 0.2s to 10e-4s.
⋆ **Medical Words.** We've assembled a list of medical terminology (e.g. name of medical devices) that are frequently misidentified by ASR and are seldom found in other domain datasets. We randomly insert the medical word in the original sentence and insert the corresponding misrecognized word in the corrupt sentence.
⋆ **Adding Repetition.** To simulate nurses' efforts to complete or rephrase words, we randomly added repeated words of varying lengths to the beginning or end of a sentence.

Our sentence alignment preprocessing algorithm and data corruption technique are original, enabling us to construct a synthetic dataset that is well-suited for our training purpose, and be able to make up for the lack of real data for our unique ICU narration errors.

3.2 Model

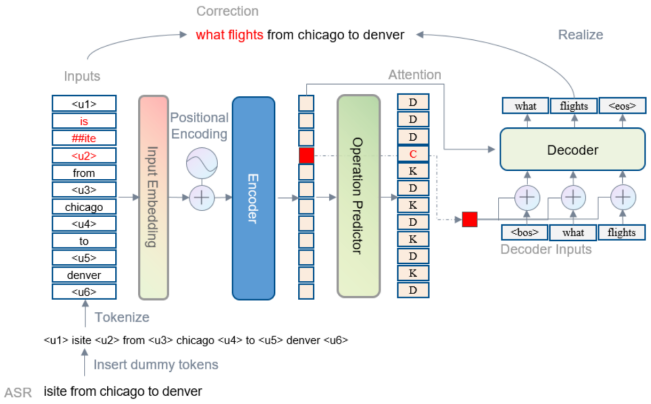

Figure 3: ConstDecoder Yang et al. (2022) model architecture. Taken from paper.

The model we choose for this project is ConstDecoder. Introduced by Yang et al. (2022), ConstDecoder uses a BERT encoder to encode the input token sequence. Then, the decoder utilizes the encoded hidden representation, the past input, and the attended context vector to output the corrected tokens. ConstDecoder introduces a novel operation predictor, which predicts whether each token in the sequence should be kept (K), deleted (D) or changed (C), allowing it to only process the C tokens during the decoding phase, which reduces latency. A combination of operation loss and generation loss is used for the model, allowing it to simultaneously optimize the operation predictor and the decoder. See 3 for diagram of the architecture.

We select ConstDecoder for our investigations due to its low latency, high efficiency, and usage of pretrained BERT encoder to encode hidden representations of the input. The pretrained BERT encoder used in the model greatly reduces the burden during training, reduces overfitting, and allows for better overall performance Yang et al. (2022). To further explore the effect of the pretrained BERT encoder, we compared two different pretrained weights: the base weights ("bert-base-uncased"), which was trained on a general corpus of Wikipedia articles and books, and BioBERT Lee et al. (2019), which was additionally trained on PubMed and PMC articles, granting it a biomedical context.

4 Experiments

4.1 Data

We employed the augmented dataset described earlier to pretrain our model, splitting it into training and validation sets in an 80:20 ratio. For finetuning, we used our own narration dataset, which was divided into training, validation, and testing sets in a 60:20:20 ratio.

The data we augmented was sourced from a dataset of 5000 medical transcripts scraped from mtsamples.com Boyle (2019). This database contains transcribed medical reports from various medical disciplines, including general medicine, radiology, surgery, and emergency care. Using this dataset not only provided the same third-person narrative tone present in our own data, but also exposed our model to a broader range of medical contexts and vocabulary, which increased the generalizability of our pretrained model in the medical domain.

Our own narration dataset comprises 473 nurse narrations of ICU videos that were transcribed using ASR, along with the corresponding corrected sentences that the nurses provided after reviewing the ASR output. The transcriptions encompass a variety of activities that take place in the ICU, such as repositioning patients in their beds, administering medication, performing oral/tracheal suctioning,

	Model	WER	WERR
ASR raw output	No correction	0.1149	-
Baselines	BART	0.128	-0.114
	T5-small	0.5810	-4.0566
	FLAN T5-small	0.5100	-3.4386
Our Methods	Base BERT + ConstDecoder	0.2021	-0.7589
	BioBERT + ConstDecoder	0.1033	0.1011

Table 2: Evaluation results with Encoder-Decoder architecture on nurse-corrected ASR Narration Dataset.

and attaching IVs to patients. By asking the nurses to correct the ASR output immediately after narrating the ICU video, we ensured that the meaning of each corrected sentence remained consistent with the original speaker's intent, which is a crucial criterion for the success of ASR correction systems.

