



Regularizing Neural Machine Translation by Target-bidirectional Agreement

Zhirui Zhang¹, Shuangzhi Wu², Shujie Liu³, Mu Li³, Ming Zhou³, Tong Xu¹

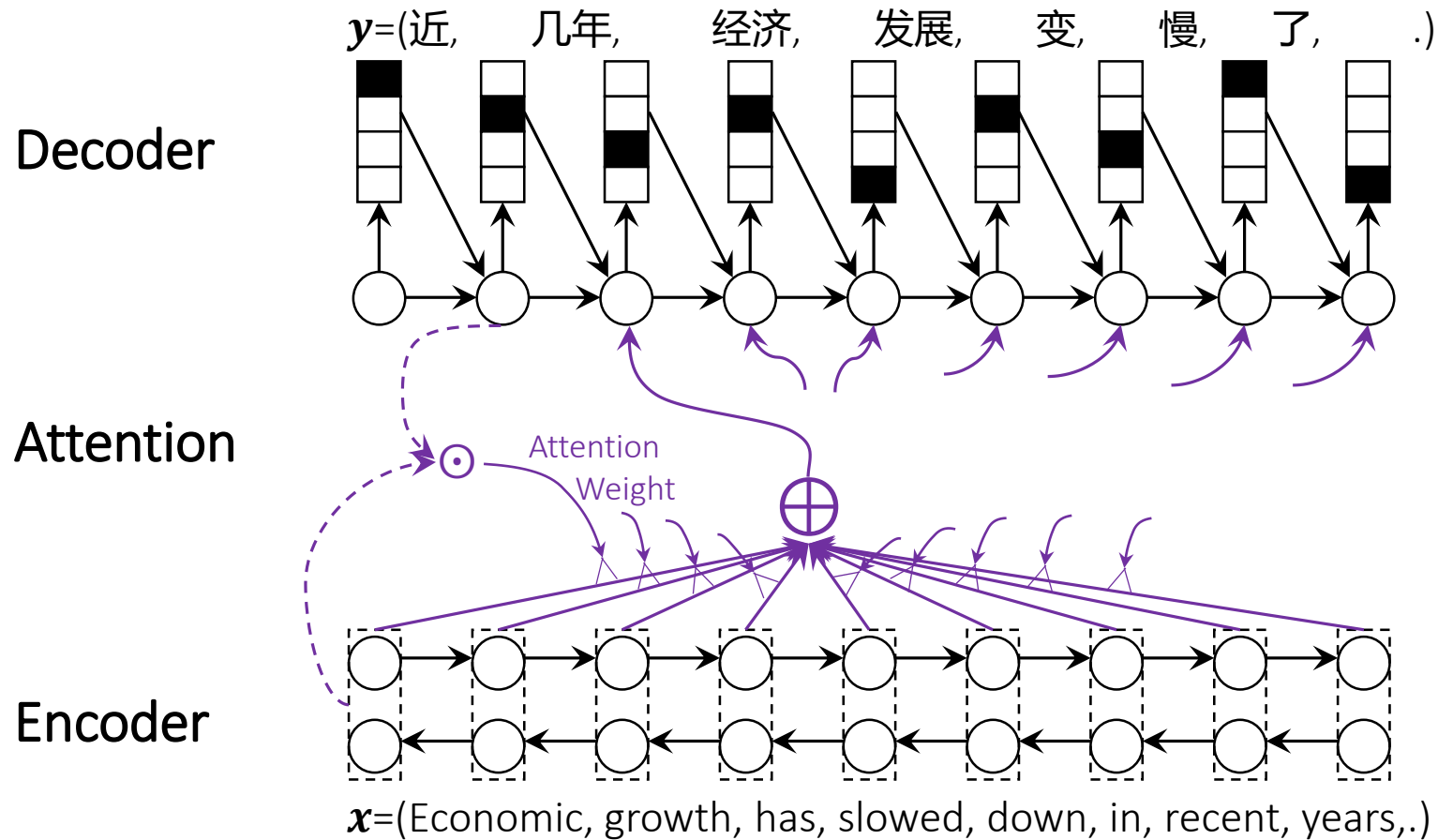
¹University of Science and Technology of China

