Personalization for BERT-based Discriminative Speech Recognition Rescoring

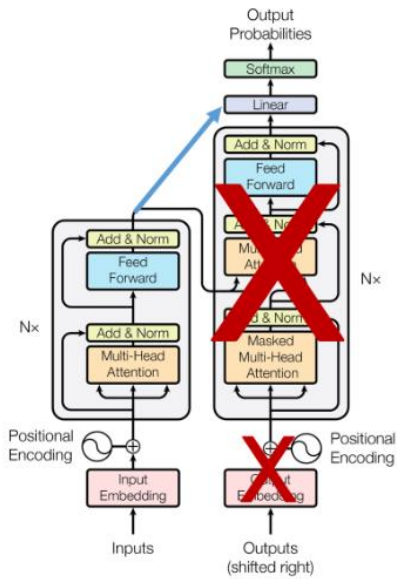
Amazon Alexa(Jari et al)

surim-lab

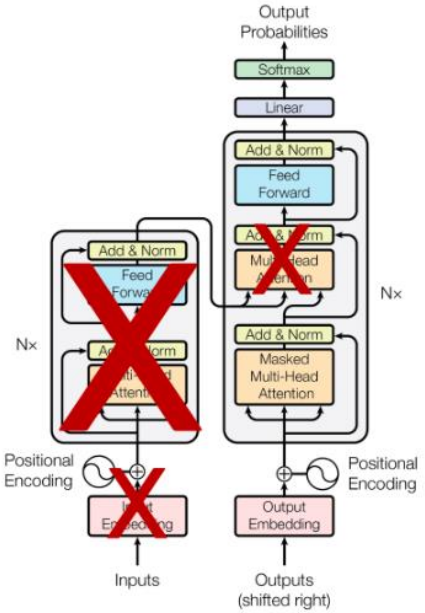
I. RescoreBERT

(1) BERT VS GPT, RescoreBERT

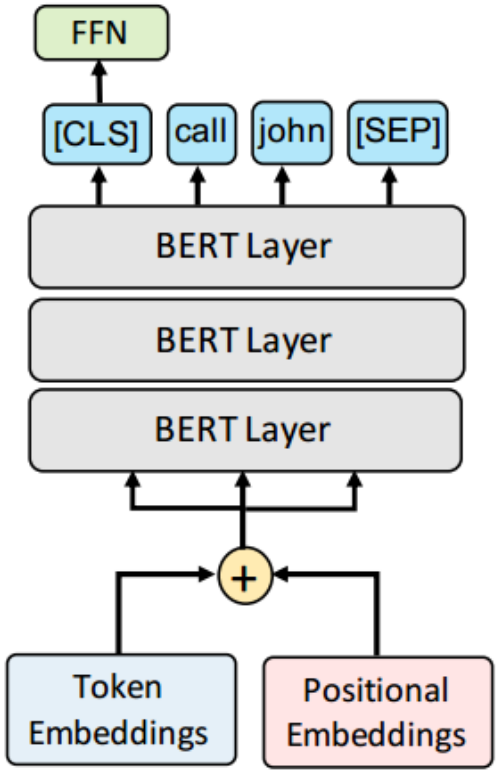
$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_\mathbf{K}}})\mathbf{V}$$



BERT
(Bidirectional Encoder Representations from **Transformer,** masked LM)



GPT
(Generative Pretrained **Transformer**, unidirectional)



RescoreBERT

Transformer

- Encoder: compress source sequence and send
- Decoder: create target sequence
- Self Attention: learning process of contextual relationship using Q, K, V -> Multi-head

Rescoring

- 1st pass in ASR: create a list of hypotheses from acoustic information
- 2nd : Rescore to identify likely transcription
- Final score: 20*1st + 2nd $v_i = \alpha u_i + \beta s_i$

RescoreBERT

- Bidirectional Transformer Rescoring Model
- Predicts a single score from hypothesis to minimize WER trained with discriminative ASR loss,
- FFN attached to CLS token

