N-best T5: Robust ASR Error Correction using Multiple Input Hypotheses and Constrained Decoding Space

UNIVERSITY OF CAMBRIDGE

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

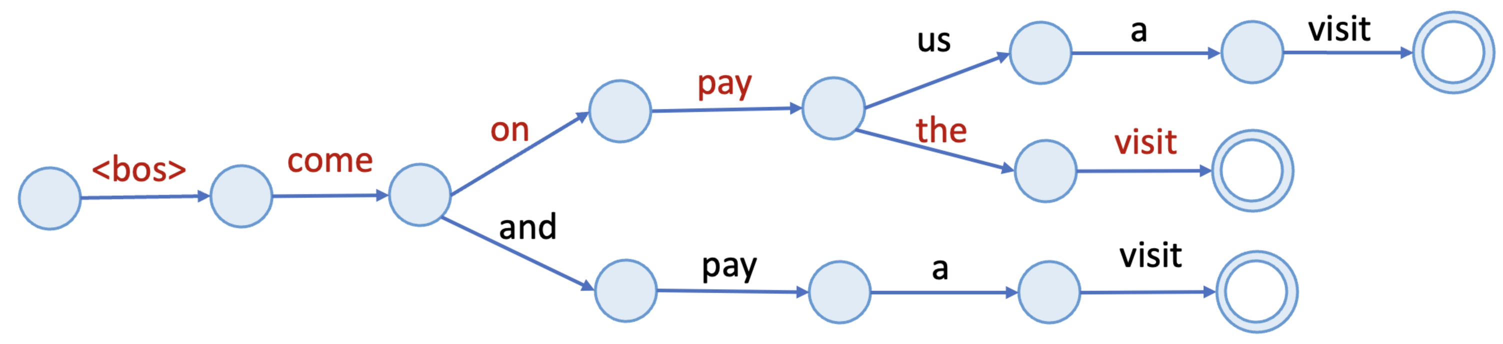
ALTA Institute / Department of Engineering, University of Cambridge

1. Introduction

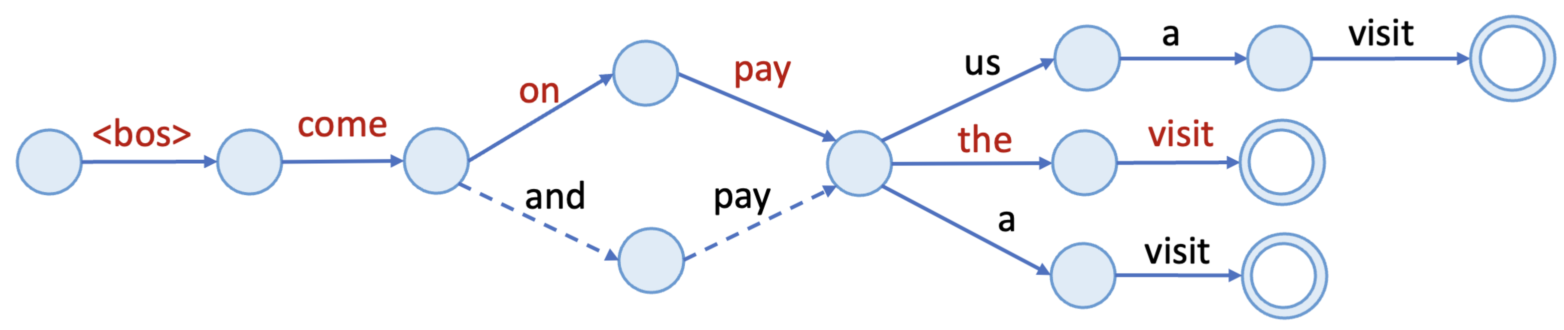
- Most error correction models use the 1-best ASR hypothesis as input and adopt beam search in the decoding.
  - ▷ context within one sentence does not provide enough information
  - ▷ the decoding process of the error correction model is not well-guided
- We propose to use the ASR N-best list as model input and testify several constrained decoding algorithms.

