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

Summary

- Non-autoregressive models produce inconsistent spellings in Code-Switching ASR.
- CCTC loss mitigates the problem by:
  - Adding language dependencies to letters without:
    - Losing parallelizability
    - Needing of frame-level alignments
    - Modifying the output units
- We show the effectiveness of CCTC in both Code-Switching (CS) ASR and monolingual ASR.

Code-Switching ASR

- Code-Switching (CS) speech alternates languages within an utterance.
  - Borrow words
    - Thai: คนที่มี follower เพียงแค่ 5000
    - Eng: The person who has only 5000 followers.
  - Borrow phrases
    - Thai: ผม work from home มาเกือบจะ 4 เดือนแล้ว
    - Eng: I have worked from home for almost 4 months.
- Fully convolutional non-autoregressive model is fast.
  - It predicts all tokens along time axis at once.
  - It lacks dependencies between predicted letters

Motivation

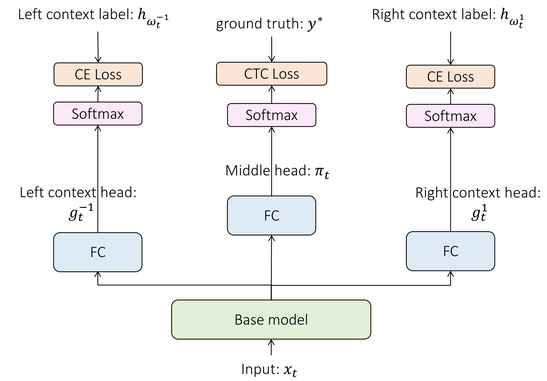
- No letter dependency raises the problem of:
  - Inconsistent spelling
    - Mixing alphabets from many languages within a single word ➔ unreadable
      - Example: คน ที่ มี โฟลเวอร์ เพียง แค่ ห้า พัน

        fol.ɐʊ.er
    - Wrong character ordering
      - Predicting `ฉนย` instead of the ground truth `ไฉน`.

        cha.na.ya cha.năy
      - The sound of the middle letter, ฉ, comes first.

CCTC model

- Introduce additional context prediction heads:
  - For predicting "*surrounding*" letters [1]
- The middle head is trained by CTC loss.
- The context heads are trained by Cross Entropy (CE) loss.
- The model is trained in a multi-task learning manner.

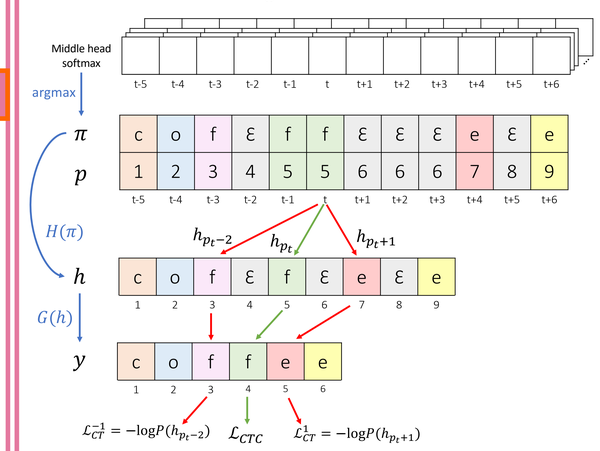


CCTC loss

- Contextualized CTC (CCTC) loss is defined below:

$$\mathcal{L}_{CCTC} = \mathcal{L}_{CTC} + \sum_{k=1}^{K} \alpha^{-k}\mathcal{L}_{CT}^{-k} + \beta^{k}\mathcal{L}_{CT}^{k}$$

- We used the prediction from the previous iteration as ground truths for training context heads ($\mathcal{L}_{CT}$).
  - Consecutive duplicates are ignored.
  - Blank tokens are ignored.



CS Corpus

	Train	Development	Test
duration	150 Hr	24 Hr	26 Hr
#total utterances	190K	30K	35K
#TH-CS utterances	8.4K	1.3K	2K
#TH letters	7M	1M	1M
#EN letters	84K	30K	19K
#TH words	1.9M	293K	333K
#EN words	14K	2K	3K
#TH vocabulary	36K	12K	13K
#EN vocabulary	3K	1K	1K

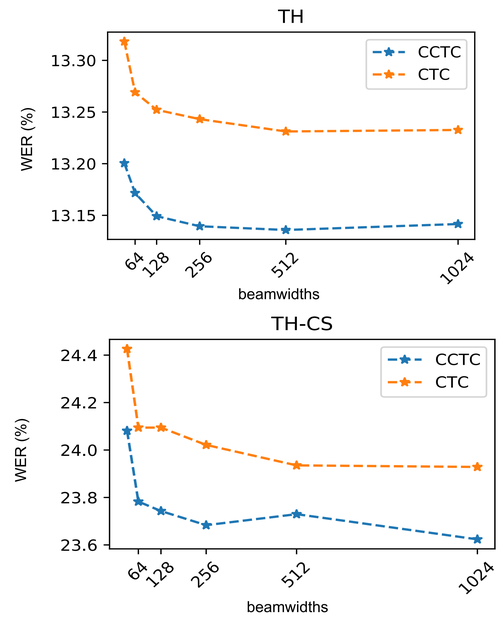
Results on CS corpus

- The training set includes both monolingual and CS utterances.
- The evaluation set is separated into:
  - TH (Thai Monolingual)
  - TH-CS (Thai-English code-switching)
- The model is a fully-convolutional Wav2Letter+ [2].
- The asterisk indicates significant difference using MAPSSWE test.

Data	Model	WER (%) argmax	beam	3-gram
Development set				
TH	CTC	15.01	14.89	13.27
	CCTC	**14.67***	**14.58***	**13.17***
TH-CS	CTC	28.02	27.76	24.09
	CCTC	**27.57***	**27.43**	**23.78**
Test set				
TH	CTC	15.66	15.52	13.47
	CCTC	**15.30***	**15.22***	**13.42**
TH-CS	CTC	27.73	27.74	25.04
	CCTC	**27.33***	**27.27***	**24.56***

Argmax	CTC	ฉนย จึง ทร่ง สามารถ 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 โฟลเวอร์ เพียง แค่ 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

Effect of beamwidth



Results on Monolingual English

- Train: LibriSpeech clean 100 hr subset
- Test: LibriSpeech clean

Model	Decoder	WER
Wav2Letter+ w/ CTC	greedy	22.00
Wav2Letter+ w/ CCTC	greedy	**21.32**
Wav2Letter++ w/ CTC [3]	beam w/ 3-gram LM	18.97
Wav2Letter+ w/ CTC	beam w/ 3-gram LM	15.72
Wav2Letter+ w/ CCTC	beam w/ 3-gram LM	**15.67**

Future Work

- Explore the effectiveness of CCTC in other sequence predicting problems
- Explore the benefits of using wider contexts

References

[1] Zhang, Yu, et al. "Speech recognition with prediction-adaptation-correction recurrent neural networks." *ICASSP* 2015.

[2] Kuchaiev, Oleksii, et al. "Mixed-precision training for nlp and speech recognition with openseq2seq." *arXiv preprint, 2018.*

[3] Pratap, Vineel, et al. "Wav2letter++: A fast open-source speech recognition system." *ICASSP 2019.*