An Enhanced Model for ASR in the Medical Field

Hsu, Wei-Chen
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
g10530021@alum.ccu.edu.tw

Lin, Pei-Xu
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
a367353933@gmail.com

Li, Chi-Jou
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
annyoao@gmail.com

Tien, Hao-Yu
Department of Information Management
National Chung Cheng University
Chiayi, Taiwan
stu710319@gmail.com

Kang, Yi-Huang
Department of Information Management
National Sun Yat-sen University
Kaohsiung, Taiwan
ykang@mis.nsysu.edu.tw

Lee, Pei-Ju
Institute of Data Science and Information Computing
National Chung Hsing University
Taichung. Taiwan
pjlee@nchu.edu.tw

***Abstract*—The application of speech recognition has been utilized to enhance work efficiency and even improve quality of life. Previous studies have indicated that using general-purpose speech recognition models for domain-specific recognition yields suboptimal results. Therefore, this study aims to enhance speech recognition models and optimize the correction of generated errors to improve the model's performance in the medical field. This study employed the PubMed dataset to optimize the speech recognition language model and dictionary. Subsequently, the output results were subjected to error correction using the Bert2Bert, BioBert2BioBert, and BART model architectures. Finally, the model was tested using out-of-domain data from the MedDialog dataset to ensure its robustness. It is anticipated that in the future, the model architecture can be applied to the medical domain to enhance efficiency in scenarios such as medical consultations.**

Keywords—text mining, speech recognition, error correction, deep learning

I. Introduction

The rise of Artificial Intelligence (AI) has seen exponential growth in various fields [1]. Among them, the performance achieved through Deep Neural Network (DNN) technology in Automatic Speech Recognition (ASR) and image recognition has been particularly outstanding [2]. In recent years, due to the continuous improvement in ASR technology, the accuracy has also been greatly enhanced, and various companies have applied ASR in various fields, packaging them into various products and services for consumers. Companies like Apple, Google, Amazon, and others have also introduced smart speakers, which understand user commands through speech recognition to control services like music playback and weather queries, providing users with convenience. In professional domains, the use of speech recognition technology has also become beneficial for specific populations. According to the World Health Organization (WHO), approximately 15% of the global population suffers from varying degrees of speech impediments.

The increasing demand for Automatic Speech Recognition (ASR) across different domains has led scholars to invest more time in improving speech recognition. Traditional ASR systems typically rely on Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM), which constitute a statistical-based recognition model. ASR models based on HMM involve the combination of multiple modules, prompting many scholars to conduct research and adjustments on various ASR modules such as signal processing, natural language preprocessing, and speech reception devices [3]. However, each module is independent and cannot consider the inter-module correlations, training is usually complex and costly. Consequently, many researchers have begun to explore end-to-end ASR models. End-to-end ASR integrates traditionally disparate modules into a single model, enabling the use of a unified evaluation metric to improve model performance, resulting in more flexible training methods and representing another direction for ASR research.

In the past, scholars have attempted to achieve better ASR results by optimizing different modules within ASR systems. For instance, [4] enhanced ASR noise robustness by testing clean and noisy speech using a DNN-HMM acoustic model. Some scholars have also sought to improve ASR by expanding the vocabulary of language models [5] [6]. Reference [7] proposed an N-best rescoring system that re-evaluates utterances by extracting attention information for confused words. Reference [8] introduced a method that separates language models from end-to-end ASR, enabling independent updates of language models when using end-to-end ASR models. As ASR often struggles with non-native speech data, [9] employed voice mapping techniques to enhance non-native German speech data by mapping it to native English data. These studies demonstrate that scholars have effectively addressed various ASR limitations, resulting in significant improvements.

