Post-decoder Biasing for End-to-End Speech Recognition of Multi-turn Medical Interview

Heyang Liu, Yu Wang and Yanfeng Wang

Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
Shanghai Artificial Intelligence Laboratory

2024.5

Contextual automatic speech recognition (CASR)

Setting

Contextual speech recognition enhances the recognition performance of a human-defined bias list from context or additional information relevant to the speech environment.

Most words show relatively high accuracy, due to their frequencies in the training corpus. But not for rare words.Contextual speech recognition is typically correlated to rare words recognition.

Wide applications: medical consultation, company meeting ...
knowledge-intensive scenerios

Bias list

Most research uses words with low frequencies in the training set.

i.e. frequency < 10
the rank of frequency > 5000

Speech corpus for CASR

Characteristic

A good speech corpus contains rare words with important meanings, i.e. knowledge-intensive corpus with massive named entities.

Such data sets are rarely open source because issues such as privacy protection are often involved, especially in medical or clinical scenarios.

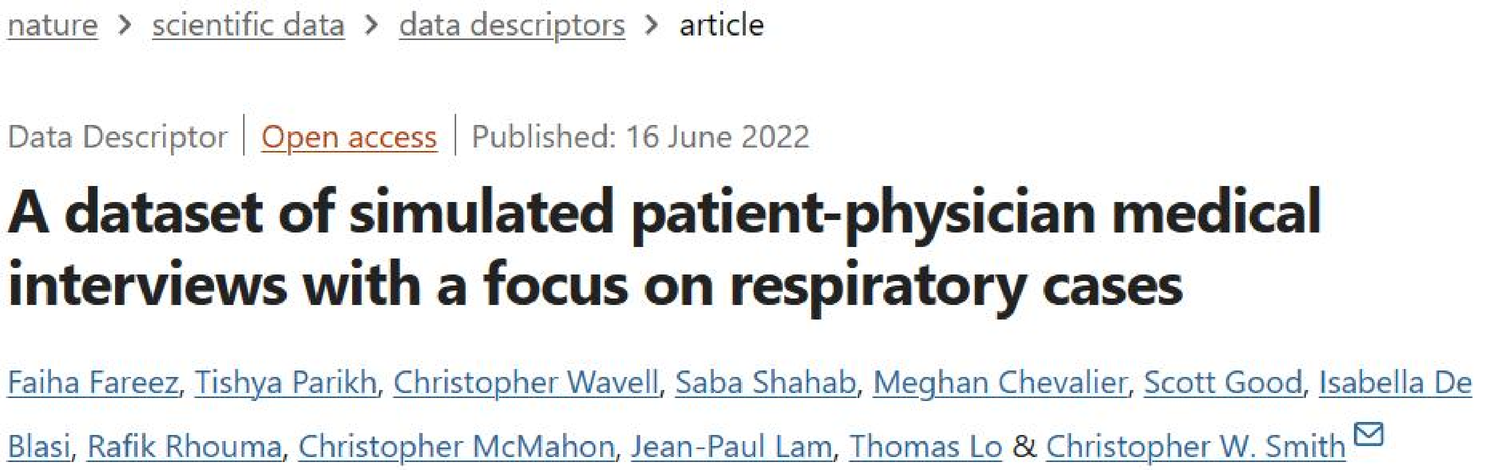
Previous work

Previous approaches adopt various alternatives.

1. Common speech corpus (LibriSpeech, GigaSpeech)
2. Collect before experiments.

This is not a good simulation of the CASR scenario.

Medical Interview (MED-IT)



nature > scientific data > data descriptors > article

Data Descriptor | Open access | Published: 16 June 2022

A dataset of simulated patient-physician medical interviews with a focus on respiratory cases

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, Thomas Lo & Christopher W. Smith

