ETH zürich

N.ERA ROBOTIC

REDUCING SPELLING INCONSISTENCIES IN CODE-SWITCHING ASR USING CONTEXTUALIZED CTC LOSS

Burin Naowarat[1], Thananchai Kongthaworn[1], Korrawe Karunratanakul[2], Sheng Hui Wu[3], and Ekapol Chuangsuwanich[1]

[1]Department of Computer Engineering, Chulalongkorn University, Bangkok, Thailand
[2]ETH Zurich, Switzerland, [3]NewEra AI Robotics, Taiwan

ICASSP 2021

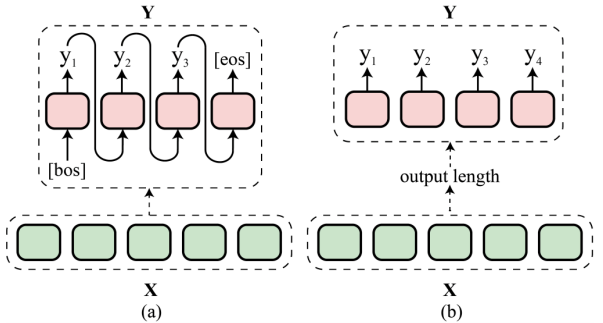
Code-Switching Speeches

- Alternating languages within a conversation
- Alternating languages within a sentence
  - Loanwords (mostly)
    - Thai: คนที่มี followers เพียงแค่ 5000
    - Eng: The person who has only 5000 followers.
  - Loanphrases
    - Thai: ผม work from home มาเกือบจะ 4 เดือนแล้ว
    - Eng: I have worked from home for almost 4 months.
- Usually found in bilingual communities

Fully Convolutional Code-Switching ASR model

- Good
  - Fast

- Bad
  - Hard for Code-Switching speech
    - Inconsistent spellings
      - คนที่มี folaเวอร์ (follower) เพียงแค่ 5000
  - Hard for Thai
    - Alphabet orderings
      - ไฉน : The sound of the middle letter, ฉ, comes first.
      - อาลัย vs กำไล : Consonant first vs vowel first (similar pronunciation).



(a) Autoregressive; (b) Non-Autoregressive

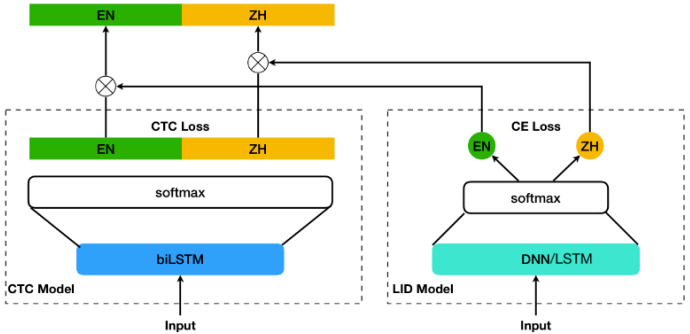
Thai word	folaเวอร์	ไฉน
IPA	fɒl.əʊ.ər	cha.năy

Kuchaiev, Oleksii, et al. "Mixed-precision training for nlp and speech recognition with openseq2seq." *arXiv preprint, 2018.*
Su, Yixuan, et al. "Non-Autoregressive Text Generation with Pre-trained Language Models." *EACL 2021.*

Possible Solutions

- Add dependencies between predicted letters
  - Do beamsearch with language model [1,2,3]
  - Do auto-regressive decoding [1,2,3]
  - Lose parallelizable ability

- Utilize language identifier [2,3,4]
  - Predicts language
  - Requires frame-level ground truths



[1] Sreeram, Ganji, and Rohit Sinha. "Exploration of End-to-End Framework for Code-Switching Speech Recognition Task: Challenges and Enhancements." *IEEE Access* 8 (2020)
[2] Shan, Changhao, et al. "Investigating end-to-end speech recognition for mandarin-english code-switching." *ICASSP 2019.*
[3] Zeng, Zhiping, et al. "On the end-to-end solution to mandarin-english code-switching speech recognition." *Interspeech 2019*
[4] Li, Ke, et al. "Towards code-switching ASR for end-to-end CTC models." *ICASSP 2019.*