²Harbin Institute of Technology ³Microsoft Research Asia

¹zrustc11@gmail.com tongxu@ustc.edu.cn

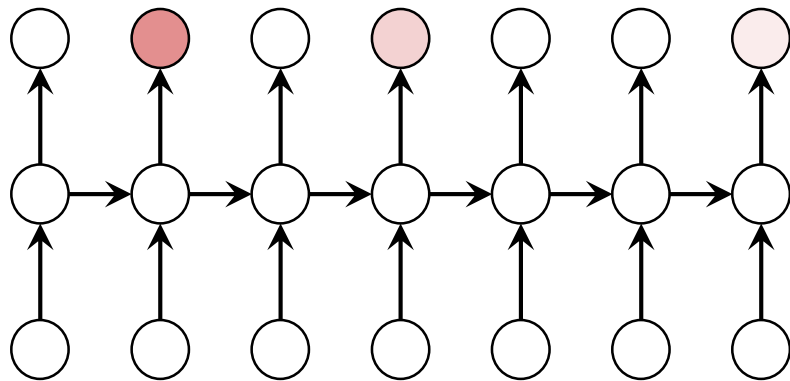
²³{v-shuawu, shujliu, muli, mingzhou}@microsoft.com

Neural Machine Translation

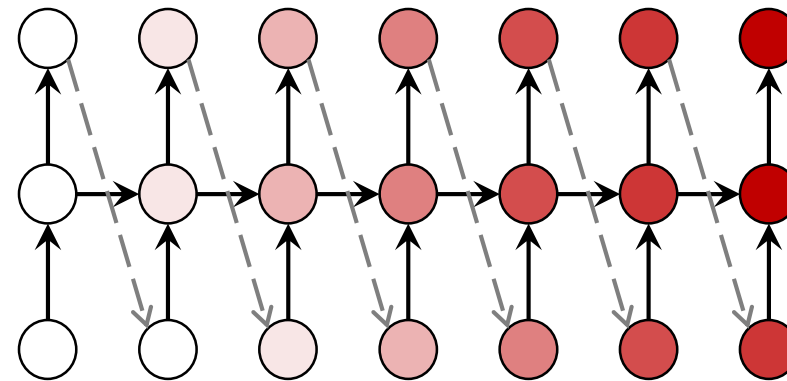


Motivation: Exposure Bias Problem

- NMT model is only trained with golden bilingual sentences
- Output sentence is auto-regressively generated word by word during decoding
- Previous errors will mislead the generation of the subsequences
- Errors will be quickly amplified



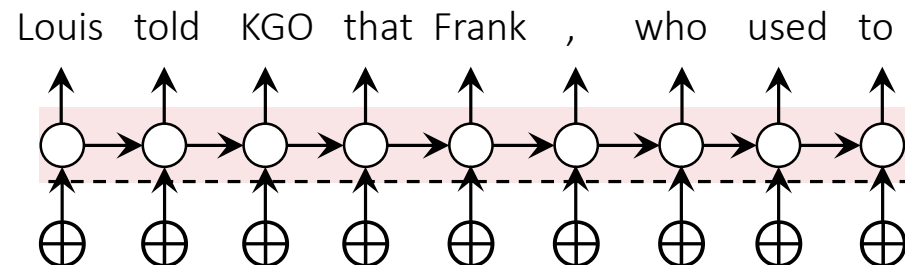
Training



Decoding

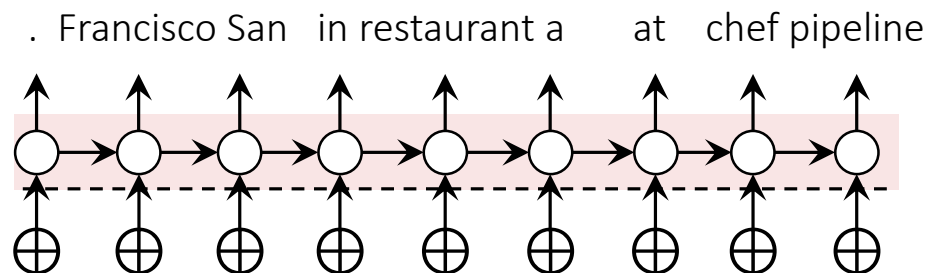
Motivation: Exposure Bias Problem

- Translate from Left to Right
Louis told KGO that Frank, who used to work as a cook in Boston, **had found a job in a restaurant.**