Original recordings

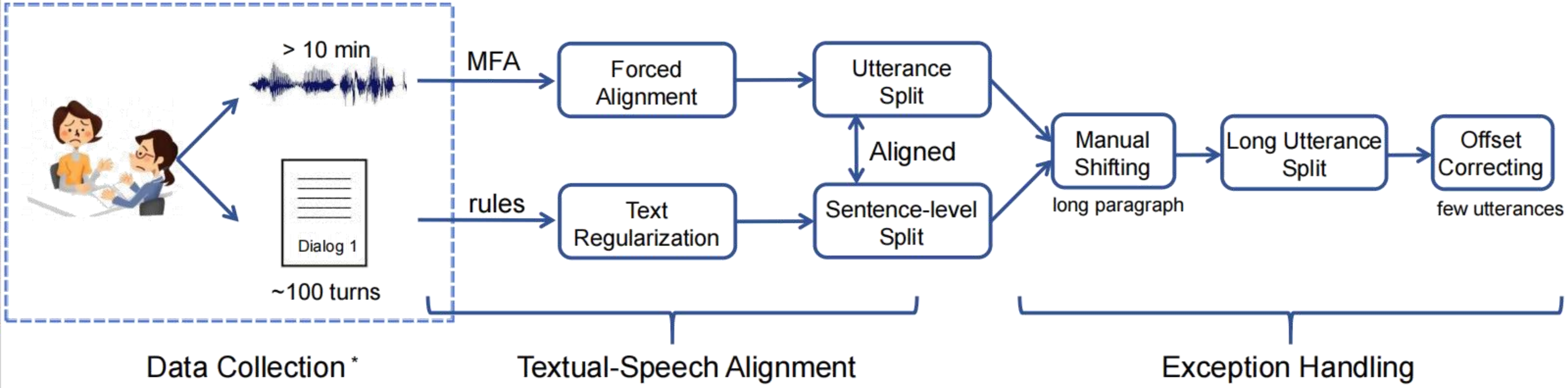
Clinical consultation speech recordings provided by previous work.

There are problems to be solved before adopting in the ASR task.

1. Long speech pieces (~10 min)
2. Bad transcription
3. Various non-text annotation
4. Hesitation and pauses

......

Medical Interview (MED-IT)



Recreation pipline

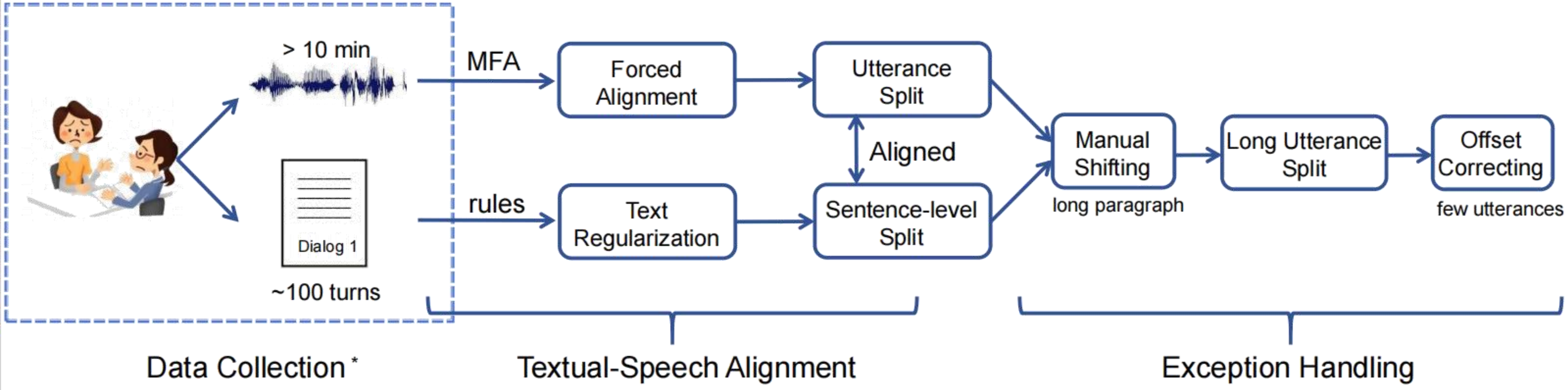
We perform textual-speech forced alignment and exception handling.

In details, three phases of preprocess are performed.