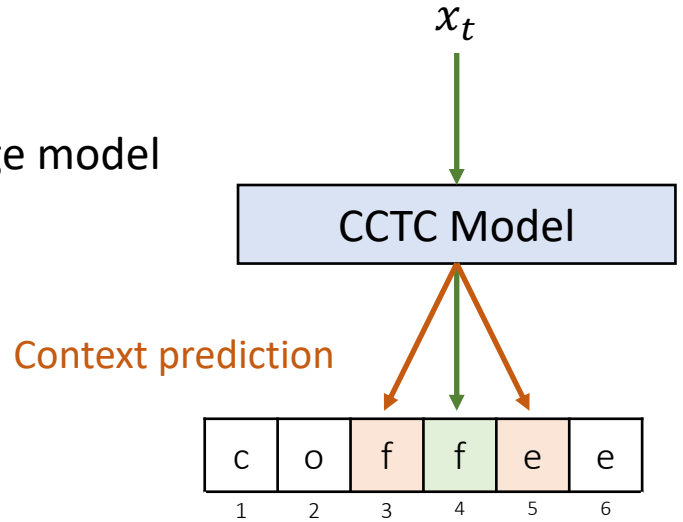
The Proposed CCTC

- Contextualized CTC (CCTC)
  - Mitigates the misspellings
  - Adds the dependencies between predicted letters without:
    - Losing parallelization ability
    - Needing external frame-level alignments
  - Does not increase the inference time