II. ER in ASR

(1)CSID, Edit Distance, WER, CER, PER, WERR

- **CSID: Correct, Substitution, Insertion, Deletion**
- **ER = Edit Distance / N(words in Reference) = min(S + D + I) / (S + D + C)**

  = {WER(Word), CER(Character), PER(Phoneme), TER(Token), SER(Sentence)…}

  *** WERR: WER Reduction (+ : WER ↑ ←→ - : WER ↓ (performance ↑))**

Model	Personalized	General
Oracle	-57%	-58%
Tiny RescoreBERT	+3.9%	-5.3%
Big RescoreBERT	+4.8%	-7.1%
Tiny RescoreBERT (fine-tuned)	+2.5%	-5.3%
Big RescoreBERT (fine-tuned)	+1.7%	-6.8%

MEDIAZEN

II. ER in ASR

(2) CSID Algorithm, Example(1) snow VS sunny

* Direction

↘ +0 : Correct
↘ +1 : Substitution
↓ +1 : Deletion
→ +1 : Insertion

For calculate Edit Distance

* CODE

```
arr[i][j] = min(
        arr[i-1][j]+1, # D
        arr[i][j 1]+1, # I
        arr[i-1][j-1] + cost)
        # S(cost=1) or C(cost=0)
```

		S	U	N	N	Y
	0	1	2	3	4	5
S	1					
N	2					
O	3					
W	4					

(2,2): min(1+1(D), 1+1(I), 0+0(C)) =0

		S	U	N	N	Y
	0	1	2	3	4	5
S	1	0	1	2	3	4
N	2					
O	3					
W	4					

C I

		S	U	N	N	Y
	0	1	2	3	4	5
S	1	0	1	2	3	4
N	2	1	1	1	2	3
O	3					
W	4					

C I C

		S	U	N	N	Y
	0	1	2	3	4	5
S	1	0	1	2	3	4
N	2	1	1	1	2	3
O	3	2	2	2	2	3
W	4					

C I C S

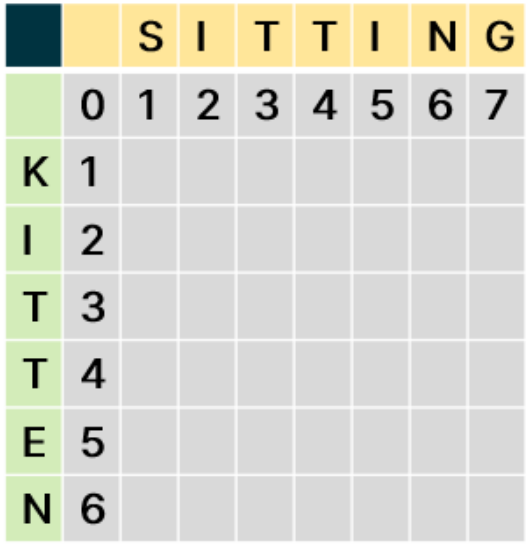
		S	U	N	N	Y
	0	1	2	3	4	5
S	1	0	1	2	3	4
N	2	1	1	1	2	3
O	3	2	2	2	2	3
W	4	3	3	3	3	3

CICSS → 3(not C)