1. Data cleaning: A standarized text transcription.

Umm, Um, Aum,...	→ special token <hesitation>
Digits and abbreviations	→ spoken format
Obvious spelling errors	→ correct transcription

Medical Interview (MED-IT)



Recreation pipline

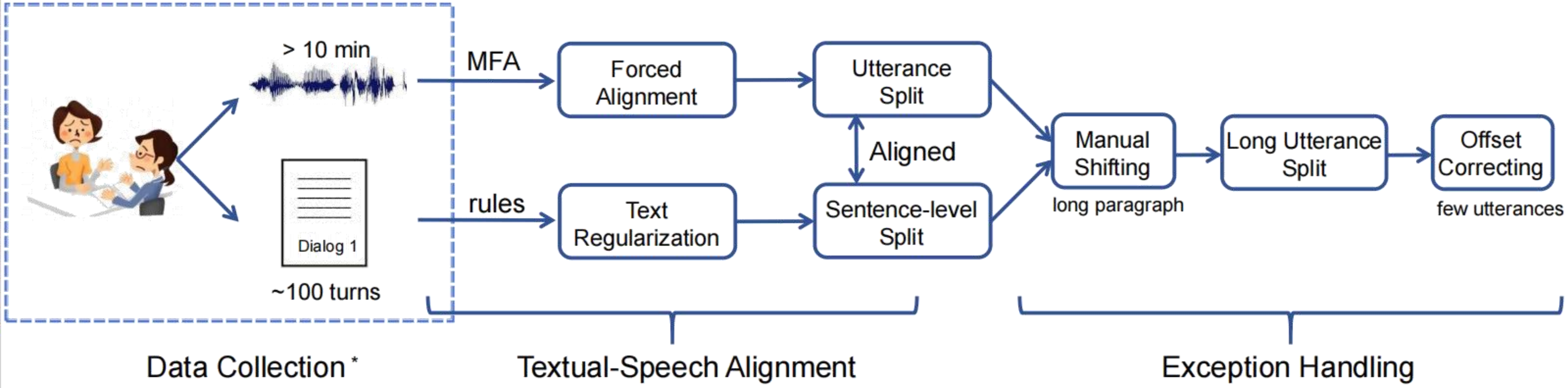
2. Data segmentation: Generate utterance level speech-text alignment for segmentation.

   Montreal Forced Aligner is applied for forced alignment, and then text and radio is segmented sentence-wised.

   For the few instances where significant alignment errors occur, we use Praat to re-segment.

   Long utterance is segmented again into speech pieces within 24s.

Medical Interview (MED-IT)



Recreation pipline

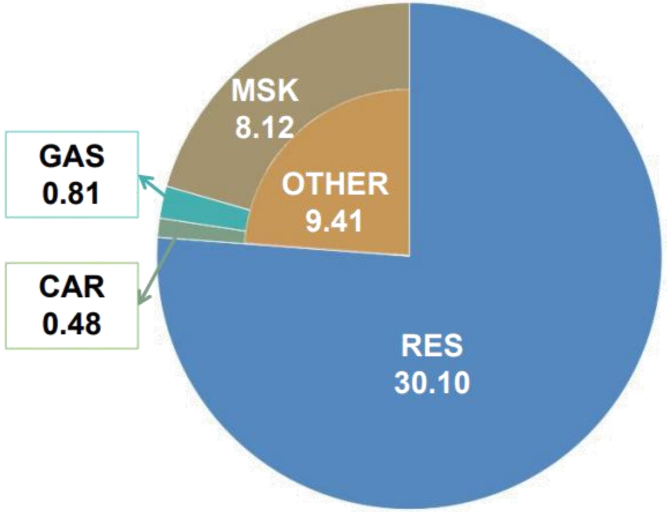
3. Data examination: Regular interval sampling check to prevent short passage offset.

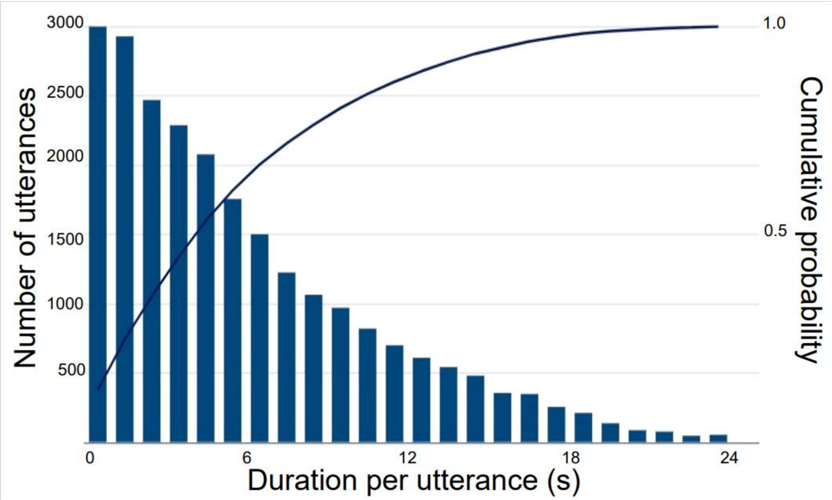
The peformance of MFA is related to the speech quality and environment noise.

