Adapting an Unadaptable ASR System

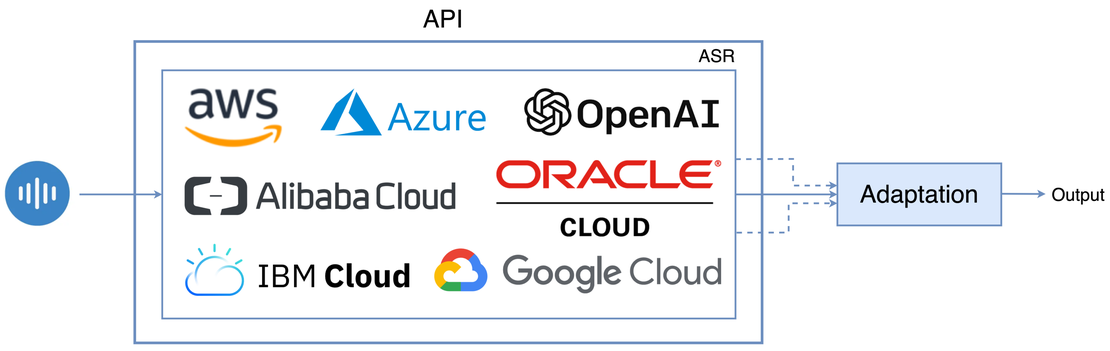
Mengjie Qian*, Rao Ma*, Mark J. F. Gales, Kate M. Knill
{mq227,rm2114,mjfg100,kmk1001}@eng.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

ALTA Institute / Department of Engineering, University of Cambridge

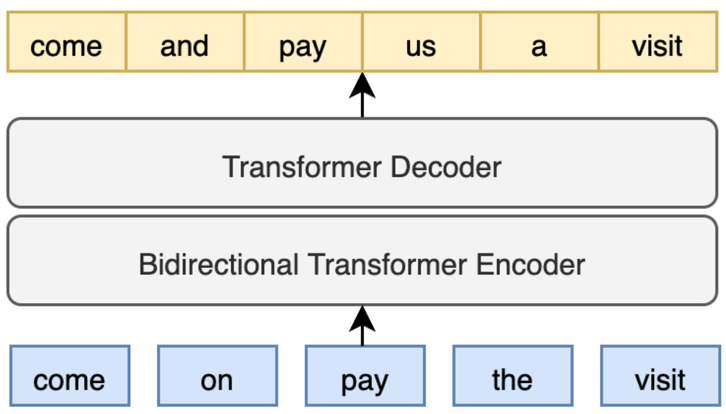
1. Introduction

- When integrating an ASR module, it is common to call APIs from online service providers rather than building the ASR model from scratch.
  - it remains challenging to adapt systems to a specific target domain
  - we utilise the Whisper model as an example to evaluate adaptation methods when the ASR model is not accessible

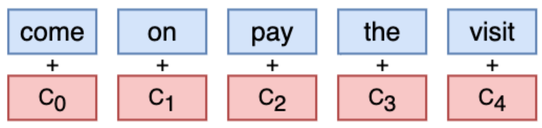


2. Error Correction Model and Variants

- E2E structure taking the ASR transcription as input and is trained to generate the corrected sentence
  - no need to access the ASR system
  - applicable to adapting a black-box, cloud-based speech-to-text system



- In addition to text transcriptions, other outputs can be returned by the ASR service providers
  - extended inputs to the error correction model to improve robustness



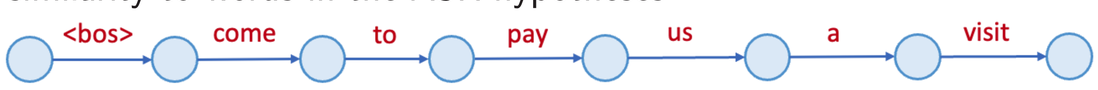
(a) Extended input with ASR confidence score embedding.

(b) Extended input with phone sequence.

(c) Extended input with ASR N-best hypotheses.

3. Unconstrained vs. N-best Constrained Decoding

- Unconstrained decoding
  - the error correction model is decoded with beam search in an unconstrained decoding space
  - the model might output synonyms that have a high embedding similarity to words in the ASR hypotheses



- N-best constrained decoding
  - force the decoding result to appear in the ASR N-best list
  - scores of the ASR model and the error correction model can be linearly combined for better performance

ASR N-best Hypotheses	ASR Score	EC Score
come on pay the visit	-0.1	-1.0
come and pay a visit	-0.3	-0.8
come on pay us a visit	-0.4	-0.5

4. Experiments

- Experimental setup
  - Whisper small.en model
  - Librispeech dataset for training the error correction model
  - PLM for error correction model: T5-base, BART-base
- Comparison of LM rescoring and error correction

Model	Method	LB_clean	LB_other
Baseline	-	3.52	7.37
GPT-2	Rescoring	4.17	7.38
LSTMLM	Rescoring	4.11	7.23
BART	Correction	3.34	7.07
T5	Correction	**3.16**	**7.03**

- Error correction results using various input features

Model	LB_clean	LB_other
1-best	3.16	7.03
+ confidence	3.11	7.04
+ phones	3.21	7.05
10-best uncon	**2.90**	**6.39**
10-best constr	3.10	6.69

- Generalisation of the approach
  - Error correction results on other datasets by models trained on LibriSpeech in the transfer setting

Model	LibriSpeech clean	LibriSpeech other	Other sets TED	Other sets Artie	Other sets MGB
ASR Baseline	3.52	7.37	3.89	9.03	13.10
1-best uncon	3.16	7.03	4.78	9.27	17.29
10-best uncon	**2.90**	**6.39**	4.56	9.16	22.88
10-best constr	3.10	6.69	**3.64**	**8.14**	**12.71**

  - Error correction results for Conformer-Transducer ASR model on LibriSpeech in the transfer setting

Model	LB_clean	LB_other
ASR Baseline	2.79	6.90
1-best uncon	2.78	6.92
10-best uncon	3.86	7.72
10-best constr	**2.62**	**6.65**

- Ablation analysis with disturbed 10-best list

10-best	LibriSpeech clean	LibriSpeech other	Other sets TED	Other sets Artie	Other sets MGB
Sorted	**2.90**	**6.39**	**3.64**	**8.14**	**12.71**
Randomized	3.31	6.82	3.74	8.50	13.01
Reversed	3.50	7.18	3.75	8.57	12.99

5. Conclusions

- It is possible to adapt an ASR system to a particular domain without direct access to the system itself.
- We used an error correction framework with 1-best and N-best outputs to adapt the ASR system.
- The error correction module trained for a specific domain can be applied to other speech domains, as well as other ASR systems.

6. Reference

[1] A. Radford, J. K. Kim, T. Xu, G. Brockman, C. McLeavey, et al, "Robust speech recognition via large-scale weak supervision", arXiv preprint arXiv:2212.04356, 2022.

[2] R. Ma, M. J. Gales, K. Knill, and M. Qian, "N-best T5: Robust ASR error correction using multiple input hypotheses and constrained decoding space", arXiv preprint arXiv:2303.00456, 2023.

[3] R. Ma, M. Qian, M. J. Gales, and K. Knill, "Adapting an Unadaptable ASR System", arXiv preprint arXiv:2306.01208, 2023.

This paper reports on research supported by Cambridge University Press & Assessment, a department of The Chancellor, Masters, and Scholars of the University of Cambridge. Mengjie Qian is supported by EPSRC Project EP/V006223/1 (Multimodal Video Search by Examples).

* Equal Contribution.