2. Discussion on ASR output

- 1-best vs. N-best list with beam search

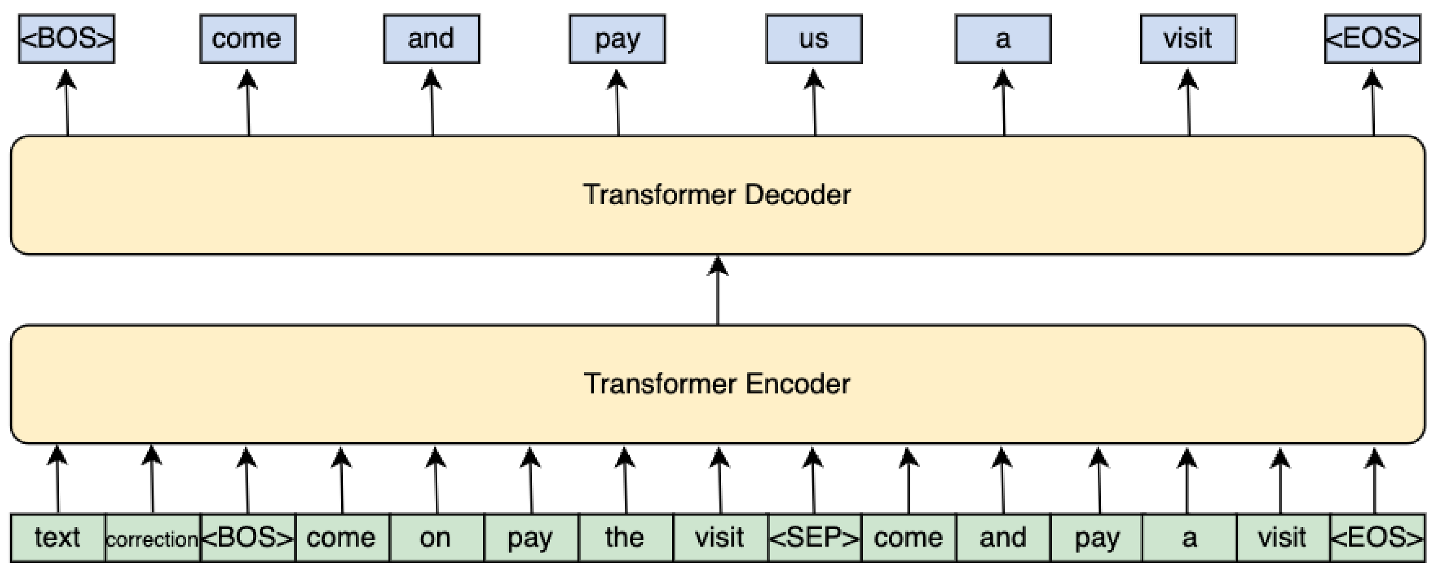


- Word lattice produced in beam search with path merging
  - ▷ merge partial hypotheses with the same N-gram context
  - ▷ the path is removed from the active beam while kept in the lattice



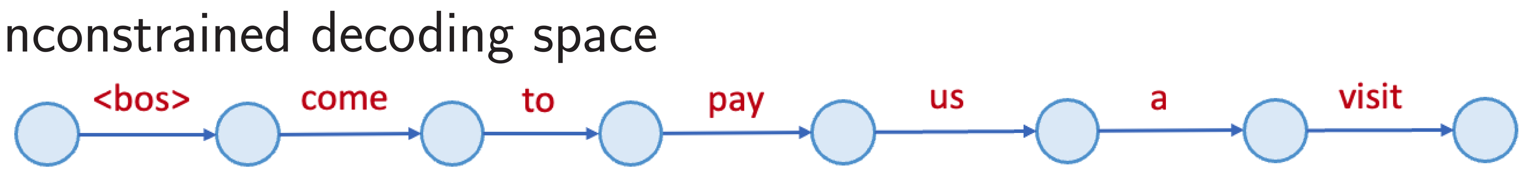
3. N-best T5 Model Structure

- Based on a pre-trained T5 model.
- Concatenate the ASR N-best list, and insert a special token between different hypotheses to denote the sentence end.



4. Constrained Decoding in Inference

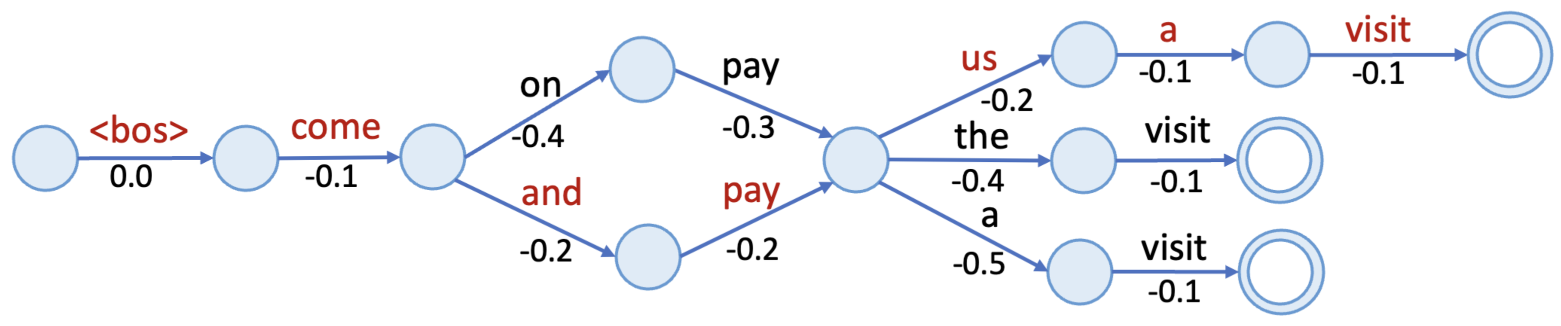
- Unconstrained Decoding
  - ▷ the error correction model is decoded with beam search in an unconstrained decoding space



- N-best Constrained Decoding
  - ▷ force the decoding result to appear in the ASR N-best list

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

- Lattice Constrained Decoding
  - ▷ force the decoding result to appear in the ASR lattice



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).