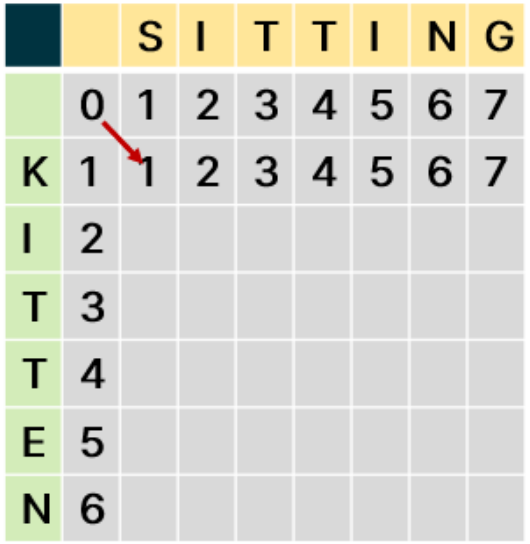
MEDIAZEN

II. ER in ASR

(2) Example(2): kitten VS sitting

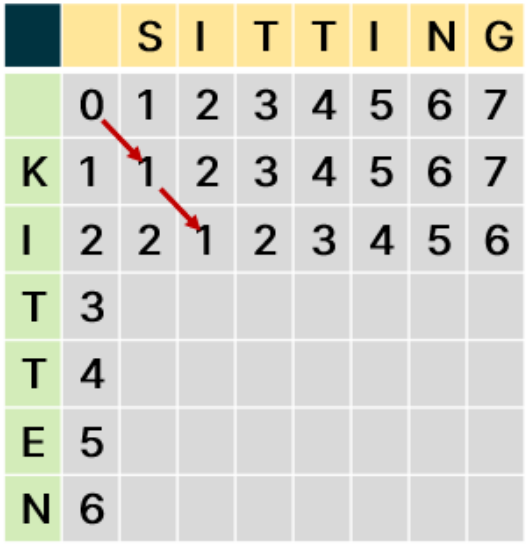


		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1							
I	2							
T	3							
T	4							
E	5							
N	6							



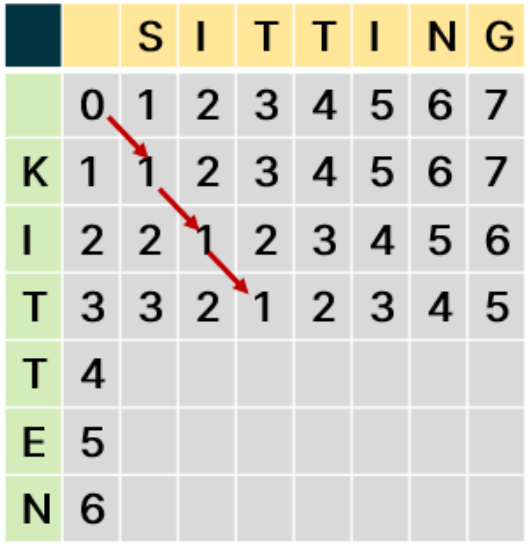
		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2							
T	3							
T	4							
E	5							
N	6							

S



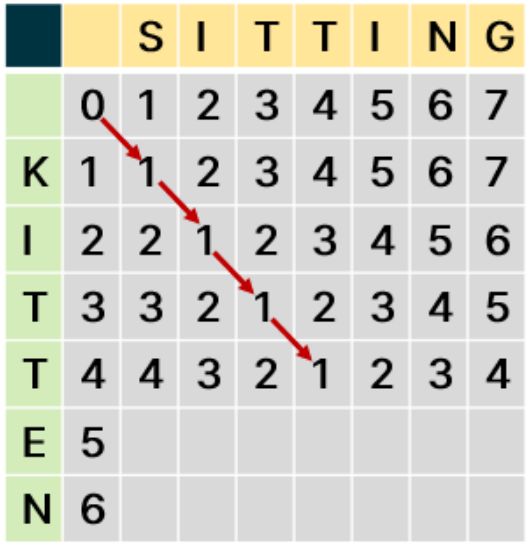
		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3							
T	4							
E	5							
N	6							

S C



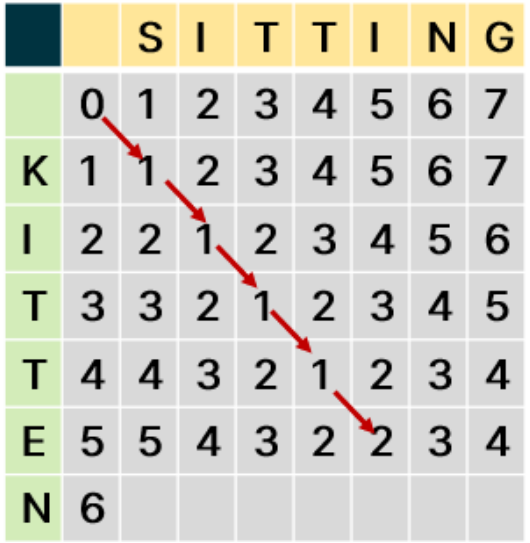
		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3	3	2	1	2	3	4	5
T	4							
E	5							
N	6							

