Joint Modeling of Code-Switched and Monolingual ASR via Conditional Factorization

Brian Yan, Chunlei Zhang, Meng Yu, Shi-Xiong Zhang, Siddharth Dalmia, Dan Berrebbi, Chao Weng, Shinji Watanabe, Dong Yu

Carnegie Mellon University

Tencent AI Lab

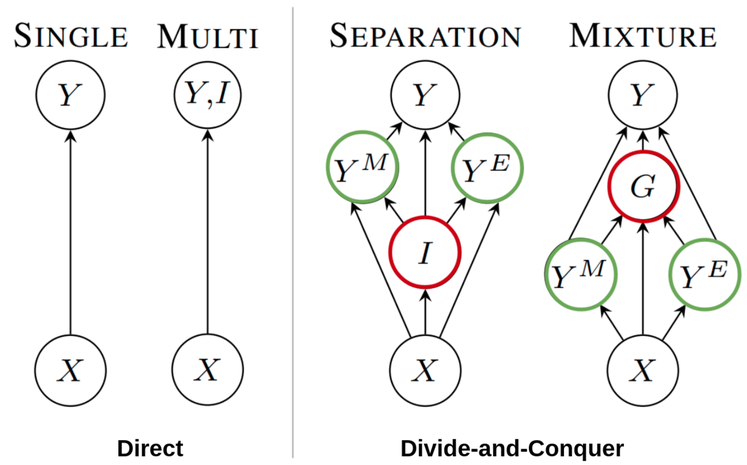
At a Glance

- We first propose a formulation of the bilingual ASR problem as a ***conditionally factorized joint model*** of monolingual and CS ASR where the final output is obtained given only monolingual label-to-frame synchronized information
- We then apply an end-to-end neural network, which we call the ***Conditional RNN-T***, to model our conditional joint formulation

Code-Switching is a Subset of Bilingualism

In this work: we evaluate our models on not only Mandarin-English CS Mixed Error Rate (MER) but also Monolingual English Character Error Rate (CER), and Monolingual English Word Error Rate (WER).

Probabilistic Graphical Models of Prior Works



Direct Graphical Models

- Pro: Simple to only model a direct dependency
- Con: Dependency becomes too complex for unrelated languages

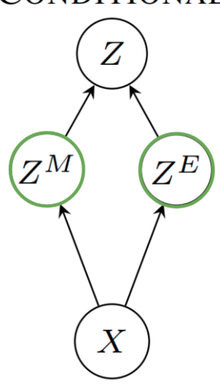
Divide-and-Conquer Models

- Pro: Division of monolingual subtasks → simpler dependencies, more compatible with monolingual data
- Con: Dependence on quality of "divider" module → risk of error propagation, increased complexity

Motivation: Simplest PGM that still has monolingual subtasks?

Conditionally Factorized Joint Model

CONDITIONAL



X : speech features

Z : bilingual transcript
Z^M: mandarin sub-task
Z^E: english sub-task

Z is label-to-frame alignment:

$$z_t = \begin{cases} z_t^M, & \text{if } z_t^M \in \mathcal{V}^M \text{ and } z_t^E = \varnothing \\ z_t^E, & \text{if } z_t^E \in \mathcal{V}^E \text{ and } z_t^M = \varnothing \\ \varnothing, & \text{if } z_t^M = \varnothing \text{ and } z_t^E = \varnothing \end{cases}$$

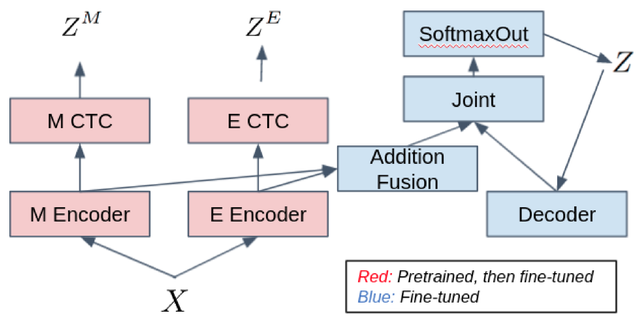
Jointly model CS and Monolingual parts, w/ **conditional factorization**:

$$\begin{aligned} p(Z|X) &= p(Z, Z^M, Z^E|X) \\ &= p(Z|Z^M, Z^E, X)p(Z^M, Z^E|X) \\ &\approx p(Z|Z^M, Z^E, \cancel{X})p(Z^M|X)p(Z^E|X) \end{aligned}$$