However, there has been limited research on improving ASR in specialized domains. According to the research by [10], using general-purpose ASR models to transcribe specialized domain speech content may lead to inaccurate results. Therefore, this study aims to retrain ASR models specifically for the medical domain, aiming to bridge the research gap in specialized domain ASR. This study aims to utilize two types of text data: PubMed as the basis for retraining ASR Language Models (LM), and the MedDialog medical question-answering dataset [11] as the testing dataset. This study focuses on adjusting ASR Language Models for medical vocabulary. Therefore, the study adopts the open-source JHU ASpIRE model [12] as the foundational ASR model due to its flexibility. The objectives of this study are as follows: (1) Retrain ASR Language Models using medical texts from PubMed to learn professional terminology in the medical field; (2) Utilize pre-trained word embeddings from various

2835-5776/24/$31.00 ©2024 IEEE
DOI 10.1109/IRI62200.2024.00021

medical domains for model training and evaluate the performance of different embeddings; (3) Optimize ASR output results using various deep learning methods to achieve better accuracy.

II. Literature Review

The basic ASR model consists of a feature extractor, an acoustic model, a language model, and a decoder. This section will sequentially introduce the aforementioned basic modules [13].

a) Feature Extractor

Feature extraction is a fundamental preprocessing step in pattern recognition and machine learning, widely applicable across various fields. When performing feature extraction, input sequences are transformed into a set of features, providing useful information for subsequent tasks. By retaining useful information from the input sequence while discarding unnecessary details, feature extraction serves as a form of dimensionality reduction technique [14]. In the field of speech recognition, feature selection is extensively applied, including basic speech analysis, synthesis, recognition, enhanced utilization, speaker identification, and speech correction, among other practical applications in daily life. As speech signals exhibit temporal variability, feature extraction aims to reduce the variability of speech signals. Reference [15] divided feature extraction in ASR into three stages: firstly, using speech analysis to perform spectro-temporal analysis of speech signals and generate raw features of short speech segments; next, compiling extended feature vectors composed of dynamic and static features; finally, transforming these extended feature vectors into more concise and robust vectors for subsequent implementation by the recognizer. A common method of feature extraction is Mel-frequency cepstral coefficients (MFCC).

b) Acoustic Model

The acoustic model is a crucial module in ASR systems. Typical systems often use Hidden Markov Model-Gaussian Mixture Model (HMM-GMM) to establish a statistical-based acoustic model. After speech input undergoes feature extraction, the transformed features serve as the input to the acoustic model. Each word may consist of one or multiple phonetic units, each assigned a phoneme label, with each phoneme having its own HMM. Combining acoustic feature vectors, the speech decoder calculates the probability of input speech signals matching different phonemes, thereby determining the most suitable result. In the past, ASR acoustic models mostly used HMM-GMM as the modeling approach. In recent years, many scholars have replaced the GMM in HMM-GMM acoustic models with various Deep Neural Networks (DNN), with research indicating that training large-scale ASR tasks using DNN yields better results than GMM [16] [17]. Because DNN is well-suited for discriminative learning, especially with sufficient annotated data, more accurate results can be achieved [18].

c) Language Model

The language model represents a set of possible given words, assigning statistical estimates to each word based on generating grammar rules or retraining a corpus. Many words have similar pronunciations, but with context, listeners can understand the meaning of homophones based on the appearance of words or phrases in context. The purpose of the language model is to provide context and phrases for the ASR recognition result to judge and improve possible word sequences. Language models calculate the probability of word sequences occurring using n-grams. Through training on millions of word-marked data, the perplexity of training data is reduced to improve ASR recognition results [19]. Common language models include bigram and trigram models, which calculate the probability of groups of two or three words in a word sequence.

d) Decoder

In the decoding stage, the optimal word sequence is obtained by providing the observation sequence and language model. The decoding process utilizes the Viterbi algorithm, a dynamic programming algorithm. The algorithm employs a greedy approach to find the best matching words in the decoding network. However, this may overlook better word sequences due to the greedy approach. Therefore, the beam search algorithm is often used for improvement. By expanding the beam width, candidate lists of preceding and following words can be sorted, effectively avoiding the possibility of missing the best output.

2) Automatic Speech Recognition Error Correction Methods

In general, ASR error correction includes error detection in the entire process. Internal ASR module optimization includes the acoustic model, alignment, and language model [22][23]. In particular, in advanced ASR systems, the language model is usually decoded twice. Initially, hypotheses are generated using an n-gram language model, followed by re-scoring the generated hypotheses using a Neural Network based Language Model (NNLM) [24]. NNLM has been proven to significantly improve ASR performance, with some models outperforming traditional n-gram LMs [25]. Research also demonstrates that re-scoring the hypothesis list of the language model using lattice rescoring is an effective strategy [25][26][27]. Therefore, [28] proposed a method to perform lattice rescoring on LSTMLM, which improved WER by 8% compared to using an n-gram LM in the recognition results of YouTube videos.