S C C



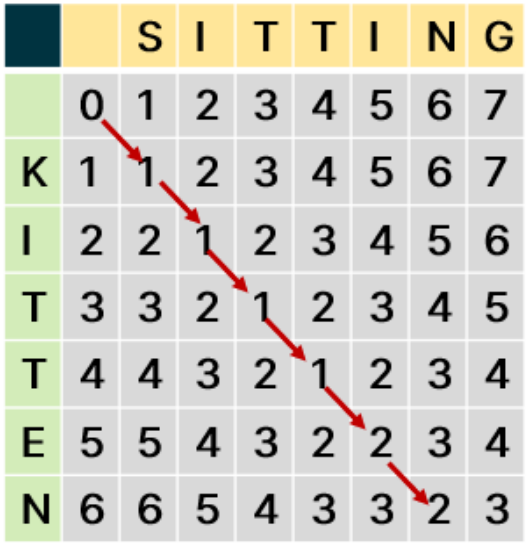
		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3	3	2	1	2	3	4	5
T	4	4	3	2	1	2	3	4
E	5							
N	6							

S C C C



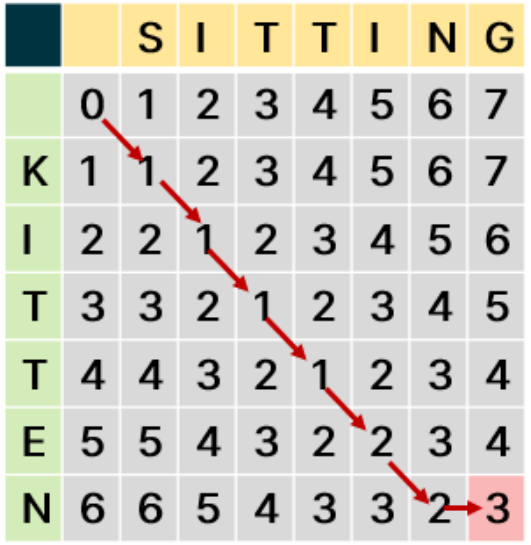
		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3	3	2	1	2	3	4	5
T	4	4	3	2	1	2	3	4
E	5	5	4	3	2	2	3	4
N	6							

S C C C S



		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3	3	2	1	2	3	4	5
T	4	4	3	2	1	2	3	4
E	5	5	4	3	2	2	3	4
N	6	6	5	4	3	3	2	3

S C C C S C

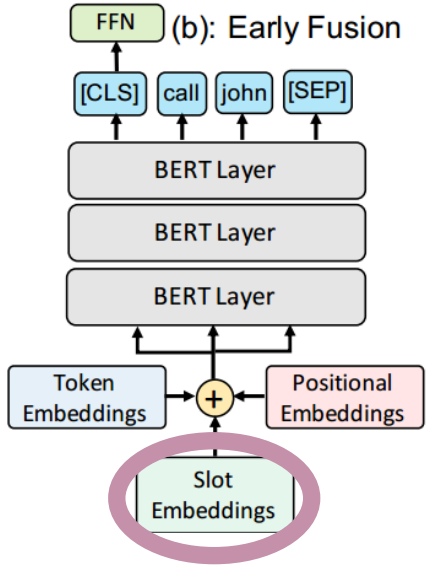


		S	I	T	T	I	N	G
	0	1	2	3	4	5	6	7
K	1	1	2	3	4	5	6	7
I	2	2	1	2	3	4	5	6
T	3	3	2	1	2	3	4	5
T	4	4	3	2	1	2	3	4
E	5	5	4	3	2	2	3	4
N	6	6	5	4	3	3	2	3