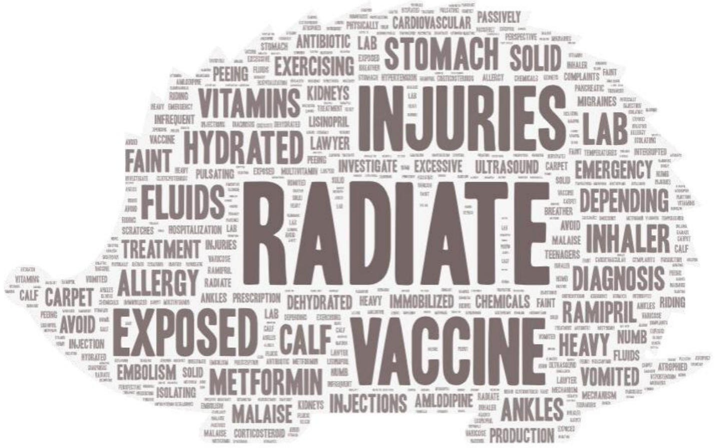
Human sampling check to ensure the quality of MED-IT.

The speech corpus after above process exhibts appropriate durations and precise alignment with the text annotations.

Overview of MED-IT







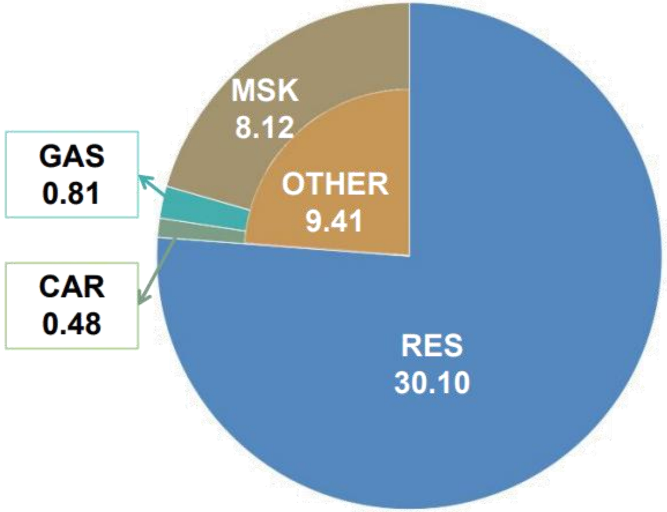
Dataset details

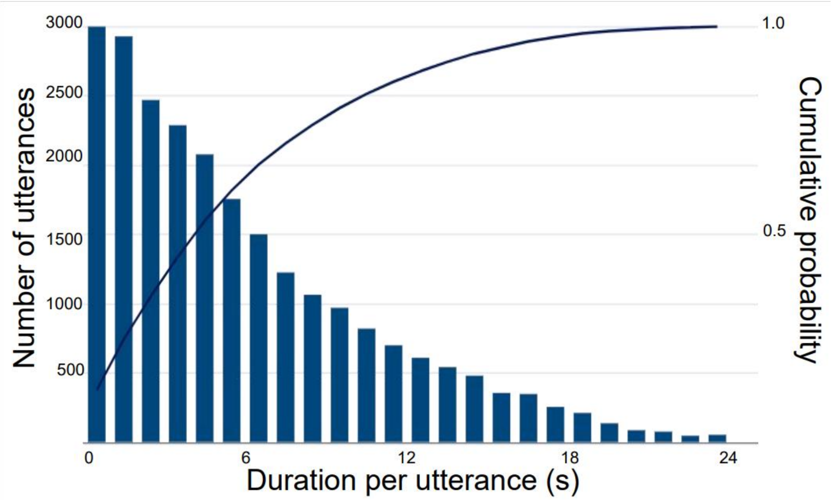
Med-IT encompasses diagnostic and treatment consultant recordings of four departments: Respiratory (RES), Musculoskeletal (MSK), Gastrointestinal (GAS) and Cardiovascular (CAR).