Reference [29] proposed a new NNLM called Neural Error Corrective Language Models (NECLM), which consists of encoder and decoder networks. The encoder constructs a context vector using N-best lists and confidence scores generated by the ASR recognizer, while the decoder uses the vector obtained by the encoder to calculate word generation probabilities to re-score recognition hypotheses, thus correcting ASR errors. Experimental results demonstrate that combining NECLM with two different ASR systems can achieve better results.

Reference [30] proposed a new error correction method by training a supervised spelling correction model to preliminarily modify errors generated by ASR. The ASR output results are then re-scored by an external LM based on the N-best lists generated by the ASR model. Experimental results show that adjusting the hypothesis generation of the ASR spelling correction model to 8 hypotheses and expanding the ASR N-best list from 8 to 64 candidates can reduce the error rate of the original low-noise test set from 6.03% to 4.52%, resulting in significant improvement.

On the other hand, the output of ASR is prone to both speech and spelling errors, so there are also studies proposing methods to improve the ASR output results. These methods typically use Sequence-to-sequence (Seq2Seq) models or Transformers as error correction models. Below, we will introduce error correction models of this type.

Reference [31] proposed a method for re-scoring the output using word-level alignment, applying the pre-trained model BART as an adaptive model. They used a dataset enhanced with common induced errors and actual errors in ASR output to denoise and correct similar errors in the model. The authors used the ROVER system combination technique [32] to build a confusion network based on BART and ASR output results. The final experimental results show that this method effectively corrects ASR errors and reduces WER.

Reference [33] proposed a Transformer architecture that is applicable to machine translation, using Jasper ASR as the foundation for ASR. They combined two different LMs: 6-gram KenLM and Transformer-XL for LM re-scoring, and corrected the ASR system output. When correcting the ASR model, the authors considered methods such as random initialization and using pre-trained BERT weights identical to the Transformer architecture as initialization weights. The results showed that using both external LMs improved the output results. Among them, using Transformer-XL yielded better results than 6-gram KenLM, with WERs of 2.95% and 3.34% respectively for low-noise speech and 8.79% and 9.62% respectively for more noisy speech.

B. Specialized Domain Speech Recognition

In specific domains, many proprietary terms cannot be accurately predicted by general ASR models. Re-training a language model for a specialized domain typically requires a large amount of training data, and collecting sufficient medical data is not easy. Therefore, common research often utilizes domain adaptation techniques to achieve better results in specific domains. Domain adaptation is one application of transfer learning, aiming to extract common latent factors between the source domain and the target domain, and adjust the target domain to reduce the distribution mismatch in feature space between domains. By leveraging knowledge learned from another relevant domain with sufficient annotated data, the performance of the model on the target domain with insufficient annotated data can be improved.

1) Domain Adaptation

Common domain adaptation methods can be divided into training the Acoustic Model (AM) and training the Language Model (LM). However, retraining an AM tailored to the medical domain requires a vast amount of medical speech data, which is not easy to obtain. Therefore, [34] suggested directly modifying the LM, which can greatly reduce the training cost of the model. By using a large-scale question-and-answer dataset, emrQA, the authors extracted medical domain question-answer pairs as unformatted data for domain adaptation. They used the open-source model JHU ASpIRE to adjust the LM. This adjustment method can maintain the original basic structure, so if it needs to be applied to different domains, there is no need to readjust the decoding network to generate different ASR systems. Only the LM output needs to be re-scored, making it an easy-to-update and versatile domain adaptation method.

