Neural Lexicon Reader: Reduce Pronunciation Errors in End-to-End TTS by Leveraging External Textual Knowledge

Mutian He^, Jingzhou Yang, Lei He, Frank K. Soong*
^The Hong Kong University of Science and Technology, *Microsoft

Code/Model/Sample
mutiann.github.io/papers/nlr

Introduction

Goal: To build fully end-to-end TTS with minimal resources

- Both minimal data and minimal human efforts, incl. linguistic expertise to build G2P pipeline

Challenge: Fully E2E TTS without preprocessed phoneme inputs often produces pronunciation errors

- Especially for non-phonemic scripts (like Chinese) and irregular orthography (like English)

E2E TTS needs to **know** how to pronounce

- …but paired data won't cover all the knowledge
- …and neural networks are inefficient to memorize "hard" knowledge

Lexicons are widely available, but G2P is more than that:

- Polyphones/heteronyms: Pronunciations dependent on contexts
- Require "soft" capability to resolve

Idea: Not to internalize knowledge, but to learn **how to** extract external knowledge directly given to the model in text forms, e.g. texts from dictionary entries

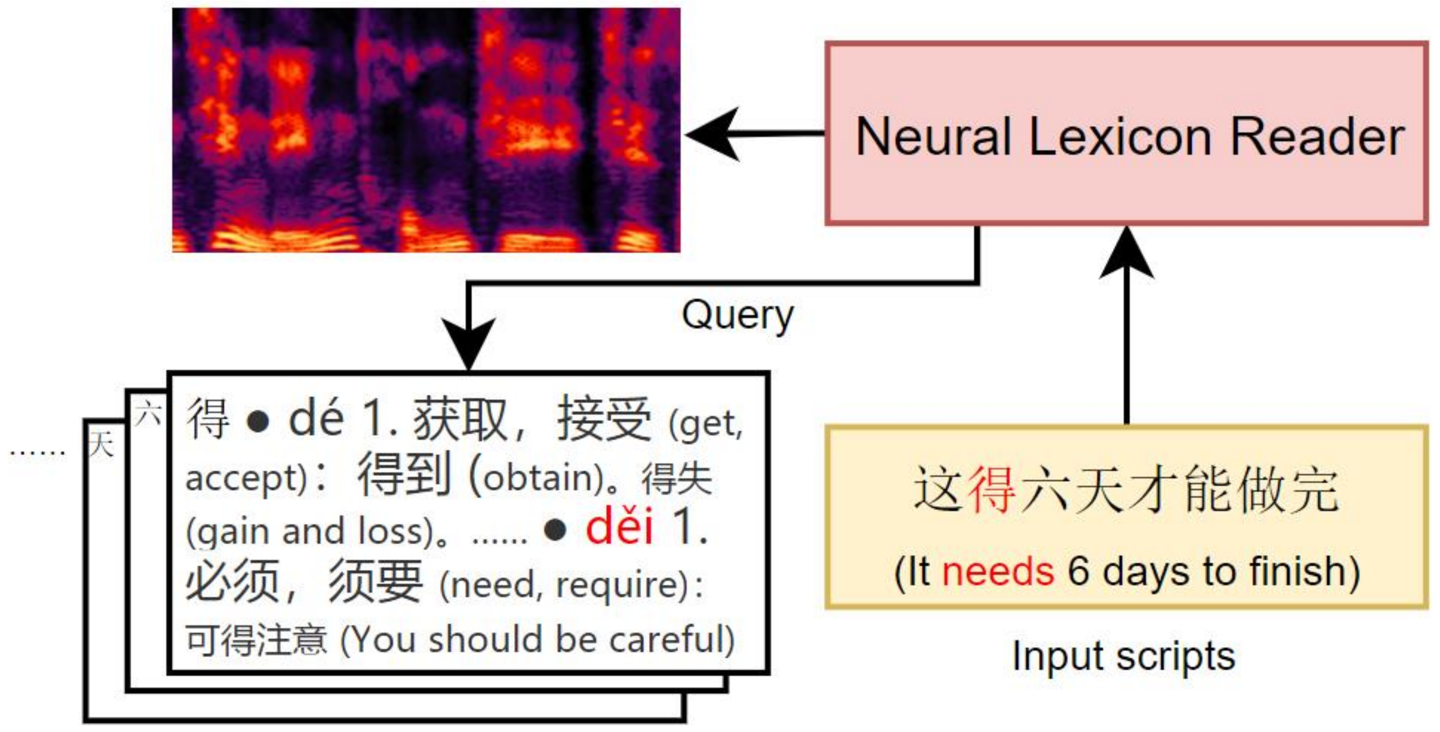
Example: Pronunciation knowledge; Given the human-readable raw dictionary entry of a word in the input script, with readings and explanations of each reading, build a model to match the explanation with the contexts and to extract and pronounce the correct reading