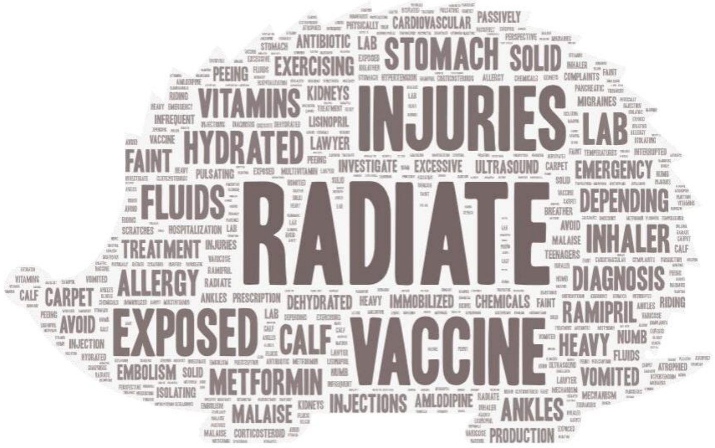
The patient-doctor dialogues are structured according to the Object Structured Clinical Examination.

The doctor and patient are both played by medical students, ensuring no potential issues related to privacy disclosure.

Overview of MED-IT







Dataset details

Most utterances in MED-IT last for less than 10s. The longest one last for about 24s.

The hedgehog-shaped word cloud diagram shows the rare words included in MED-IT, where the size of the word represents the relative frequency of each word in the rare word list.

A large amount of rare words for MED-IT are named entities which are important for downstream tasks like natural language understanding.

MED-IT is a well-designed speech corpus for contextual speech recognition.

Post-decoder Biasing

Intuition

Common words are easily to recognize, for their occurance in the training set is enough for the model to learn a good feature representation.

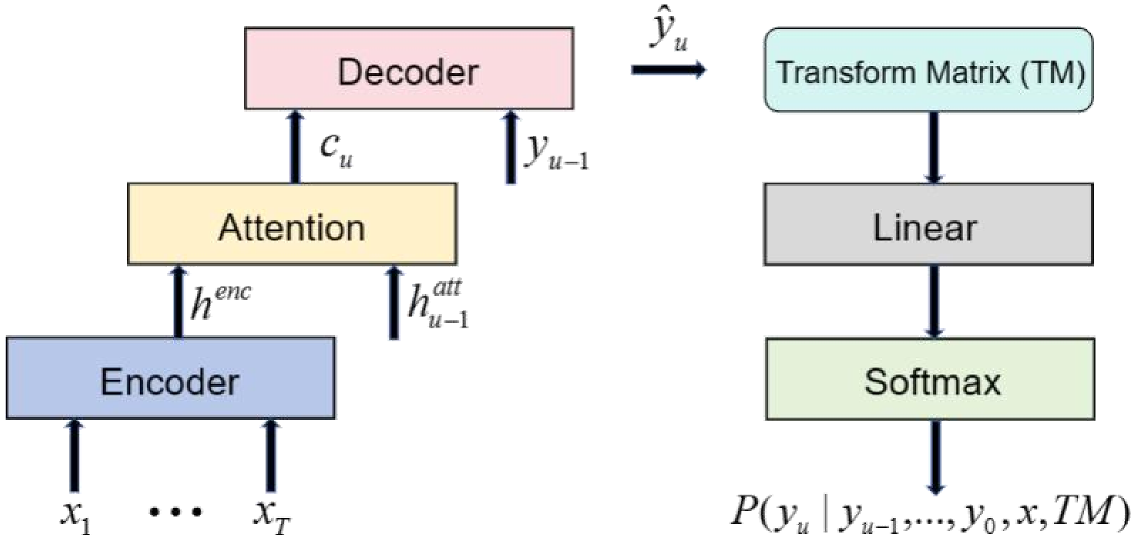
When recognizing common words, the end-to-end model is likely to produce a large probability. But when recognizing rare words, the model tends to produce a low probability, so the correct rare words are often wrongly recognized as other common words.

When one top-k hypothesis contains rare words, you should pay more attention to this sentence.

Can we integrate this idea into the end-to-end model?