Reference [35] used a dataset containing 3807 de-identified dialogues between doctors and patients for training, validation, and testing in domain adaptation. To learn the proprietary terms in the medical domain, the authors curated 20,000 medical professional terms from the Unified Medical Language System (UMLS) and improved 200 common medical terms. Training was conducted using Sequence-to-Sequence (Seq2Seq) models and Transformers for domain adaptation, and scoring was performed using WER and BLEU score. The Seq2Seq model reduced the WER from 41.0% to 34.1% in the Google speech API and from 35.8% to 34.5% in JHU ASpIRE. The BLEU score for Google speech API increased from 52.6 to 56.4, and for JHU ASpIRE, it increased from 54.3 to 55.8. These results demonstrate a significant improvement in ASR transcription quality with the Seq2Seq model. However, the Transformer model did not perform well, with a WER of 92.2 and BLEU score of 0.06 for the Google speech API, possibly due to the limited training data.

III. Research Method

A. Research framework

The framework of this study mainly includes four parts: data source, data preprocessing, model establishment and evaluation methods. Figure 3.1 is the process framework of this study, and the following will introduce the process framework of this study.

Past research voice texts will use videos [36], or open source voice datasets [30][31][33]. However, the voice data set in medical field is difficult to obtain, so this study uses Google Text-To-Speech API to convert the text data collected from PubMed into voice format to establish a voice data set in medical field. This data set will be used as the basis for the subsequent speech recognition model retraining, and can also be provided for the post-processing model. In addition to using PubMed as the voice data set, we also set up 981 MedDialog test sets to test the robustness of the model to foreign data.

In this study, after obtaining the output results of ASR, two different network architectures will be used to establish the post-modified model, namely Seq2Seq model and transformer model. Seq2Seq model corrects the text from sequence to sequence, and improves the accuracy of the correction by considering the context and syntactic structure of the text. In the encoder part, we use BERT and BioBERT to encode the text. By comparing these two embedding models, we can know whether the pre-trained embedding model in medical field is helpful to better understand the text. To enhance the model's performance in handling long texts, this study incorporates an attention mechanism into the output of the encoder. This approach allows it to make more accurate judgements based upon contextual features. In the decoder part, LSTM and Transformer Decoder are used. LSTM can capture the time series relationship in the text, while Transformer Decoder can handle the long-distance dependence. By leveraging these diverse network architectures, we aim to fully utilize contextual and syntactic embeddings, thereby enhancing the correction model's ability to comprehend text.

The other model framework is Transformer. This study chooses the pre-training model BART as the basis, and fine-tunes it, so that it can have a more robust performance on the words used in the medical field. The evaluation metrics employed in this study are the WER and BLEU score. These two distinct metrics will be utilized to assess and compare the original output results of ASR with post-correction results.

B. Data Sources

The data for constructing and evaluating our model were obtained from two distinct sources: PubMed and the MedDialog medical question-and-answer dataset. Firstly, PubMed is utilized for training and testing the error correction model, leveraging its wealth of medical expertise to enhance the accuracy and efficiency of our model. Secondly, MedDialog is employed as an out-of-domain medical data test set to evaluate our model's performance in real-world medical scenarios, as well as to verify its generalization ability and practical value.