4.2 Baseline and Evaluation method

We used multiple baselines to compare the performance of the Whisper model. We directly used the raw ASR sentences without correction as a naive baseline reference. In addition, we adopted seq2seq Transfer Transformer(T5)-smallRaffel et al. (2019), FLAN-T5-smallChung et al. (2022), and BART Lewis et al. (2019) models as our other baselines. We feed the Whisper ASR output directly to these three models, after passing it through their tokenizer. This method does not require retraining or fine tuning and is only invoked during inference time. This baseline method is consistent with Dutta et al. (2022).

We adopted Word Error Rate (WER) as the performance metric to measure the performance of error correction. WER is a commonly used metric to evaluate the performance of a speech recognition or machine translation system and its calculation is shown as

$$WER = \frac{S + D + I}{N}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference. We removed all punctuation from the sentence when evaluating this metric.
We also use Word Error Rate Reduction Leng et al. (2021b), calculated as

$$WERR = \frac{WER_{raw} - WER_{corr}}{WER_{raw}}$$

where WER_{raw} is the WER without correction and WER_{corr} is the new WER after correction, with a more positive WERR indicating that the model is more effective at reducing the WER, and a negative WERR indicating that a method is actually making the transcription worse.

4.3 Experimental details

We pretrained the model for 10 epochs with a learning rate of $1e-6$ and a batch size of 32 on our augmented dataset. The epoch with the lowest overall validation loss is kept. Then, the model is fine tuned on our narration dataset for 50 epochs, with a learning rate of $1e-7$ and a batch size of 16. Training was performed on a NVIDIA A100 GPU. Pretraining on our large augmented dataset takes around 2 hours, while finetuning on our small dataset only took approximately 5 minutes.

4.4 Results

We first compare our overall model results between the BioBERT Lee et al. (2019) and the Base BERT encoder to measure the effect of a medical context in our encoder, and the overall performance of our model as a whole against our baselines. Results can be seen in 2.

Then, to assess the effectiveness of each augmentation approach we use, we conduct an ablative study to measure the effect of the number of augmentations on final model performance, which can be seen in 3. We conduct withhold individual augmentations from our process to assess the effect on our results.

Training Dataset	OC	FW	MA	RA	WD	HW	Model and Training	WER	WERR
ICU narration							ASR Raw Output	0.1149	
MtSamples	✓						BioBERT + ConstDecoder PT on Augmented dataset FT on ICU dataset	0.1007	0.1236
	✓	✓						0.0988	0.1401
	✓	✓	✓					0.0968	0.1575
	✓	✓	✓	✓				**0.0962**	**0.1628**
	✓	✓	✓		✓			0.1039	0.0957
	✓	✓	✓			✓		0.1149	0
	✓	✓	✓	✓	✓	✓		0.1033	0.1010

Table 3: Augmentation methods ablation results using BioBERT + ConstDecoder, fine tuned on nurse narration dataset. OC stands for Out of Context Conversaton Insertions, FW stands for Filler Words Insertions, MA stands for Medical Words Addition, RA stands for Repetition Addition, WD stands for Words Deletion, HW stands for Homonym Words Modification.

Finally, we measure the effect of a more general augmented dataset on overall performance 4. To do so, we utilize QuerYD, a video dataset with textual and audio narrations in very general contexts Oncescu et al. (2020). The transcripts in this dataset describe general scenes from YouTube videos in a narrative tone that is also present in our validation data. We directly use the transcripts for this experiment, running each raw sentence through our full augmentation pipeline to create a pretraining dataset.

5 Analysis

As can be seen in 2, our model with BioBERT encoder is the only model among our models and baselines to actually improve the word error rate, reducing it by more than 10%. This confirms our hypothesis that using a medically-relevant encoder assists our model in performing error correction in our solely medical context. Surprisingly, many of our baselines made the resulting transcription worse. We largely attribute this to the fact that these baselines are not trained on our nurse annotated dataset, and are not adapted to correct common ASR errors.