Louis 对 KGO 表示, 之前在波士顿做厨师的 Frank 在 **旧金山** 的餐馆找到一份流水线厨师的工作。

- Translate from Right to Left
Louis told KGO, had found a job as a pipeline chef at a restaurant in San Francisco .



Louis 对 KGO 表示, 之前在波士顿做厨师的 **Frank** 在旧金山的餐馆找到一份流水线厨师的工作。


Target-bidirectional Agreement Regularization

- Introduce two Kullback-Leibler (KL) divergence regularization terms

$$L(\vec{\theta}) = \sum_{n=1}^N \log P(y^{(n)}|x^{(n)}; \vec{\theta})$$


$$- \lambda \sum_{n=1}^N \text{KL}(P(y|x^{(n)}; \overleftarrow{\theta}) || P(y|x^{(n)}; \vec{\theta}))$$


$$- \lambda \sum_{n=1}^N \text{KL}(P(y|x^{(n)}; \vec{\theta}) || P(y|x^{(n)}; \overleftarrow{\theta}))$$



$$\frac{\partial \text{KL}(P(y|x^{(n)}; \overleftarrow{\theta}) || P(y|x^{(n)}; \vec{\theta}))}{\partial \vec{\theta}}$$

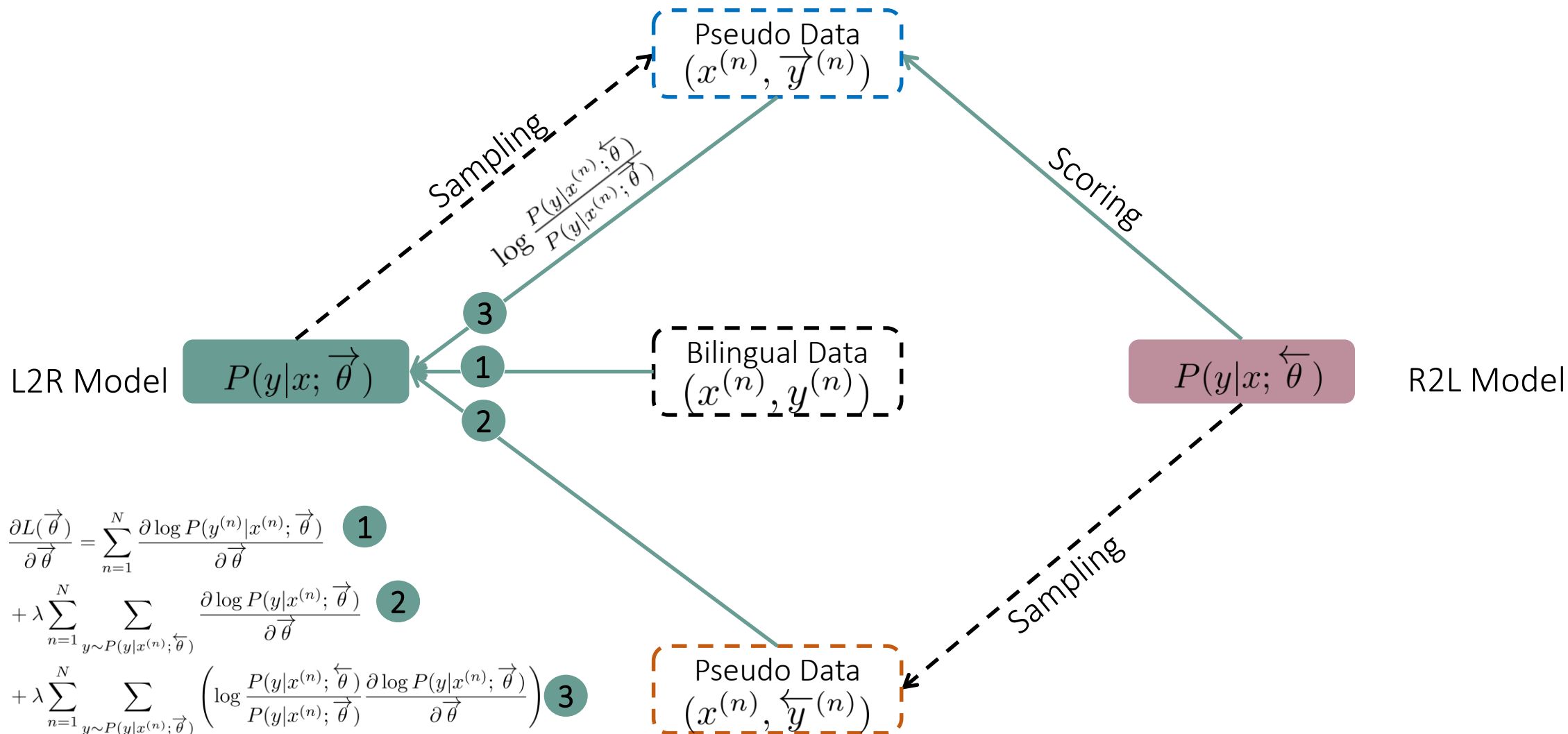
$$= - \mathbb{E}_{y \sim P(y|x^{(n)}; \overleftarrow{\theta})} \frac{\partial \log P(y|x^{(n)}; \vec{\theta})}{\partial \vec{\theta}}$$