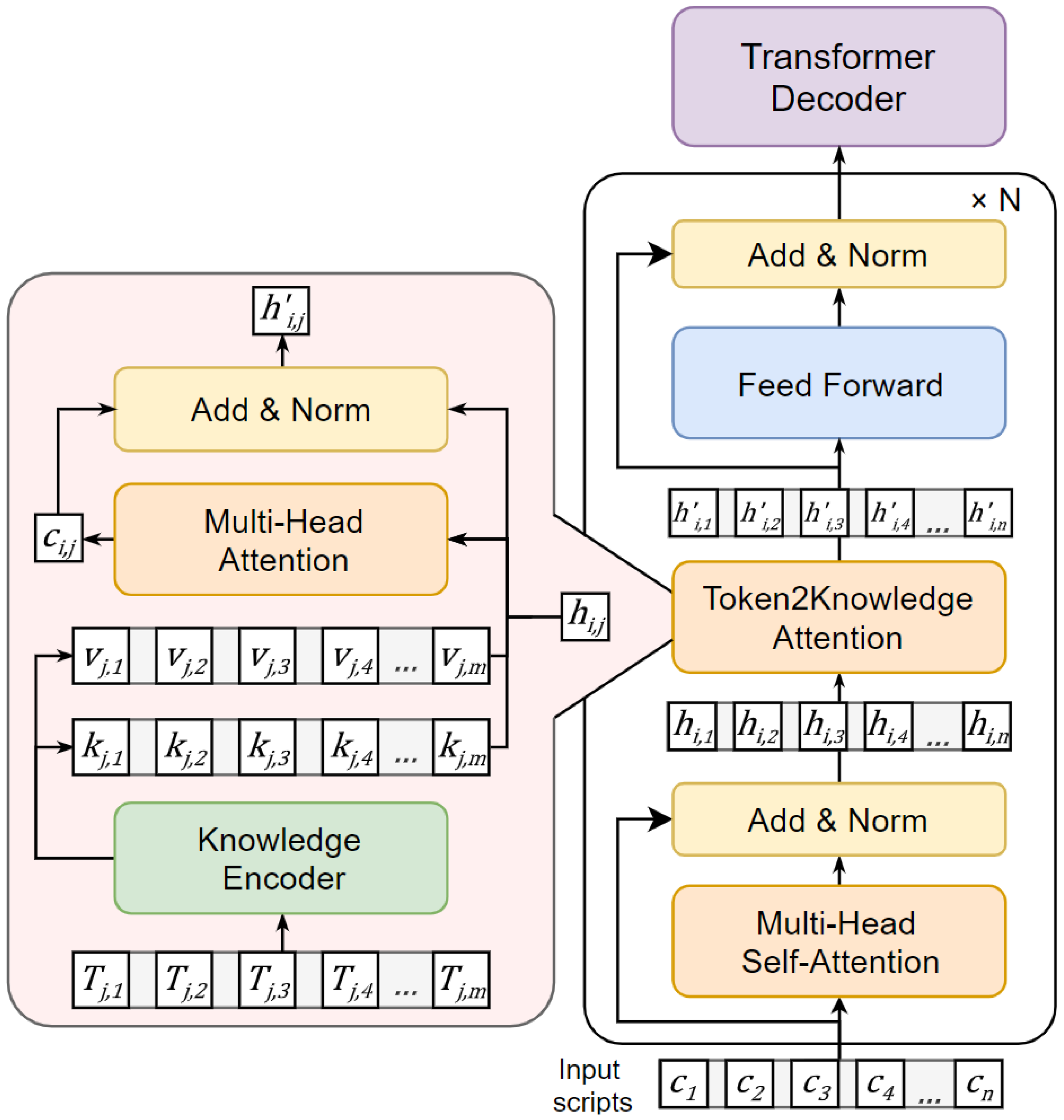
Methods

Transformer TTS, with additional Token2Knowledge Attention in each encoder layer i after self-attention

- Similar to Encoder-Decoder Attention in Decoder
- Difference: Each token attends to its own relevant knowledge. For example, a character c_j attends to its own lexicon entry text T_j
- Texts in lexicon entries encoded by the knowledge encoder, e.g. the pretrained language model XLM-R
- Encoded texts v_j for the pronunciation matching the context $h_{i,j}$ will be extracted through attention

Using online lexicons; pronunciation errors evaluated with subjective character error rates (CER) by human listeners, and objective ones by Azure Speech-To-Text





Experiments

Starting from Mandarin dataset with 18K utterances, down-sampled to different sizes to simulate low-resource conditions, and evaluated on

- **General** domain, held-out from the dataset
- Texts with **Rare** characters, only appeared once in whole data, half of them unseen in the 10K split
- A particularly challenging **Heteronyms** test set

Findings: NLR learns to speak according to the lexicon

- NLR has significantly fewer errors than the baseline under low-resource conditions
- The lexicon-reading capability can be transferred to another language Cantonese/Japanese to achieve E2E low-resource adaptation on non-phonemic scripts
- Generalizable to characters totally unseen in training, as long as their lexicon entries are given in inference
- NLR is better at resolving heteronyms
- Attention heatmap shows that correct pronunciations have much higher energy cf. incorrect ones
- Pronunciations can be easily manipulated by changing the readings or the explanations of a reading in the lexicon

Table 1: *Objective CER(%) for Mandarin systems*

Dataset size	18K	10K	7.5K	5K
Baseline	4.65	9.40	18.33	Fail
NLR	4.82	5.86	**7.14**	**13.64**

Table 2: *Subjective error rate (%) on different test sets*

	Rare		Heteronyms	
Dataset size	18K	10K	18K	10K
Baseline	8.0	62.0	75.5	80.9
NLR	2.0	4.0	72.6	76.4

	General			
Dataset size	18K	10K	7.5K	5K
Baseline	0.9	5.4	10.9	Fail
NLR	0.3	1.9	4.1	8.5

Table 3: *CER(%) for low-resource adaptation to a different language, with different dataset size for each column*

Cantonese	2K	1K	750	500	250
Baseline	12.32	14.89	17.64	22.27	35.08
NLR	8.79	10.04	10.26	10.73	12.85
Japanese	**5K**	**3K**	**2K**	**1K**	**750**
Baseline	12.46	15.99	18.66	26.85	33.76
NLR	10.50	12.48	13.95	19.29	21.81