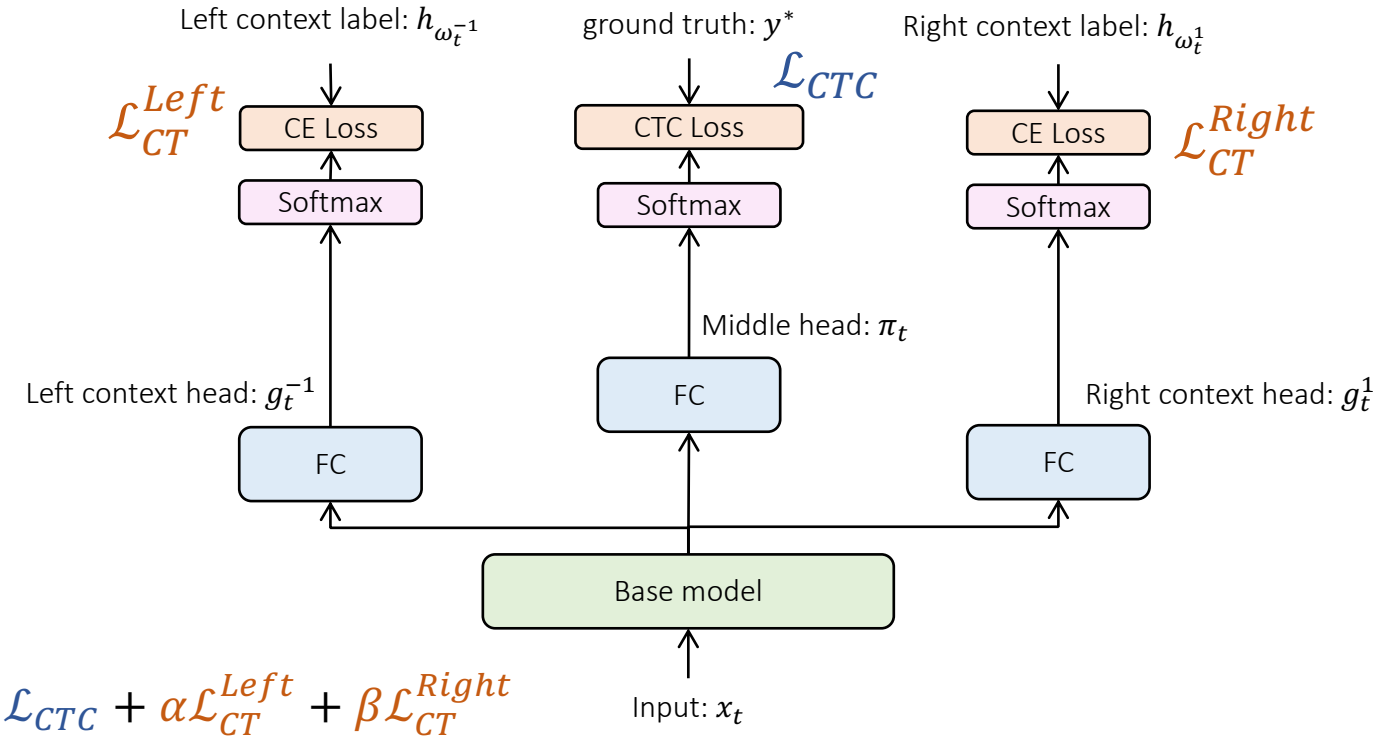
The Proposed CCTC

Add weak dependencies to the predicted letters

- Predict *surrounding* letters
  - Can be done in parallel
- Have similar effects to a low-order language model