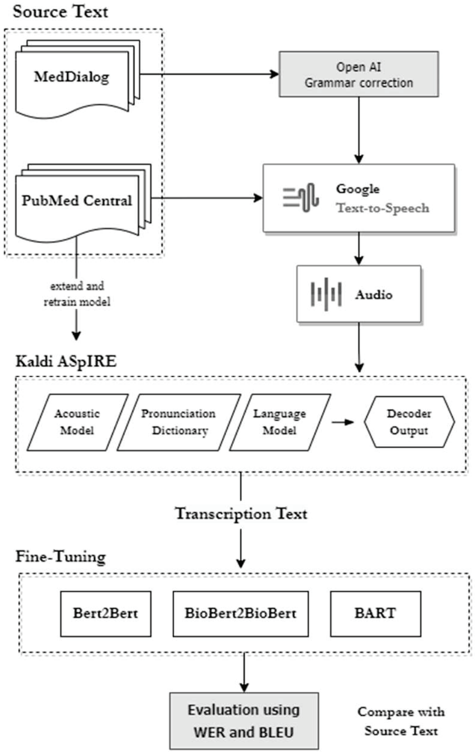

Fig. 3.1 Research Framework Diagram

1) PubMed

PubMed is a medical literature database maintained by the National Center for Biotechnology Information (NCBI) in the United States. It encompasses academic journal articles from various medical fields globally, including medicine, biomedicine, nursing, dentistry, pharmacy, public health, and numerous other specialized areas. As one of the vital resources in the medical industry, PubMed provides a wealth of literature for medical researchers, physicians, and academic institutions alike. This study gathered data from three sub-databases within PubMed, specifically PubMed BMC, PubMed JMIR, and PubMed PLOS-1, yielding a total of 57,967 articles. In order to avoid data containing mathematical operators, bullet points, and examples that do not typically reflect colloquial expressions, we specifically targeted the 'Introduction' section of the abstracts. After careful observation, it was concluded that the content within these sections more closely corresponds with general language usage. Consequently, this selected content will serve as our primary dataset for training the speech recognition model and for post-processing treatments.

2) MedDialog

MedDialog, introduced in 2020, is a large-scale medical question-and-answer dataset. It was created to address the issues of existing medical dialogue datasets, such as small scale, insufficient coverage, and bias towards certain diseases. MedDialog provides datasets in both Chinese and English. The Chinese dataset includes 172 specialized diseases, 3.4 million doctor-patient dialogues, 11.3 million sentences and 660.2 million annotations. Conversely, the English dataset covers 96 specialized diseases, comprising 260,000 dialogues, 510,000 sentences and 44.53 million annotations.

However, in this study, we identified several issues within the patient-doctor dialogues in the MedDialog dataset that need to be addressed such as words were found to have no spaces in between, patients and doctors tend to use abbreviations to describe conditions, doctor responses might include URLs, Q&As may contain usernames of patients or doctors, as well as personal and clinic information, etc.

C. Data Preprocessing

The purpose of data preprocessing is to convert text into a form that is usable for experiments and models while eliminating redundant and irrelevant noise. The data preprocessing in this study will address the error types found in the PubMed and MedDialog medical question-and-answer datasets. The preprocessing methods used in this study included sentence segmentation and punctuation removal, which were performed using the Python natural language processing toolkit, spaCy. This study opted to re-perform web crawling for data acquisition of MedDialog. This approach allowed for preliminary error handling and improvements during the crawling process, which included adding extra spaces to each paragraph of text, deleting URLs within the text, and removing content contained within brackets in the text. Such modifications have resulted in cleaner text that is more conducive for subsequent research analysis. Through grammar correction using OpenAI API, we obtained a test set consisting of 981 pieces of data, termed Med981, which includes conversations between doctors and patients, totaling to 1,962 sentences, used to evaluate the effect of the speech recognition and post-processing models in this study.

D. ASR Domain Adaptation

This study will conduct the adaptation of the speech model to the medical domain in two steps: dictionary creation and language model establishment. First, we will embark on the step of dictionary creation. The purpose of this step is to collect and categorize professional vocabulary and technical terms specific to the medical field, with the aim to establish a lexicon for medicine. The creation of this dictionary will assist us in better understanding and managing medical text, providing precise definitions of vocabulary, and term translations, thereby ensuring a solid foundation for the subsequent establishment of the language model. Secondly, we will carry out the step of language model establishment. In this phase, we will employ the professional vocabulary and technical terms found in the dictionary, combined with extensive medical text data, to create a language model custom-tailored to the medical field. The language model will take into account the contextual elements and syntactic structure of medical texts, which aims to better comprehend and generate medical information. Through the training of this specialized medical domain language model, we aspire to provide more accurate, professional, and fluent speech recognition and generation capabilities, ultimately satisfying the needs of the medical domain.

To enable the ASpIRE speech recognition model to adapt to specific medical terminologies, this study utilized text collected from PubMed BMC, PubMed JMIR and PubMed PLOS-1 for retraining and testing. The collected 57,987 articles were divided into training and testing sets in an 8:2 ratio. After the division, we obtained 46,374 articles as training data, and 11,593 articles for testing. This study performs domain adaptation on the model's own dictionary and language model, aiming to strengthen both components to enhance the performance of ASpIRE on medical texts with more robustness. By expanding the vocabulary in the original speech recognition model and reducing the number of OOV terms, we hope to achieve domain adaptation. Before training the model, we will first tokenise the text to generate a word set, and then commence the creation of the dictionary and language model. To consider context, we employ 3-gram for training the language model. Upon completion of the new dictionary and language model, we will merge them with the original model, thus obtaining the retrained speech recognition model for the medical domain.