Our model can perform corrections on almost all out of context conversation removal, which can be seen from 5. It is also able to perform quite well on medical word modification, as it was able to successfully change the word "Aleven" to "allevyn". However, our model still has trouble inferring incomplete sentences, such as inferring "The nurse covers the patient with a blanket" from the incomplete sentence "The nurse". This task is exceedingly difficult given the limited context given to our model. Further errors discovered in our model's performance include an inability to infer punctuation symbols from their words (like comma and ,), and inability to remove uncertainty. These errors leave more room for future augmentations and modifications to our approach in future work.

Our augmentation ablation study performed in 3 demonstrates that the optimal combination of augmentations includes out of context conversation insertions, filler word insertions, medical words addition and repetition addition. The usage of word deletion and homonym word modification negatively affects the performance of our model. We attribute this to the nature of our fine-tuning

	Model	**WER**	**WERR**
ASR raw output	No correction	0.1149	-
Augmented Dataset	BioBERT + ConstDecoder	**0.1033**	**0.1011**
QuerYD	Base BERT + ConstDecoder	0.2034	-0.770
	BioBERT + ConstDecoder	0.1129	0.0174

Table 4: Evaluation results with Encoder-Decoder architecture on nurse-corrected ASR Narration Dataset vs. model trained on Oncescu et al. (2020)

Table 5: Qualitative Results of Model Correction vs Nurse Correction

Error Type	Raw Sentence	Corrected by Model	Corrected by Nurse
Out of Context Removal	"I think it froze"	DELETE	DELETE
Filler Words Removal	"Alright"	DELETE	DELETE
Medical Words Correction	"The patient had a sacral Aleven in place"	"The patient had a sacral allevyn in place"	"The patient had a sacral Allevyn in place"
Repetition Removal	The nurse is placing a second bio patch on central line dressing site the central line site	The nurse is placing a second bio patch on the central line dressing site	The nurse is placing a second bio patch on the central line site
Punctuation Removal	The patient has sat down comma They are readjusting the lines	The patient has sat down comma They are readjusting the lines	The patient has sat down, they are readjusting the lines
Incomplete Sentences	"The nurse"	"The nurse"	"The nurse covers the patient with a blanket"
Reduce Ambiguity	"The nurse opened a I think that might have been a bio patch container"	"The nurse opened a I think that might have been a bio patch container"	"The nurse opened a biopatch package"

narration dataset, which contains an unproportional number of extraneous words that necessitate deletion. However, as seen in 5, random word deletion and homonym modification still appear in real-world data, so we believe further investigation is needed in gathering a larger dataset and assessing the robustness of these methods.

Our investigations of training on a general video narration dataset (4), demonstrate that the video narration dataset does not provide sufficient medical context to create a good error corrector. The model's performance with Base BERT actually increases the WER. While BioBERT does assist in helping the model's performance, it still does not reduce the WER sufficiently, proving that using Mtsamples Boyle (2019) dataset in pretraining does indeed help our model's performance.

6 Conclusion

In conclusion, we propose, implement, and evaluate an ASR error correction approach based on augmentation of data. We gathered our own clinical ICU narration dataset, conducted augmentations on a third-party medical dataset to create a pretraining dataset, and fine tuned on our narration dataset. Our contributions are three-fold: ***(i)*** We are among the first in the field to study ASR error correction for the narration of ICU activities, especially for the removal of disfluencies and irrelevancies; ***(ii)*** We propose an augmentation pipeline that enables the creation of a pretraining dataset, which in turn, enhances the overall performance of our model; ***(iii)*** We demonstrate that our methods achieve significant quantitative and qualitative improvements compared to our baselines, making it a good candidate for deployment in further downstream tasks.

One major limitation of our project is lack of data. We had original plans to gather a larger test set, but due to reasons beyond our control, our annotation sessions with nurses repeatedly were delayed. Having access to a larger fine tuning set will further improve the performance of our model and make it less biased and more robust and reliable for deployment. We were also not able to explore model architectures beyond ConstDecoder Yang et al. (2022) in this work due to time constraints (architectures like FastCorrect Leng et al. (2021b) required training everything, including the tokenizer, completely from scratch).

Future work includes further exploring the potential of nurse-annotations, potentially through considering multiple candidate corrections. By exploring reinforcement learning with human feedback as in Ouyang et al. (2022), we can further utilize the expertise of the nurses to ask them to rank these candidate corrections, which can further inform our model create final corrections that are even more consistent with nurse intent.