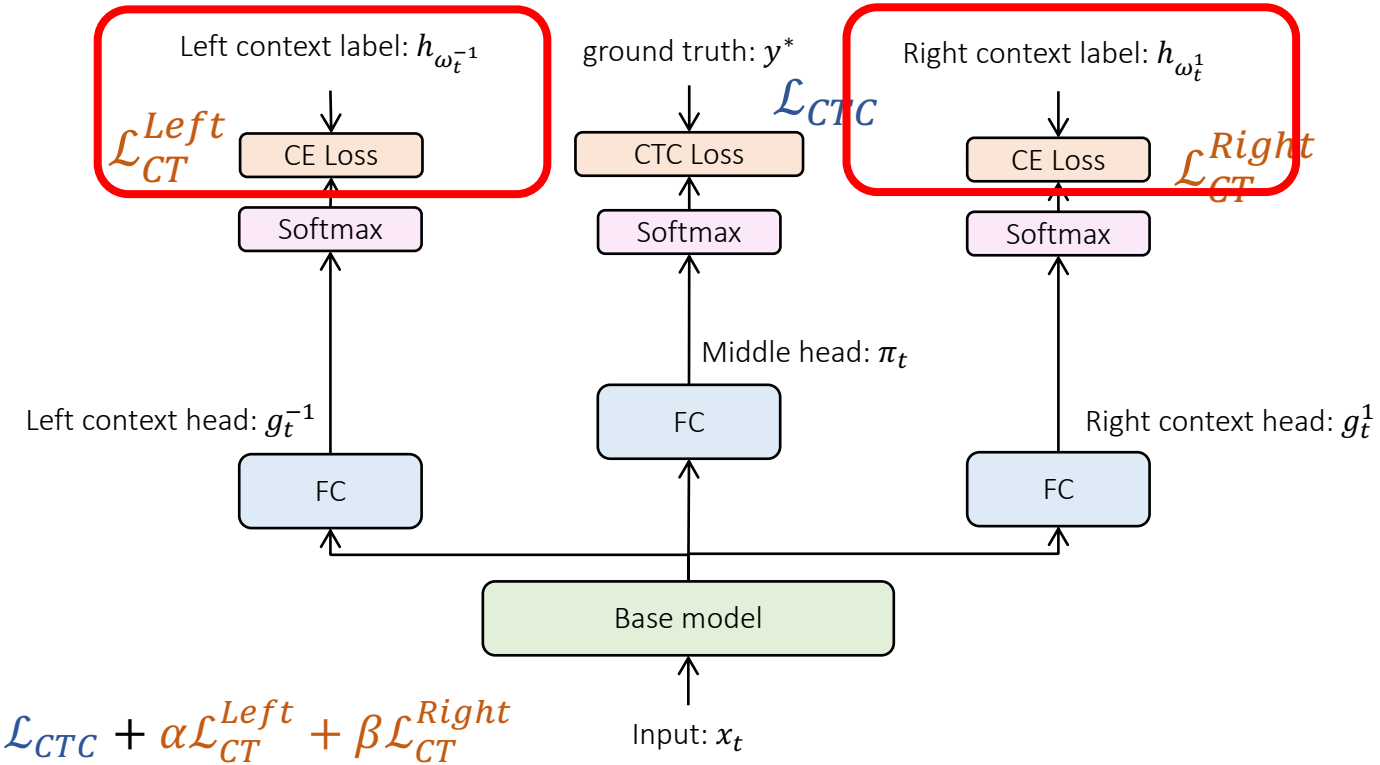


CCTC Model (Training)



$$\mathcal{L}_{CCTC} = \mathcal{L}_{CTC} + \alpha\mathcal{L}_{CT}^{Left} + \beta\mathcal{L}_{CT}^{Right}$$