E. Error Correction Techniques

After the text is decoded by ASR in this study, we will proceed with the correction of erroneous words in the text. We plan to employ both the Seq2Seq and Transformer models for post-processing, with the aim of improving the quality of the text, minimizing the misinterpretation caused by incorrect words, and ultimately enhancing overall performance and usability.

F. Experimental Setup

This study designed three distinct experiments to evaluate the outcomes of different error correction models through their respective implementations. In the first experiment, the speech recognition model will be retrained based on the medical dataset. The second experiment involves the creation of two different error correction models: Seq2Seq and Transformer, in order to evaluate the performance of the models after training. The third experiment will evaluate the fine-tuned models using out-of-domain datasets.

G. Data Validation and Evaluation Metrics

In this study, K-fold cross-validation is employed for model validation assessment, with the intention to use ten-fold cross-validation. In this study, both WER and BLEU scores will be used as evaluation metrics to compare the performance of various models on ASR output text. The WER and BLEU scores will be computed to observe the performance and results of different models.

IV. Experimental Results and Discussion

Experiment 1 involved the retraining of the ASpIRE speech recognition model to adapt it to medical domain text. To achieve this, the PubMed medical text training set was utilized to expand ASpIRE's dictionary and language model, followed by retraining ASpIRE using the expanded dictionary and language model. The performance of the retrained ASpIRE on medical domain texts was evaluated using the PubMed medical text testing set. Table 4.1 presents the comparison of predictions on the PubMed testing set between the basic ASpIRE model and the retrained ASpIRE model. The results indicate that the WER of the retrained ASpIRE, after expansion and retraining, is 27.40%, which is a significant improvement of 13.54% compared to the original ASpIRE model. Similarly, the BLEU score increased from 0.42 to 0.52, indicating a significant enhancement in the readability of generated medical domain speech and the accuracy of sentences after retraining.

Table 4.1
Comparison of Basic ASpIRE and Retrained ASpIRE Predictions on the PubMed Testing Set

	WER	*BLEU*
BASE ASpIRE	40.94	0.42
Retrained ASpIRE	27.40	0.52

Experiment 2 employed two different types of Seq2Seq models for training: Bert2Bert, BioBert2BioBert, and fine-tuning of the Transformer-based pre-trained model BART. These two categories of models were used as error word correction models, and predictions and evaluations were conducted on three different models and the retrained ASpIRE using the PubMed testing set. Finally, evaluations and comparisons of WER and BLEU scores were performed.

Table 4.2
Comparison of Predictions Results between Seq2Seq and Transformer Models

	WER	*BLEU*
Retrained ASR	27.40	0.52
Retrained ASR+Bert2Bert	25.00	0.63
Retrained ASR+BioBert2BioBert	19.44	0.68
Retrained ASR+BART	16.36	0.71

Table 4.2 displays the prediction results of the Seq2Seq and Transformer models. Based on Table 4.2, we can observe that the results after post-processing are improved compared to ASpIRE solely retrained. Using Bert2Bert resulted in a decrease in WER by 2.40% and an increase in BLEU score by 0.11%. Similarly, employing BioBert2BioBert led to a reduction in WER by 6.69% and an increase in BLEU score by 0.15%.

From these results, we can conclude that using the pre-trained medical embedding model BioBert for both encoder and decoder yields better results in correcting recognition errors in medical texts. This indicates that BioBert2BioBert demonstrates a better understanding of semantics and context in medical domain texts, thereby improving prediction accuracy.

Experiment 3 aimed to evaluate the performance and robustness of the Seq2Seq and BART models. To assess the performance and robustness of the models, we utilized the Med981 dataset as the medical text. Since the Med981 dataset differs from the dataset used in model training, it provides an evaluation of the model's performance on unseen data. By comparing the predicted results generated by the models with the actual answers, we can further assess the accuracy and robustness of the models on out-of-domain data.

