Context-Aware Multi-task Learning End-to-End Code-Switching Speech Recognition

Zimeng Qiu, Yiyuan Li, Jiaxu Zou,
Electrical & Computer Engineering Department
Carnegie Mellon University

Motivation

Previous ASR task mainly focused on monolingual task, we explored employing ASR in code-switching in multi-task framework, and try to mine intention of code-switching instance in sentimental perspective. Namely, the work is concluded as

- Extending ASR to Code-Switching setting.
- Exploring different model architectures.
- Sentimental Evaluation of ASR result.

What is Code-Switching

#1: 我还挺guilty 的我每次都骂我爸cause 我爸整天烦我
I feel pretty guilty each time I scold my dad cause my dad annoys me all the time.

 # 2: Oh shit 这个真的confidential 不可以乱讲
Oh shit this is really confidential and could not be distributed.

- Speakers alternate between two or more languages, or language varieties and styles, in the context of a single utterance or discourse.
- Provides a convenient way for speakers to express themselves.

Dataset

	Train	Dev	Test
Utt.	129,217	16,156	16,152
Tkn.	1,879,778	232,360	237,487
EN	583,210	73,390	73,470
CN	1,296,568	158,970	164,017

Table 1: SEAME dataset train/dev/test splits.

- The main dataset we're going to use is SEAME, which is a a Mandarin-English code-switching speech corpus in South-East Asia. The statistics of the dataset is shown above.
- There are several settings when code-switching takes place in this dataset: code-switching in the **linking words**; choosing the **terminology** that frequently used in the daily life and code-switching words with **strong sentiment**.

Proposed Models

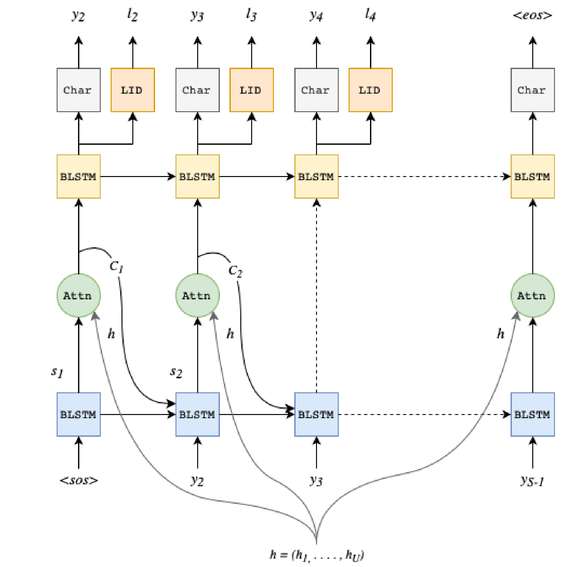


Figure 1: Proposed model 1 (MTL-LAS): Multi-task learning language ID and speech to character sequence.

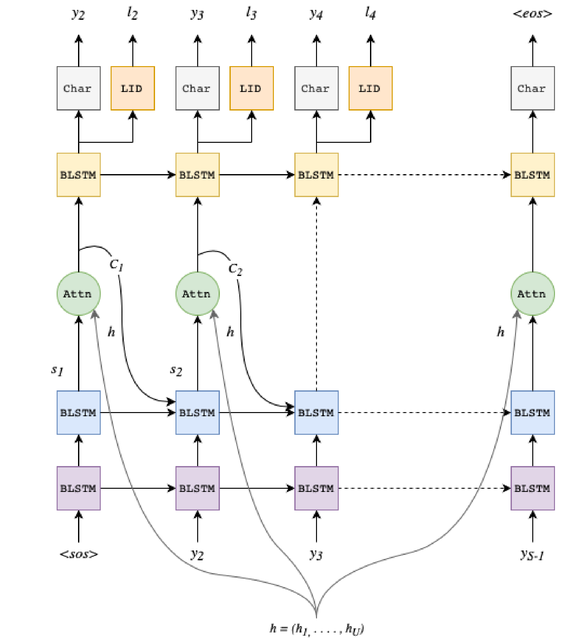


Figure 2: Proposed model 2 (MTL-Context-LAS): Multi-task learning language ID and speech to character sequence with awareness of context.

Experiment Results

We conducted experiments on SEAME dataset to test our models according to the train, dev and test sets split mentioned in Dataset section.
We also develop a tool to auto-correct English tokens in the generated transcript. Our intuition is that since the model is on character level, while avoiding out-of-vocabulary words, it generates many non-sense tokens, for example:

Generated English Tokens	Ground Truth
gin mamony	gig memory
forr tweee two one	four three two one
e mhink	i think

Table 2: Incorrect generated English tokens vs ground truth.

Therefore, for each generated English token, we first use PyEnchant to decide if this token is an English word, if not, then searching within the training corpus to find the word that has least edit distance with the genrated token.

Model	CER	Correction CER
Baseline (LAS)	48.25%	53.11%
MTL-LAS	27.06%	29.06%
MTL-Context-LAS	**26.84%**	28.47%

Table 3: Experiment results.

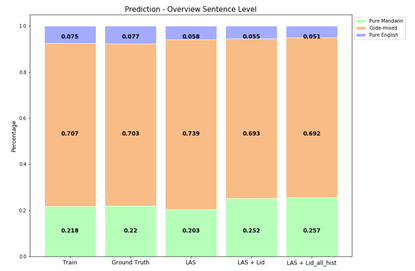
Unfortunately the auto-correct tool did not help get lower CER, it might because the tool we use to check English words is not good enough, or because the corpus is not large enough.

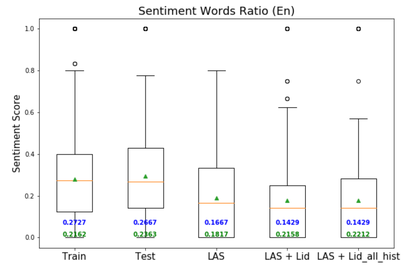
Conclusion

- We proposed a context-aware multi-task learning end-to-end code-switching speech recognition model that improves character error rate significantly on SEAME dataset compared with baseline LAS model.
- Jointly learning language ID and speech recognition boost model performance tremendously.
- Decoder attention mechanism conditioned on previous output character history helped to generate more reasonable English tokens.

Analysis

- Sentimental Analysis (SentiWord)





- Case Study Comparison

Example #1	
GT:	then erm 反正就是那些erm 不是很high end but 也不是很low class 的那种我还满喜欢的
LAS:	then erm undred 最很睦是很hard d. e. 也不是很low class 的这个
MTL-LAS:	then erm 反是就是那些erm 不是很higd and but 也不是很low class 的那种万还蛮蛮喜欢的这么我我我自o 有玩玩玩玩
MTL-Context-LAS:	then erm 饭正就是那些erm 不是很high and 那yt 也不是很low class 的那种话还满喜欢的
Example #2	
GT:	你看那个他们在那边subscribe for i. phone la maybe 你只是pay for
LAS:	每看那一我们在那边理是screft for i. phone la main bay 你就是被four hor
MTL-LAS:	你看那个他们真那边你nee irieesfor e.fphone 啊ahmaybe 你只是pay fou
MTL-Context-LAS:	你看那个他们真那边nisstriae for i. phone aarmaybe 你只是被pay for

Future Work

- Conduct experiments on other code-switching datsets, e.g. Miami corpus.
- Pre-train the model on monolingual English speech dataset to boost the performance on generate English tokens.