CCTC Model (Training)



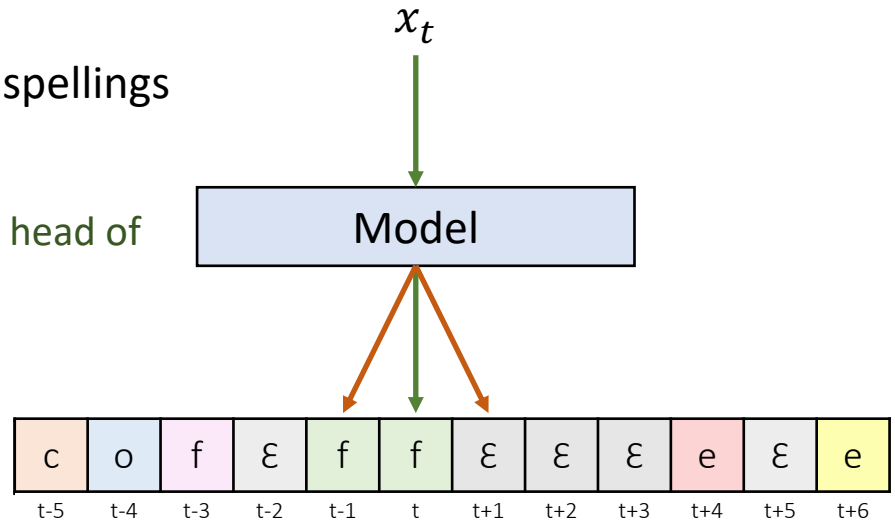
$$\mathcal{L}_{CCTC} = \mathcal{L}_{CTC} + \alpha\mathcal{L}_{CT}^{Left} + \beta\mathcal{L}_{CT}^{Right}$$

Ground truths of context predictions

Have no ground truth alignment

CCTC intend to encourage consistent spellings

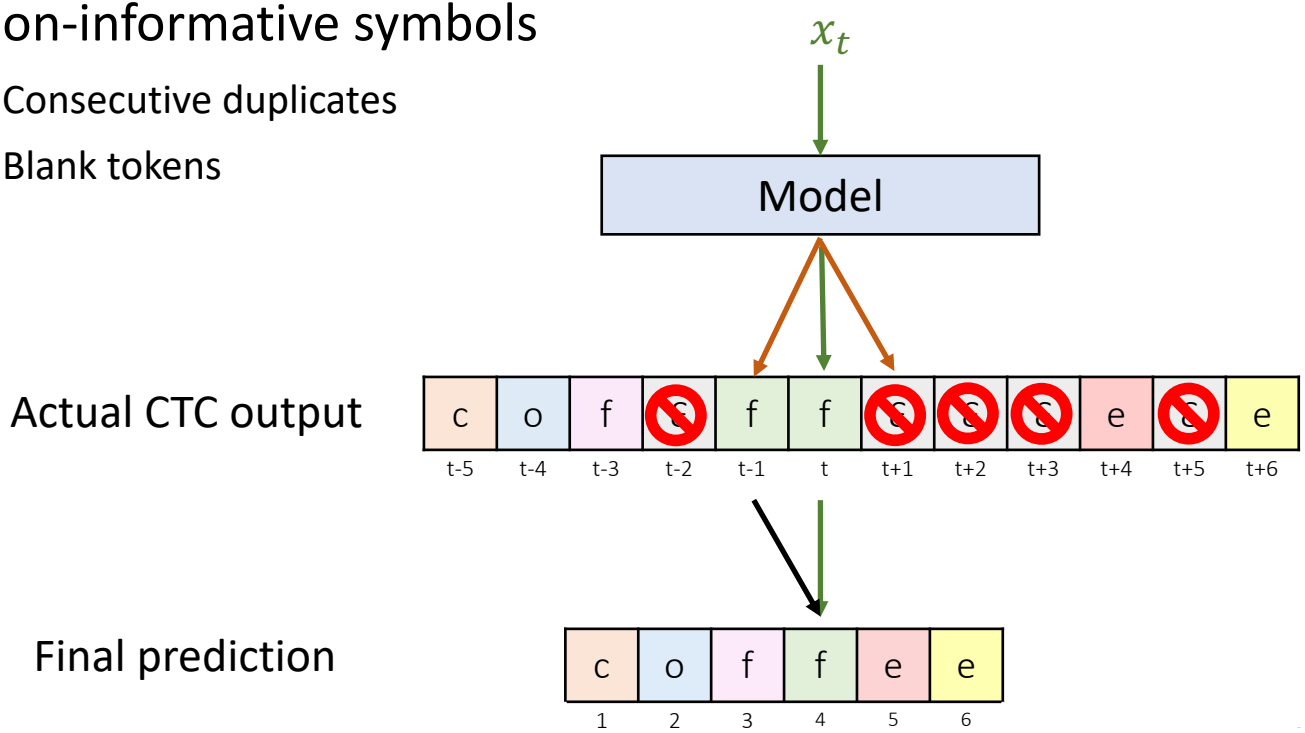
Solution: use prediction from the middle head of previous iterations as ground truths