TABLE 4.3
PERFORMANCE RESULTS OF ERROR CORRECTION MODELS ON OUT-OF-DOMAIN DATA

	Doctor WER	***Patient WER***	***Doctor BLEU***	***Patient BLEU***
Retrained ASR	**29.98**	**26.47**	0.49	0.54
Retrained ASR+ Bert2Bert	98.13	97.26	0.01	0.01
Retrained ASR+ BioBert2BioBert	52.18	56.44	0.32	0.27
Retrained ASR+ BART	32.42	27.26	**0.5**	**0.55**

According to Table 4.3, it can be observed that when using the out-of-domain medical dataset Med981 as input for model prediction, the Doctor WER and Patient WER are best achieved by the output of the retrained ASpIRE model alone, with scores of 29.98% and 26.47% respectively. However, incorporating post-processing models resulted in an increase in WER. Among them, the Bert2Bert model exhibited the poorest performance, with Doctor WER and Patient WER of 98.13% and 97.26% respectively, with BLEU scores of 0.01.

This study hypothesizes possible reasons for the increase in WER as follows: (1) Differences in data types, although only the Introduction section of PubMed abstracts was used as training data for the model, there are still differences in the nature of the oral question-and-answer dataset Med981. (2) In the Med981 dataset, doctors typically respond to patient medical inquiries, which may include different types of drug names and units of drug consumption, differing from the medical terminology in PubMed texts, resulting in poor output.

The improvement in BLEU score for BART compared to using only the retrained ASpIRE model might be due to the better contextual understanding ability of the model, resulting in improved language fluency and semantic accuracy, thus enhancing the BLEU score.

V. CONCLUSIONS AND DISCUSSIONS

This study aimed to establish a medical speech recognition model by retraining the dictionary and language model of the ASpIRE model using PubMed medical texts to achieve adaptation to the medical domain. Subsequently, error word correction was performed using Seq2Seq and Transformer models, resulting in improvements of 21.50% and 24.58% respectively.

Based on the experimental results of this study, there are several contributions to both academia and practice. In academia, we provide a process for establishing a domain-specific speech recognition model. In practice, we establish a speech recognition model adapted to the medical domain. Furthermore, based on our research model, if further work is needed to establish medical consultation summaries and medical records, our model can serve as a foundation. Through further expansion and application, we can apply speech recognition technology to the process of automatically generating medical consultation summaries and medical records, further improving the efficiency of organizing and managing medical information.

For future research directions, it is suggested to consider utilizing authentic medical consultation records as the dataset to better capture common medical vocabulary and terminology prevalent in practical healthcare settings. This approach can enhance the effectiveness of training the speech recognition model, thereby improving its adaptability and accuracy in real-world medical scenarios.

REFERENCES

[1] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

[2] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

[3] Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic speech recognition: Systematic literature review. IEEE Access, 9, 131858-131876.

[4] Shahnawazuddin, S., Deepak, K., Pradhan, G., & Sinha, R. (2017). Enhancing noise and pitch robustness of children's ASR. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[5] Egorova, E., & Burget, L. (2018). Out-of-vocabulary word recovery using fst-based subword unit clustering in a hybrid asr system. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[6] Orosanu, L., & Jouvet, D. (2018). Adding new words into a language model using parameters of known words with similar behavior. Procedia Computer Science, 128, 18-24.

[7] Kim, H.-G., Lee, H., Kim, G., Oh, S.-H., & Lee, S.-Y. (2017). Rescoring of N-best hypotheses using top-down selective attention for automatic speech recognition. IEEE Signal Processing Letters, 25(2), 199-203.

[8] Xu, H., Khassanov, Y., Zeng, Z., Chng, E. S., Ni, C., Ma, B., & Li, H. (2020). Independent language modeling architecture for end-to-end asr. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[9] Goronzy, S., Rapp, S., & Kompe, R. (2004). Generating non-native pronunciation variants for lexicon adaptation. Speech Communication, 42(1), 109-123.