When the rare word path score that should output extremely low probability exceeds a certain threshold, the replacement of non-rare words to rare words is completed.

Model Architecture



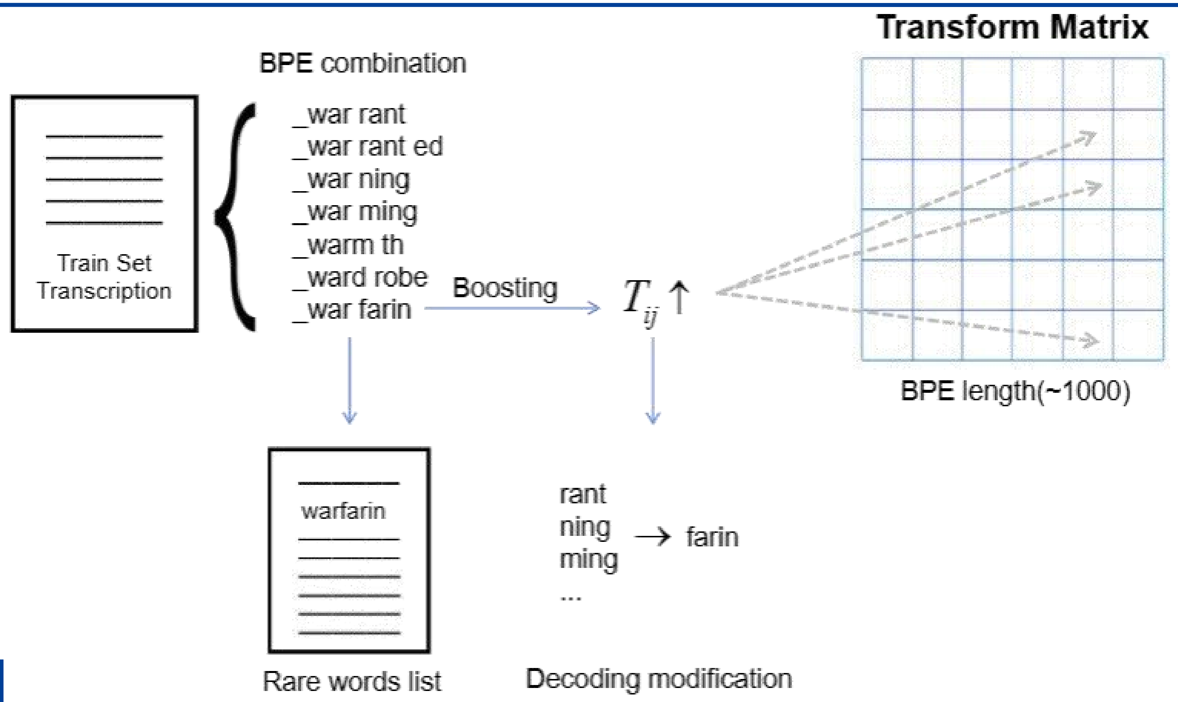
LAS

The LAS model produce unbiased probability sequence y_u. y_u can be viewed as the approximate confidence score of the model output.

A transform matrix that favors rare words replaces non-rare words if and only if the confidence of non-rare words is not high enough and the probability of rare words is relatively high.

So the problem is how to design a proper transform matrix.

Transform Matrix



BPE level

Word-level replacement seems more reasonable, but the computation cost is unbearable for ASR.

The transform matrix and recognition process is performed on sub-word level, the BPE level.

If rare word ‘warfarin’ (_war farin) is wrongly recognized as “warning” (_war ning), the transform matrix should replace “ning” with “farin”.

Unbiased Transform Matrix

Calculation

The BPEs generated by Sentencepiece contains two types, with ‘_’ and without ‘_’.

The unbiased transform matrix ultilizes the simplest stastical probability for the two types respectively.

1) BPE connection probability

For BPEs with ‘_’:

$$p_i^j = \frac{n_{ij}}{\sum_k n_{ik}}$$

For BPEs without ‘_’:

$$p_i^j = \frac{n_{ji}}{\sum_k n_{ki}}$$

n_{ij} refers to the training frequency of BPE connection $b_i b_j$ and p_i^j refers to the probability of BPE b_i is connected with another type BPE b_j.