Ground truths of context predictions

Non-informative symbols

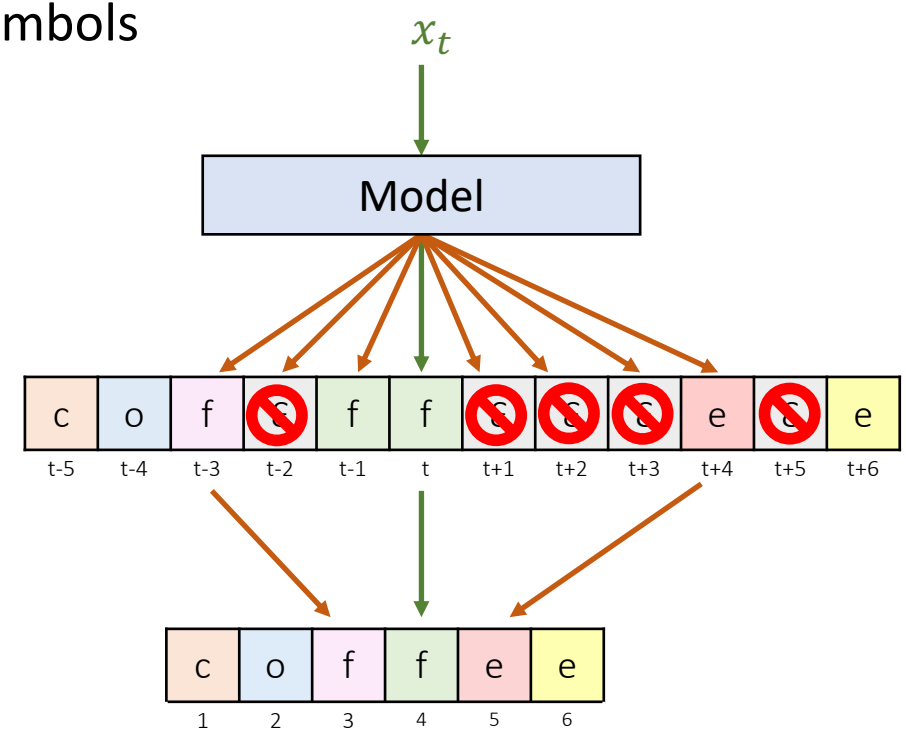
- Consecutive duplicates
- Blank tokens



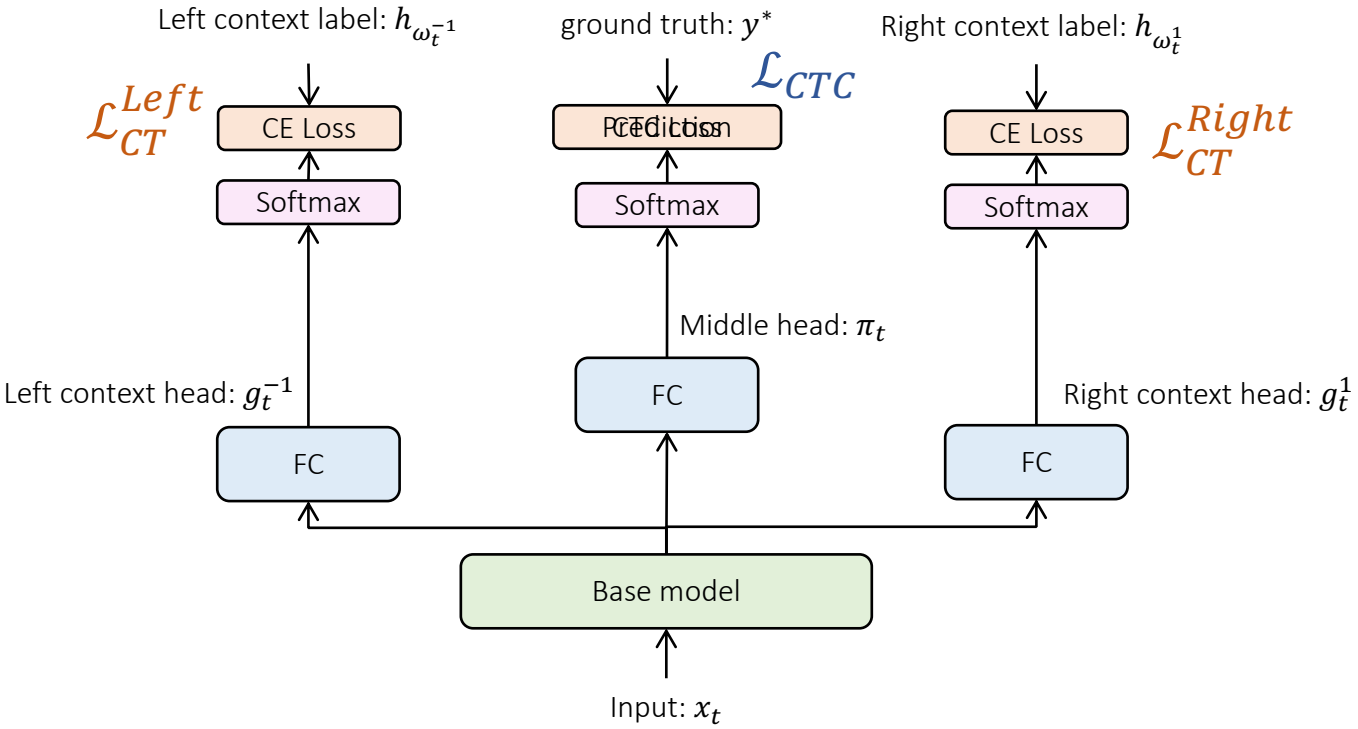
Ground truths of context predictions

Non-informative symbols

Solution: Skip all of them

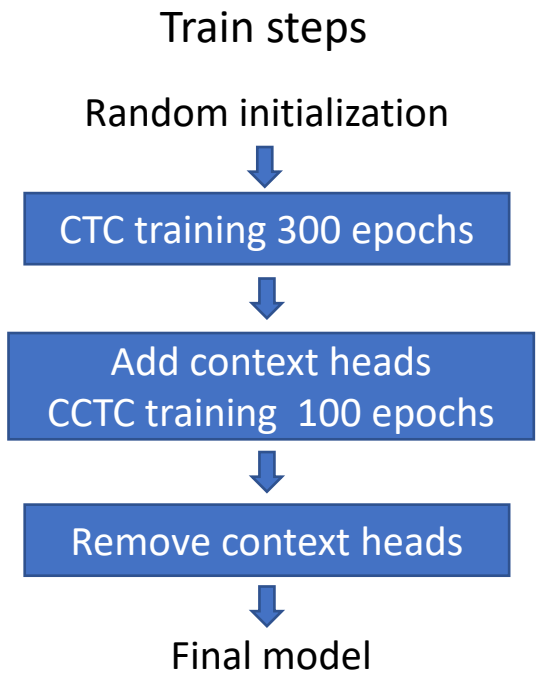


CCTC Model (Decoding)