Conditional independence *Independence*

Conditional RNN-T

$$p(Y|X) \approx \underbrace{\sum_{\mathcal{Z}} p(Z|Z^M, Z^E)}_{\triangleq p_{\text{rnnt}}(Y|Z^M, Z^E)} \underbrace{\sum_{\mathcal{Z}^M} p(Z^M|X)}_{\triangleq p_{\text{ctc}}(Y^M|X)} \underbrace{\sum_{\mathcal{Z}^E} p(Z^E|X)}_{\triangleq p_{\text{ctc}}(Y^E|X)}$$



$$\mathcal{L}_{\text{LS}} = \lambda\mathcal{L}_{\text{RNNT}} + (1-\lambda)(\mathcal{L}_{\text{M\_CTC}} + \mathcal{L}_{\text{E\_CTC}})$$

Original Bilingual g.t. (*Z*) 什么是 Code-Switching
Masked Mandarin g.t. (*Z^M*) 什么是 <en>
Masked English g.t. (*Z^E*) <zh> Code-Switching

Code-Switched and Monolingual Results

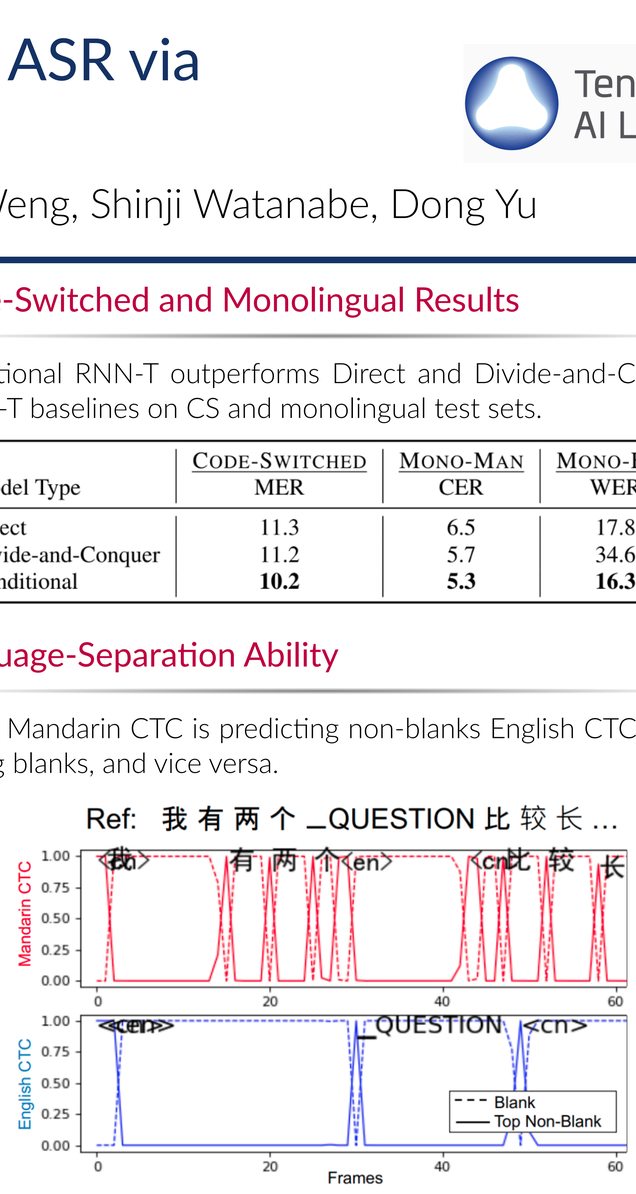
Conditional RNN-T outperforms Direct and Divide-and-Conquer RNNT-T baselines on CS and monolingual test sets.

Model Type	CODE-SWITCHED MER	MONO-MAN CER	MONO-ENG WER
Direct	11.3	6.5	17.8
Divide-and-Conquer	11.2	5.7	34.6
Conditional	**10.2**	**5.3**	**16.3**

Language-Separation Ability

When Mandarin CTC is predicting non-blanks English CTC is predicting blanks, and vice versa.

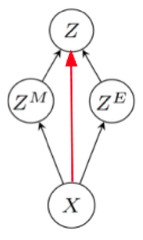
Ref: 我 有 两 个 _QUESTION 比 较 长 …



Experimental Validation of Conditional Independence

No additional information from X is required given Z^M and Z^E.

Model	Bilingual Condition	CODE-SWITCHED MER	CER	WER
Cond. RNN-T + LS	$p(Z\|Z^M, Z^E)$	**11.1**	**8.9**	**31.1**
3-Enc. RNN-T + LS	$p(Z\|Z^M, Z^E, X)$	11.2	9.0	**31.1**



3-encoder variant passes *X* to the bilingual module, creating the dependency in red

$$p(Z|Z^M, Z^E, X)$$