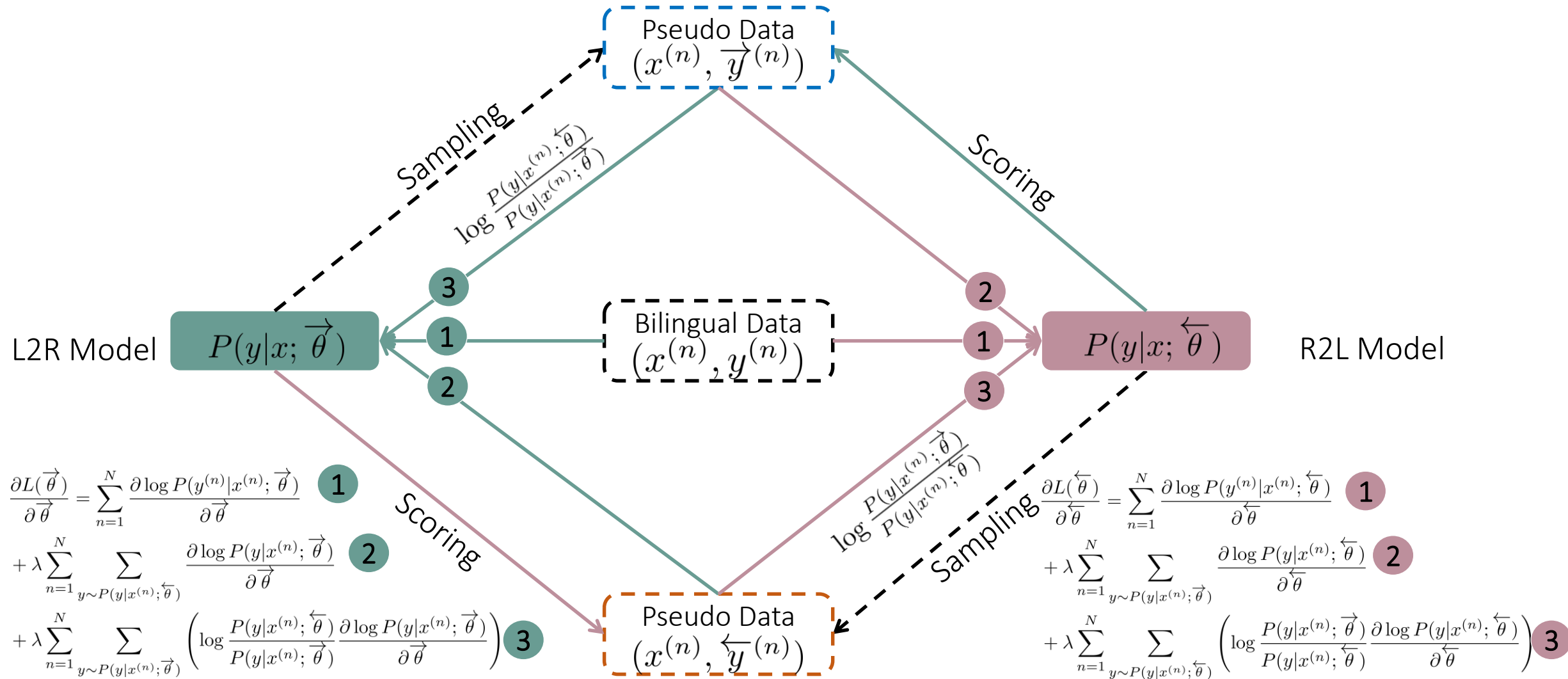
$$\frac{\partial \text{KL}(P(y|x^{(n)}; \vec{\theta}) || P(y|x^{(n)}; \overleftarrow{\theta}))}{\partial \vec{\theta}}$$