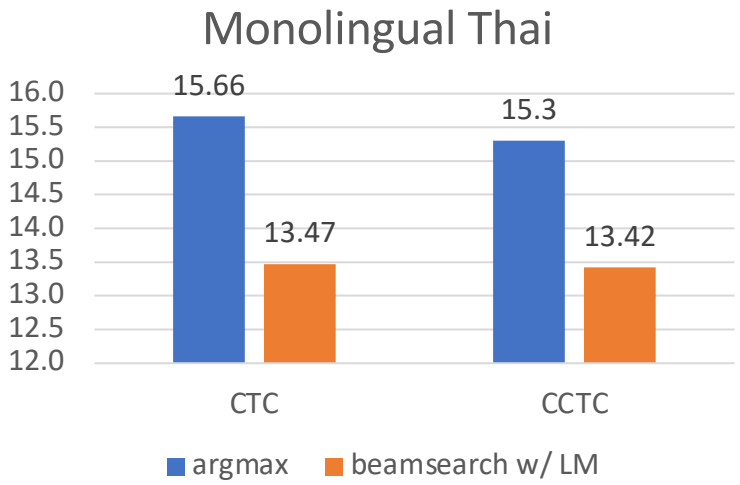
Experimental Setups

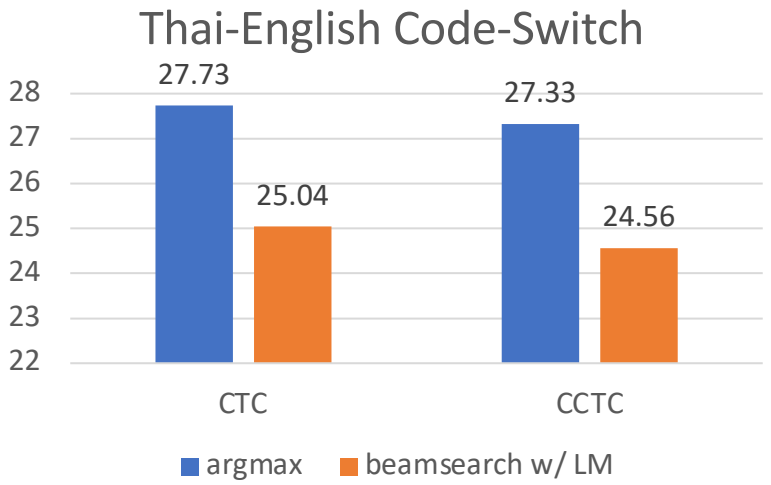
- Speech Corpus
  - 150 Hours
  - 180k monolingual utterances
  - 8.4k CS utterances
- Text Corpus
  - 96M words
  - 3-gram language model
- Base model
  - Wav2Letter+



Results (CS ASR)

CCTC consistently outperforms CTC





Results (CS ASR)

In Thai

- Fix wrong ordering

In Code-Switching

- Fix alphabets mixing

Argmax	CTC	ฉนย จึง ทร่ง สา งารถ แผ ศาสนา ของ พระองค์
	CCTC	ไฉไหน จึง ทรง สามารถ แผ่ ศาสนา ของ พระองค์
3-gram LM	CTC	ทน จึง ทรง สามารถ แผ่ ศาสนา ของ พระองค์
	CCTC	ไฉน จึง ทรง สามารถ แผ่ ศาสนา ของ พระองค์
Ground truth		ไฉน จึง ทรง สามารถ แผ่ ศาสนา ของ พระองค์
Argmax	CTC	คน ที่ มี folเวอร์ เพียง แค ห้า พัน
	CCTC	คน ที่ มี polower เพียง แค่ ห้า พัน
3-gram LM	CTC	คน ที่ มี ฟลัวร์ เพียง แค่ ห้า พัน
	CCTC	คน ที่ มี follower เพียง แค่ ห้า พัน
Ground truth		คน ที่ มี follower เพียง แค่ ห้า พัน

Conclusions

- We proposed CCTC for reducing inconsistent spellings of non-autoregressive Code-Switching ASR models.
- CCTC adds language dependencies to the predicted letters without:
  - Modifying the output units
  - Losing parallelizable ability
  - Needing for frame-level alignments
- CCTC can be integrated with any existing CTC models without increasing the inference time.