Unbiased Transform Matrix

2) BPE replacement probability

For BPE b_i, assuming the overall replacement probability is p_i.

The transform matrix value modeling the BPE replacement probability is

$$T_{ij} = p_i \sum\nolimits_k p_i^k p_k^j \qquad (i \neq k, k \neq j, i \neq j)$$

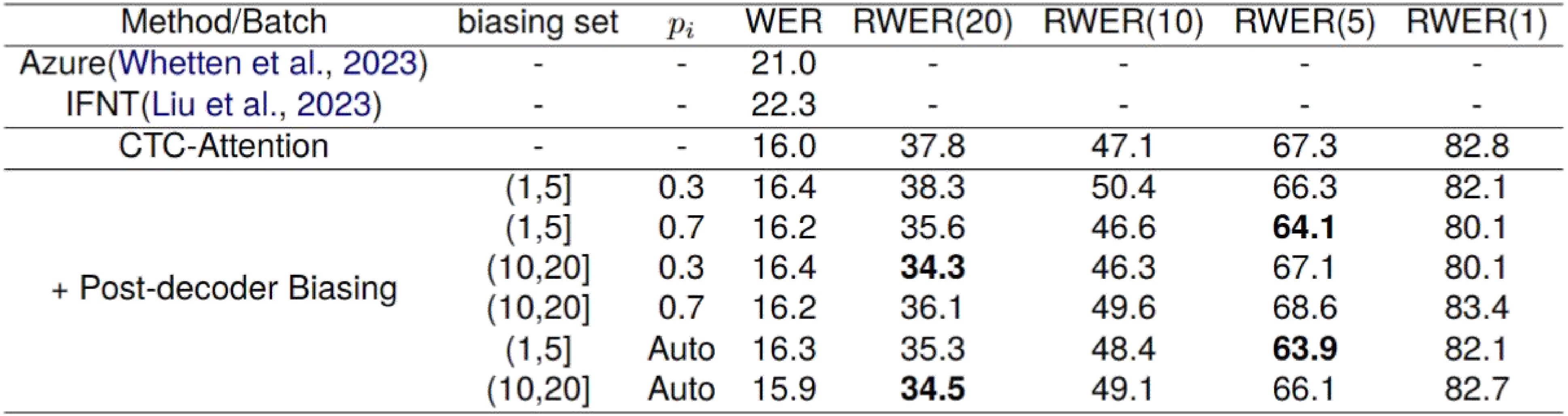
Parameter p_i

This parameter can be regarded as a hyperparameter in the experiment, or can be adjusted based on the training set word frequency.

Biased matrix

The simplest way to construct a transform matrix favoring rare words is to increase the frequency of rare words by a specific quantity (100) when calculating the matrix, thereby increasing the value of the corresponding matrix.

Experiment results



Method/Batch	biasing set	p_i	WER	RWER(20)	RWER(10)	RWER(5)	RWER(1)
Azure(Whetten et al., 2023)	-	-	21.0	-	-	-	-
IFNT(Liu et al., 2023)	-	-	22.3	-	-	-	-
CTC-Attention	-	-	16.0	37.8	47.1	67.3	82.8
+ Post-decoder Biasing	(1,5]	0.3	16.4	38.3	50.4	66.3	82.1
	(1,5]	0.7	16.2	35.6	46.6	**64.1**	80.1
	(10,20]	0.3	16.4	**34.3**	46.3	67.1	80.1
	(10,20]	0.7	16.2	36.1	49.6	68.6	83.4
	(1,5]	Auto	16.3	35.3	48.4	**63.9**	82.1
	(10,20]	Auto	15.9	**34.5**	49.1	66.1	82.7