$$= - \mathbb{E}_{y \sim P(y|x^{(n)}; \vec{\theta})} \left(\boxed{\log \frac{P(y|x^{(n)}; \overleftarrow{\theta})}{P(y|x^{(n)}; \vec{\theta})}} \frac{\partial \log P(y|x^{(n)}; \vec{\theta})}{\partial \vec{\theta}} \right)$$

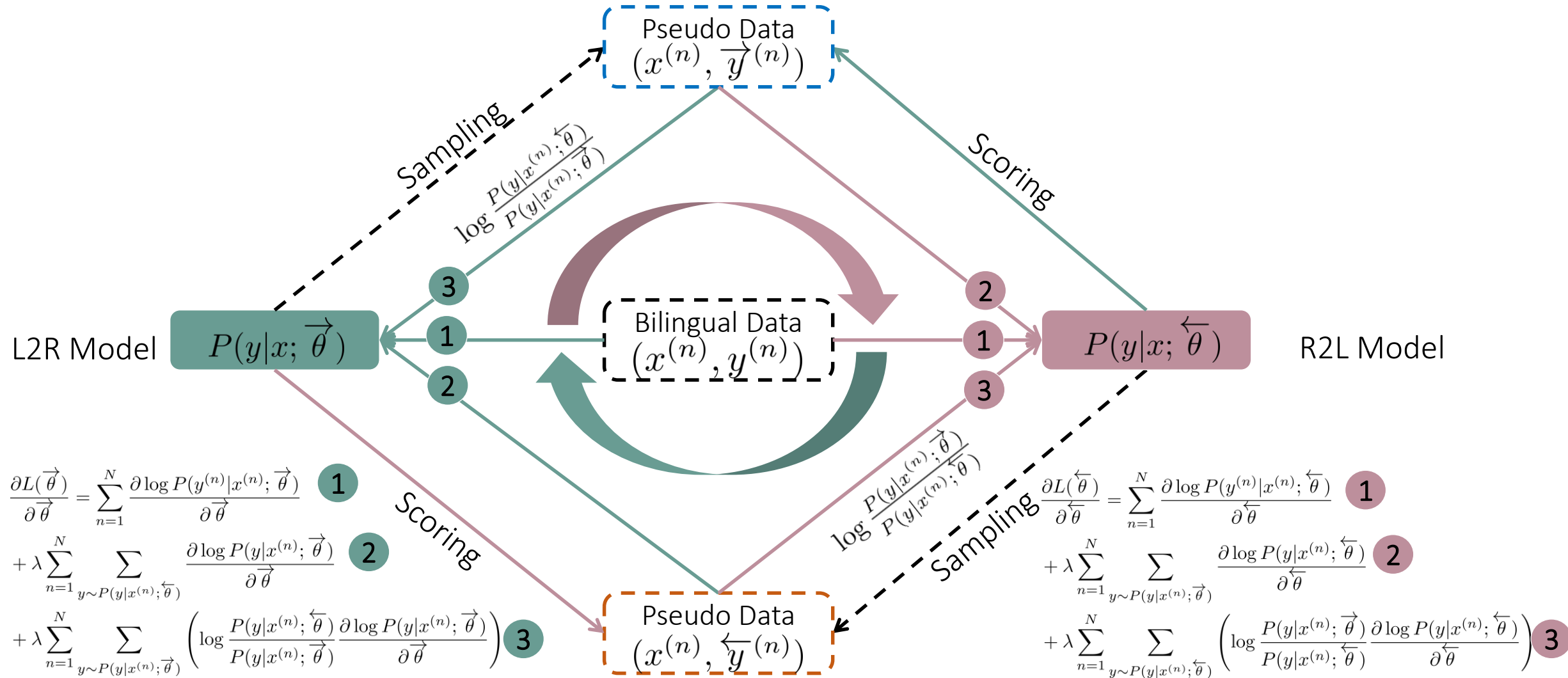
Agreement Regularization (R2L \rightarrow L2R)