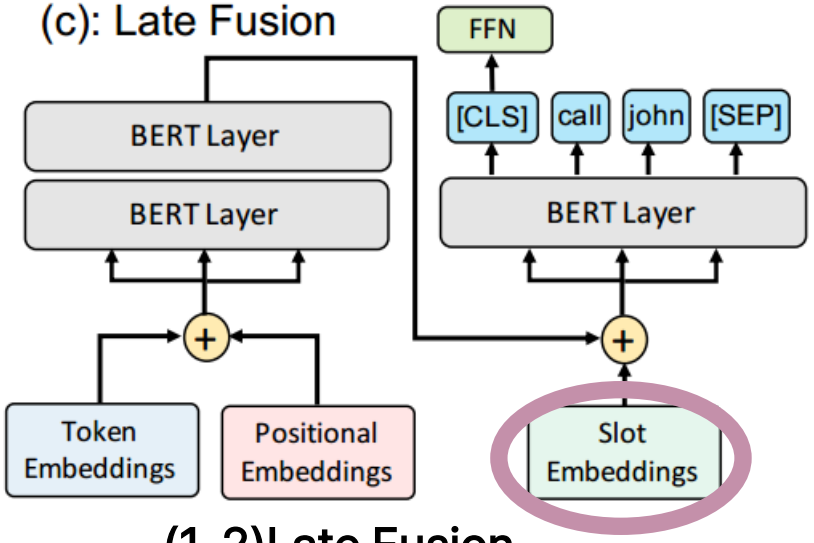
S C C C S C I → 3(not C)

MEDIAZEN

III. Personalized Approaches



(1-1)Early Fusion



(1-2)Late Fusion

(1) Gazetteers

- Original: $s_i = f \circ g(E_t(x_i) + E_p).$

➔ f(FFN), g(BERT layer), x_i = hypothesis token, $E_t(x_i)$=TE, E_p=PE

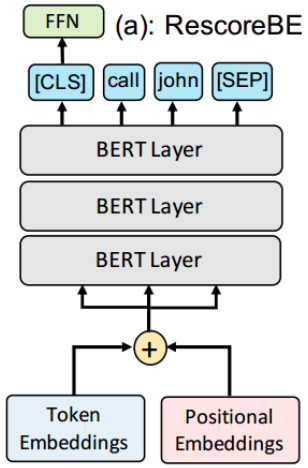
* Ex of slot: theme of movie, title of song

- Gazetteers: Token + Positional + Slot Embedding
  - Early: $s_i = f \circ g(E_t(x_i) + E_p + E_s(y_i)),$
  - Late : $s_i = f \circ g_n\left(E_s + g_{n-1} \circ \cdots \circ g_1(E_t + E_p)\right).$

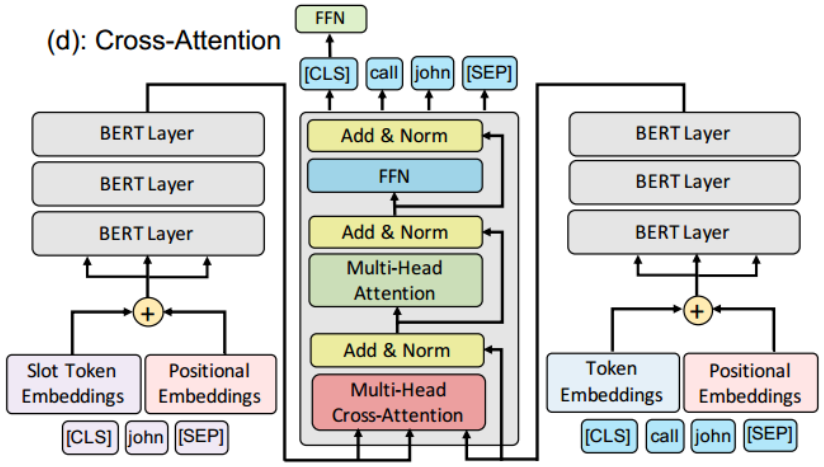
MEDIAZEN

III. Personalized Approaches

(2) NL prompting, (3)Cross Attention based encoder-decoder model



(0)RescoreBERT(baseline)