[10] Mani, A., Palaskar, S., & Konam, S. (2020). Towards understanding ASR error correction for medical conversations Proceedings of the First Workshop on Natural Language Processing for Medical Conversations, http://dx.doi.org/10.18653/v1/2020.nlpmc-1.2

[11] Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., Zhou, M., Zeng, J., Dong, X., & Zhang, R. (2020). MedDialog: Large-scale medical dialogue dataset. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP),

[12] Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., & Khudanpur, S. (2015). Jhu aspire system: Robust lvcsr with tdnns, ivector adaptation and rnn-lms. 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU),

[13] Errattahi, R., El Hannani, A., & Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. Procedia Computer Science, 128, 32-37.

[14] Kurzekar, P. K., Deshmukh, R. R., Waghmare, V. B., & Shrishrimal, P. P. (2014). A comparative study of feature extraction techniques for speech recognition system. International Journal of Innovative Research in Science, Engineering and Technology, 3(12), 18006-18016.

[15] Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. International Journal of Signal Processing, Image Processing and Pattern Recognition, 9(4), 393-404.

[16] Sak, H., Senior, A. W., & Beaufays, F. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling.

[17] Sak, H., Vinyals, O., Heigold, G., Senior, A., McDermott, E., Monga, R., & Mao, M. (2014). Sequence discriminative distributed training of long short-term memory recurrent neural networks.

[18] Sak, H., Senior, A., Rao, K., & Beaufays, F. (2015). Fast and accurate recurrent neural network acoustic models for speech recognition. arXiv preprint arXiv:1507.06947.

[19] Jelinek, F. (1998). Statistical methods for speech recognition. MIT press.

[20] Ainsworth, W. A., & Pratt, S. (1992). Feedback strategies for error correction in speech recognition systems. International Journal of Man-Machine Studies, 36(6), 833-842.

[21] Yu, D., Hwang, M.-Y., Mau, P., Acero, A., & Deng, L. (2004). Unsupervised learning from users' error correction in speech dictation. Eighth International Conference on Spoken Language Processing,

[22] Long, Y., Li, Y., Wei, S., Zhang, Q., & Yang, C. (2019). Large-scale semi-supervised training in deep learning acoustic model for asr. IEEE Access, 7, 133615-133627.

[23] Rasipuram, R., Razavi, M., & Magimai-Doss, M. (2015). Integrated pronunciation learning for automatic speech recognition using probabilistic lexical modeling. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[24] Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. arXiv preprint arXiv:1602.02410.

[25] Sundermeyer, M., Ney, H., & Schlüter, R. (2015). From feedforward to recurrent LSTM neural networks for language modeling. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23(3), 517-529.

[26] Arisoy, E., Sainath, T. N., Kingsbury, B., & Ramabhadran, B. (2012). Deep neural network language models. Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT,

[27] Liu, X., Chen, X., Wang, Y., Gales, M. J., & Woodland, P. C. (2016). Two efficient lattice rescoring methods using recurrent neural network language models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(8), 1438-1449.

[28] Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., & McDermott, M. (2019). Publicly available clinical BERT embeddings. arXiv preprint arXiv:1904.03323.

[29] Tanaka, T., Masumura, R., Masataki, H., & Aono, Y. (2018). Neural Error Corrective Language Models for Automatic Speech Recognition. INTERSPEECH,

[30] Guo, J., Sainath, T. N., & Weiss, R. J. (2019). A spelling correction model for end-to-end speech recognition. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[31] Dutta, S., Jain, S., Maheshwari, A., Ramakrishnan, G., & Jyothi, P. (2022). Error correction in asr using sequence-to-sequence models. arXiv preprint arXiv:2202.01157.

[32] Fiscus, J. G. (1997). A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). 1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings,

[33] Hrinchuk, O., Popova, M., & Ginsburg, B. (2020). Correction of automatic speech recognition with transformer sequence-to-sequence model. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[34] Jiang, Y., & Poellabauer, C. (2021). A Sequence-to-sequence Based Error Correction Model for Medical Automatic Speech Recognition. 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),

[35] Mani, A., Palaskar, S., Meripo, N. V., Konam, S., & Metze, F. (2020). Asr error correction and domain adaptation using machine translation. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

[36] Kumar, S., Nirschl, M., Holtmann-Rice, D., Liao, H., Suresh, A. T., & Yu, F. (2017). Lattice rescoring strategies for long short term memory language models in speech recognition. 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),