Agreement Regularization (L2R \rightarrow R2L)



Agreement Regularization (Joint Training)



Experiments

- Evaluate our approach on three datasets, including NIST OpenMT for Chinese-English, WMT17 for English-German and Chinese-English
- NMT systems
 - **Transformer**: the-state-of-art NMT system (Vaswani et al. 2017)
 - **Transformer + MRT**: minimax risk training (Shen et al. 2016)
 - **Transformer + JS**: re-ranking the results from bidirectional models (Liu et al. 2016)
 - **Transformer + RT**: our method

Experimental Result

- Evaluation on NIST Corpora

System	NIST2006	NIST2003	NIST2005	NIST2008	NIST2012	Average
Transformer	44.33	45.69	43.94	34.80	32.63	40.28
Transformer+MRT	45.21	46.60	45.11	36.77	34.78	41.69
Transformer+JS	45.04	46.32	44.58	36.81	35.02	41.51
Transformer+RT	46.14	48.28	46.24	38.07	36.31	43.01

- Evaluation on WMT17 Corpora

System	English-German		Chinese-English	
	newstest2016	newstest2017	newsdev2017	newstest2017
Transformer	32.58	25.48	20.87	23.01
Transformer+MRT	33.27	25.87	21.66	24.24
Transformer+JS	32.91	25.93	21.25	23.59
Transformer+RT	34.56	27.18	22.50	25.38

Analysis

- Effect of Joint Training

	English-German	Chinese-English
Iteration 0	32.58	20.87
Iteration 1	33.86	21.92
Iteration 2	34.56	22.50
Iteration 3	34.58	22.47

Table 4: Translation performance of our method on WMT validation sets during training process. “Iteration 0” denotes baseline Transformer model.