(3)Cross Attention based encoder-decoder model

(2) Natural Language Prompting

Match Condition: sentence includes one in tokenized string set D

Ex) I want you to call Felix(call [entity])

Prompt: original sentence + simple phrase(augmented prompt)

Ex) I want you call to Felix as I need to contact Felix(as I need to contact [Entity])

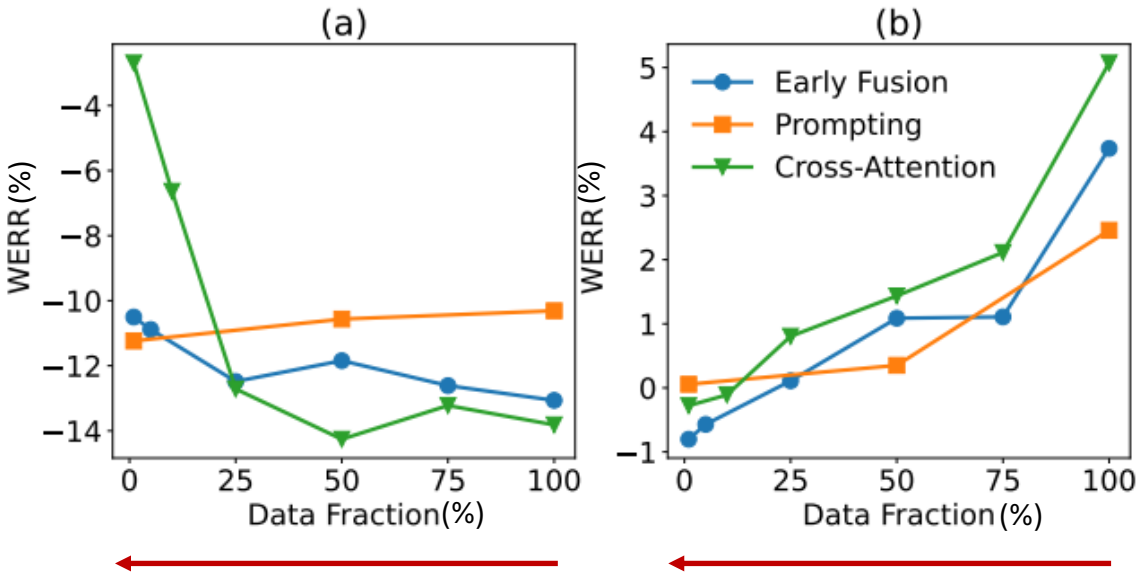
(3) Cross Attention based encoder-decoder model

z_i: slot token sequence

ex) [CLS] [entity #1] [SEP] [entity #2] [SEP] ⋯ [SEP].

$$X_i = g(E_t(x_i) + E_p)$$
$$Z_i = g(E_t(z_i) + E_p)$$
$$s_i = f \circ m(Z_i, X_i),$$

IV. Result of Experiments



Explanation

- Data: from Alexa with personalized NE
- Object: WER ↓ in personalized content
- X axis(personalized entity data)
- Progress: mixed general data (opp)
- L/R: Personalized entity / General test set

(1) NL Prompting

- Moderate performance between (2) and (3)
- WERR -7% without training with a marginal loss in generalization(useful when train data ↓)

(2) Cross Attention

- Sensitive to the Data Fraction(rapid rise in L)
- Degradation in General set

(3) Gazetteers(Early Fusion)

- L: <= -10 % WERR
- R: the only approach with WER reduction in general data

Thank you

rabit331@naver.com

2024.01.26