References

Suzanne V Blackley, Jessica Huynh, Liqin Wang, Zfania Korach, and Li Zhou. 2019. Speech recognition for clinical documentation from 1990 to 2018: a systematic review. *Journal of the american medical informatics association*, 26(4):324–338.

Tara Boyle. 2019. Medical transcriptions.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Samrat Dutta, Shreyansh Jain, Ayush Maheshwari, Ganesh Ramakrishnan, and Preethi Jyothi. 2022. Error correction in asr using sequence-to-sequence models. *arXiv preprint arXiv:2202.01157*.

Luis Fernando D'Haro and Rafael E Banchs. 2016. Automatic correction of asr outputs by using machine translation. In *Interspeech*, volume 2016, pages 3469–3473.

Imane Hammana, Luigi Lepanto, Thomas Poder, Christian Bellemare, and My-Sandra Ly. 2015. Speech recognition in the radiology department: a systematic review. *Health Information Management Journal*, 44(2):4–10.

Caner Hazirbas, Joanna Bitton, Brian Dolhansky, Jacqueline Pan, Albert Gordo, and Cristian Canton-Ferrer. 2021. Towards measuring fairness in AI: the casual conversations dataset. *CoRR*, abs/2104.02821.

Tobias Hodgson and Enrico Coiera. 2016. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the american medical informatics association*, 23(e1):e169–e179.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. 2014. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14(1):1–14.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yichong Leng, Xu Tan, Rui Wang, Linchen Zhu, Jin Xu, Wenjie Liu, Linquan Liu, Tao Qin, Xiang-Yang Li, Edward Lin, et al. 2021a. Fastcorrect 2: Fast error correction on multiple candidates for automatic speech recognition. *arXiv preprint arXiv:2109.14420*.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiang-Yang Li, Edward Lin, and Tie-Yan Liu. 2021b. Fastcorrect: Fast error correction with edit alignment for automatic speech recognition. *CoRR*, abs/2105.03842.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Wenkun Li, Hui Di, Lina Wang, Kazushige Ouchi, and Jing Lu. 2021. Boost transformer with bert and copying mechanism for asr error correction. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–6. IEEE.

Junwei Liao, Sefik Emre Eskimez, Liyang Lu, Yu Shi, Ming Gong, Linjun Shou, Hong Qu, and Michael Zeng. 2020. Improving readability for automatic speech recognition transcription. *Transactions on Asian and Low-Resource Language Information Processing*.

Kevin Lybarger, Mari Ostendorf, and Meliha Yetisgen. 2017. Automatically detecting likely edits in clinical notes created using automatic speech recognition. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1186. American Medical Informatics Association.

Kevin J Lybarger, Mari Ostendorf, Eve Riskin, Thomas H Payne, Andrew A White, and Meliha Yetisgen. 2018. Asynchronous speech recognition affects physician editing of notes. *Applied Clinical Informatics*, 9(04):782–790.

Anirudh Mani, Shruti Palaskar, Nimshi Venkat Meripo, Sandeep Konam, and Florian Metze. 2020. Asr error correction and domain adaptation using machine translation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6344–6348. IEEE.

Andreea-Maria Oncescu, Jõao F. Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. 2020. Queryd: A video dataset with high-quality textual and audio narrations. *arXiv:2011.11071*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Yue Weng, Sai Sumanth Miryala, Chandra Khatri, Runze Wang, Huaixiu Zheng, Piero Molino, Mahdi Namazifar, Alexandros Papangelis, Hugh Williams, Franziska Bell, et al. 2020. Joint contextual modeling for asr correction and language understanding. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6349–6353. IEEE.

Jingyuan Yang, Rongjun Li, and Wei Peng. 2022. Asr error correction with constrained decoding on operation prediction.

Dong Yu and Li Deng. 2016. *Automatic speech recognition*, volume 1. Springer.

Ziji Zhang, Zhehui Wang, Rajesh Kamma, Sharanya Eswaran, and Narayanan Sadagopan. 2023. Patcorrect: Non-autoregressive phoneme-augmented transformer for asr error correction. *arXiv preprint arXiv:2302.05040*.

Li Zhou, Suzanne V Blackley, Leigh Kowalski, Raymond Doan, Warren W Acker, Adam B Landman, Evgeni Kontrient, David Mack, Marie Meteer, David W Bates, et al. 2018. Analysis of errors in dictated clinical documents assisted by speech recognition software and professional transcriptionists. *JAMA network open*, 1(3):e180530–e180530.