- Performance on Long Sentences

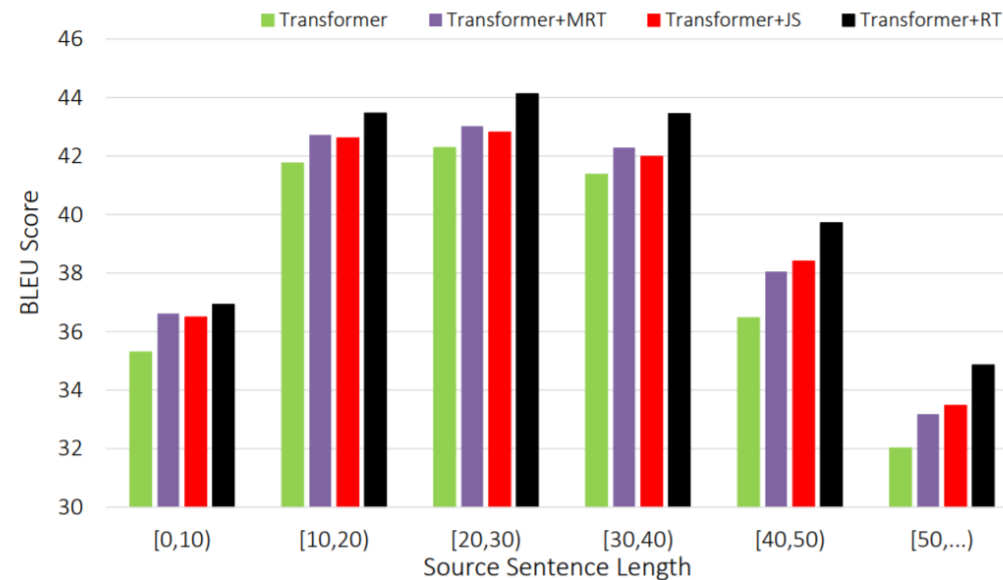


Figure 2: Performance of the generated translations with respect to the length of source sentences on NIST datasets.

Combining with Back-Translation Method

- Dataset
 - Monolingual data: randomly select 5M German sentences and 12M English sentences from “New Crawl: articles from 2016”
- Result

System	English-German		Chinese-English	
	newstest2016	newstest2017	newsdev2017	newstest2017
Transformer-big	33.58	27.13	21.91	24.03
Transformer-big+BT	35.06	28.34	23.59	25.53
Transformer-big+BT+RT	36.78	29.46	24.84	27.21
Edinburgh’s NMT System (ensemble)	36.20	28.30	24.00	25.70
Sogou’s NMT System (ensemble)	-	-	22.90	26.40

Table 3: Case-sensitive BLEU scores (%) for English-German and Chinese-English translation on WMT test sets. Edinburgh (Sennrich et al. 2017) and Sogou (Wang et al. 2017b) NMT systems are No.1 system in leaderboard of WMT 2017’s English-German and Chinese-English translation tasks respectively.

Example

Source	shòuhàirén dē gēgē Louis Galicia duì měiguó guǎnbō gōngsī wèiyú jiùjīnshān dē diàntái KGO biǎoshì, zhīqián zài bōshìdùn zuò liúshuǐxiàn chúshī dē Frank yú liùgèyuè qián zài jiùjīnshān dē Sons & Daughters cānguān zhǎodào yīfèn liúshuǐxiàn chúshī dē lǐxiǎng gōngzuò.
Reference	The victim's brother, Louis Galicia, told ABC station KGO in San Francisco that Frank, previously a line cook in Boston, had landed his dream job as line chef at San Franciscos Sons & Daughters restaurant six months ago.
Transformer	Louis Galicia, the victim's brother, told ABC radio station KGO in San Francisco that Frank, who used to work as an assembly line cook in Boston, <u>had found an ideal job in the Sons & Daughters restaurant in San Francisco</u> six months ago.
Transformer (R2L)	The victim's brother, Louis Galia, told ABC's station KGO, <u>in San Francisco, Frank</u> had found an ideal job as a pipeline chef six months ago at Sons & Daughters restaurant in San Francisco .
Transformer+RT	The victim's brother, Louis Galicia, told ABC radio station KGO in San Francisco that Frank, who previously worked as an assembly line cook in Boston, found an ideal job as an assembly line cook six months ago at Sons & Daughters restaurant in San Francisco.

Table 5: Translation examples of different systems. Text highlighted in wavy lines is incorrectly translated.

Summary

- Proposed a simple and efficient regularization approach
 - Introduce the agreement constraints between L2R and R2L NMT models
 - Design an approximation algorithm to enable fast training of the regularized terms and optimize L2R and R2L NMT models with joint training strategy
- Significant improvements on Chinese-English/English-German tasks
- Future work
 - Test this method on other sequence-to-sequence tasks
 - Try to integrate this method with other semi-supervised methods

Thank you!

Q & A