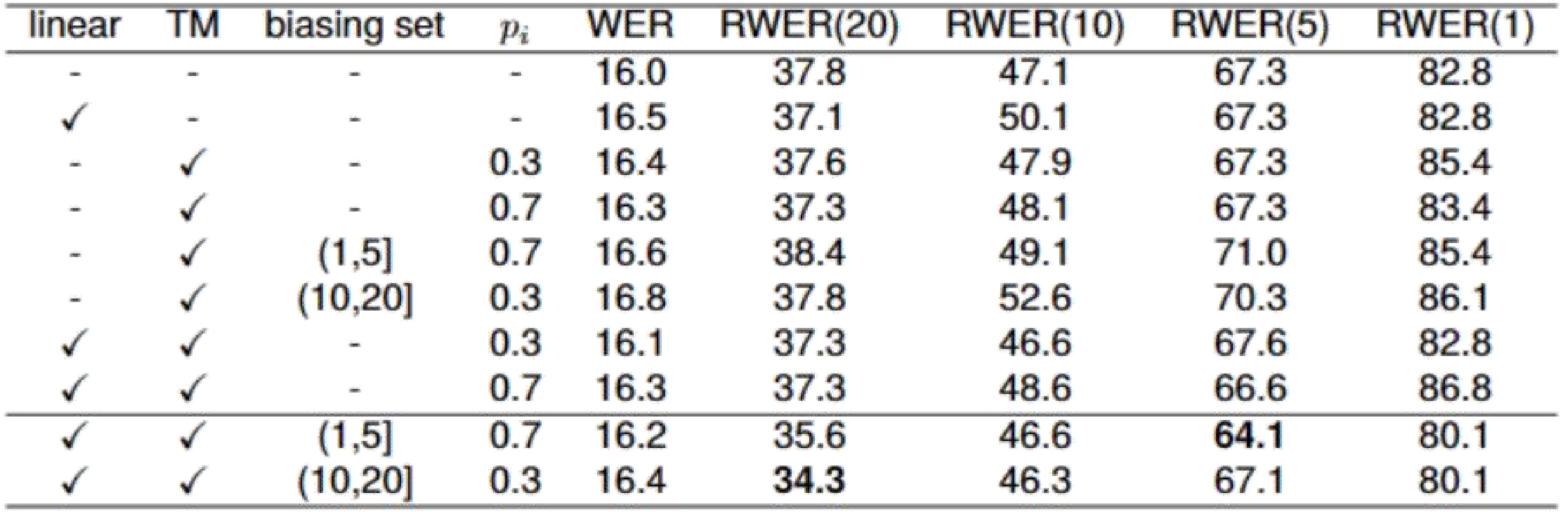
Overall performance

Our model performs better than previous speech recognition results on this dataset, demonstrating the effectiveness of our data processing.

CASR performance

For subsets of rare words appearing in the training speech between 10 and 20 times, and between 1 and 5 times, the proposed method achieves a relative improvement of 9.3% and 5.1%, respectively.

Ablation Study



linear	TM	biasing set	p_i	WER	RWER(20)	RWER(10)	RWER(5)	RWER(1)
-	-	-	-	16.0	37.8	47.1	67.3	82.8
✓	-	-	-	16.5	37.1	50.1	67.3	82.8
-	✓	-	0.3	16.4	37.6	47.9	67.3	85.4
-	✓	-	0.7	16.3	37.3	48.1	67.3	83.4
-	✓	(1,5]	0.7	16.6	38.4	49.1	71.0	85.4
-	✓	(10,20]	0.3	16.8	37.8	52.6	70.3	86.1
✓	✓	-	0.3	16.1	37.3	46.6	67.6	82.8
✓	✓	-	0.7	16.3	37.3	48.6	66.6	86.8
✓	✓	(1,5]	0.7	16.2	35.6	46.6	**64.1**	80.1
✓	✓	(10,20]	0.3	16.4	**34.3**	46.3	67.1	80.1

Ablation findings

The transform matrix and and linear layer are indispensable, and recognition improvement of biasing set can only be accomplished when the probability of rare words is improved.

We believe that the biased transform matrix changes the distribution tendency of the model's output, but its quantification is achieved through the linear layer.

Without rare words boosting, post-decoder shows a relatively small inpact on the overall recognition performance.

Contribution and Future work

Contribution

We are dedicated to medical consultations and segment authentic speech from four departments to build a knowledge-intensive speech corpus MED-IT, which will benefit the CASR research.

We propose a novel lightweight and portable scheme called post-decoder biasing to enhance the recognition of rare words.

Post-decoder biasing can be applied to E2E models without significant computational cost or latency. It does not change the paradigm of end-to-end optimization.

Future work

We use estimations during the construction of transform matrix, which is obviously not the optimal solution.

We haven’t consider the acoustic similarities, but words with similar pronunciation are often confused.

Post-decoder biasing is just a preliminary approach to inspire more in-depth research.

Thanks for listening.