5. Experiments

- Experimental setup
  - ▷ Data: LibriSpeech, audiobook reading
    - ⋆ Training set: 960hr, SpecAugment
    - ⋆ Test sets: dev_clean, dev_other, test_clean, and test_other
  - ▷ Conformer-Transducer ASR
    - ⋆ Encoder: 12 Conformer layers with a hidden size of 512
    - ⋆ Predictor: 1 LSTM layer
  - ▷ Error Correction Model
    - ⋆ A pre-trained T5 base model
    - ⋆ 6 Transformer blocks for encoder/decoder, hidden dimension: 768
- Oracle WER results in ASR outputs

ASR Output	Dev clean	Dev other	Test clean	Test other
5-best List	1.35	4.72	1.44	4.66
10-best List	1.24	4.43	1.34	4.34
Lattice	0.79	2.98	0.89	3.00

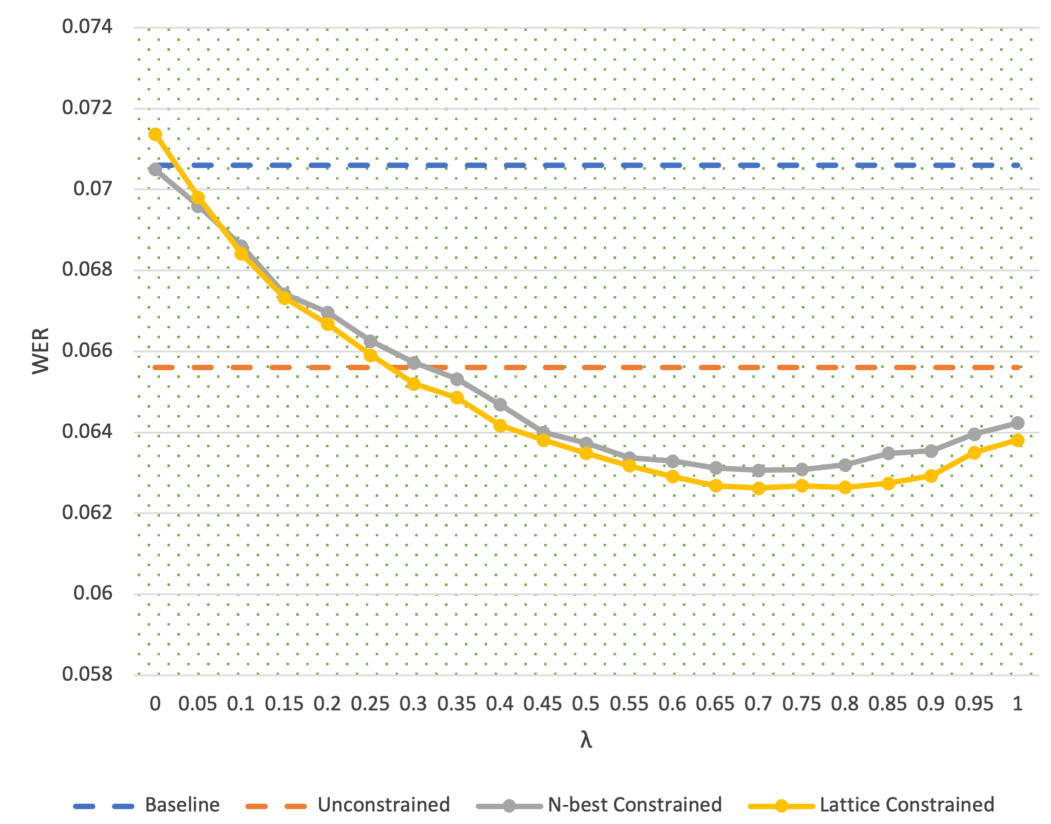
- ASR baseline and error correction results

Model	Dev clean	Dev other	Test clean	Test other
Baseline	2.71	6.99	2.88	7.06
1-best T5	2.89	6.94	3.02	7.18
5-best T5	2.62	6.40	2.77	6.67
10-best T5	**2.60**	**6.25**	**2.67**	**6.56**

- Comparison of Decoding Algorithms

Model	Decoding Method	Dev clean	Dev other	Test clean	Test other
Baseline	-	2.71	6.99	2.88	7.06
5-best T5	Unconstrained	2.62	6.40	2.77	6.67
	N-best Constrained	**2.38**	6.25	2.55	6.38
	Lattice Constrained	2.40	6.21	2.54	6.34
10-best T5	Unconstrained	2.60	6.25	2.67	6.56
	N-best Constrained	2.39	6.17	2.54	6.31
	Lattice Constrained	2.41	**6.11**	**2.53**	**6.27**

  - ▷ The effect of interpolation weight in the constrained decoding



6. Conclusions

- The proposed error correction model consistently improves over the performance of a strong ASR model on the LibriSpeech test sets.
- The first to use ASR N-best list as input to PLMs and constrained decoding algorithms based on the output from E2E ASR models for error correction.

7. Reference

[1] 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.

[2] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," Journal of Machine Learning Research, 2020.