神经机器翻译的一致性表示学习

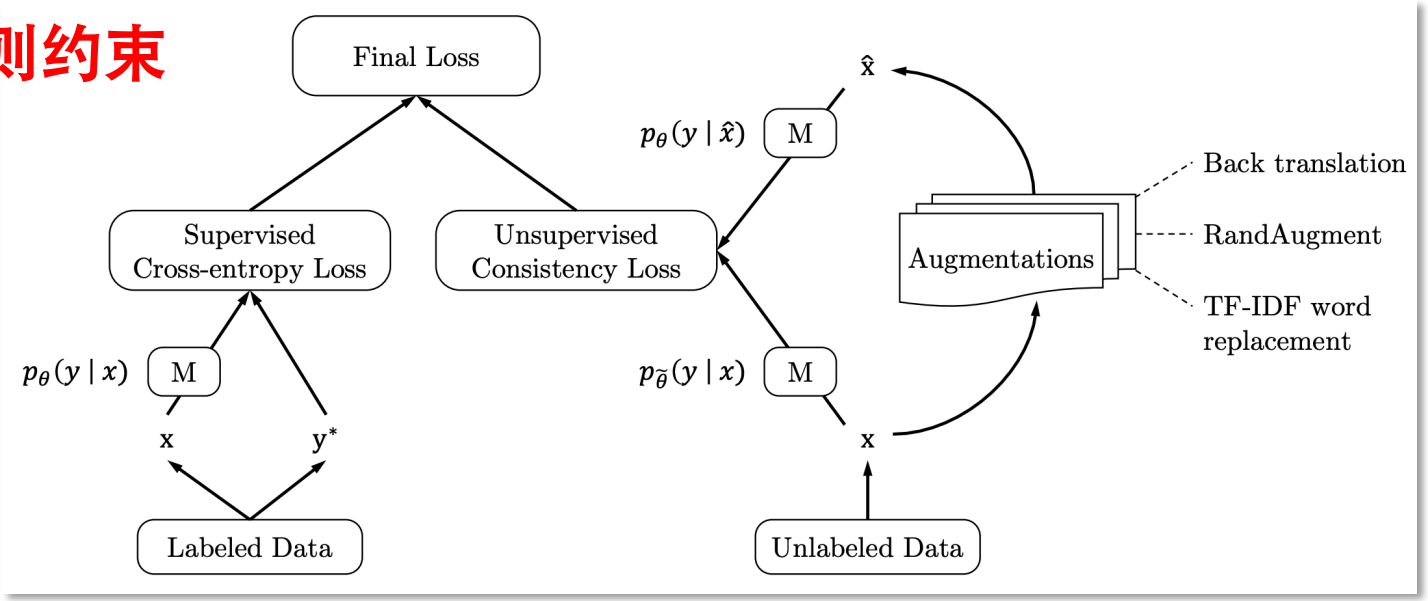
刘学博

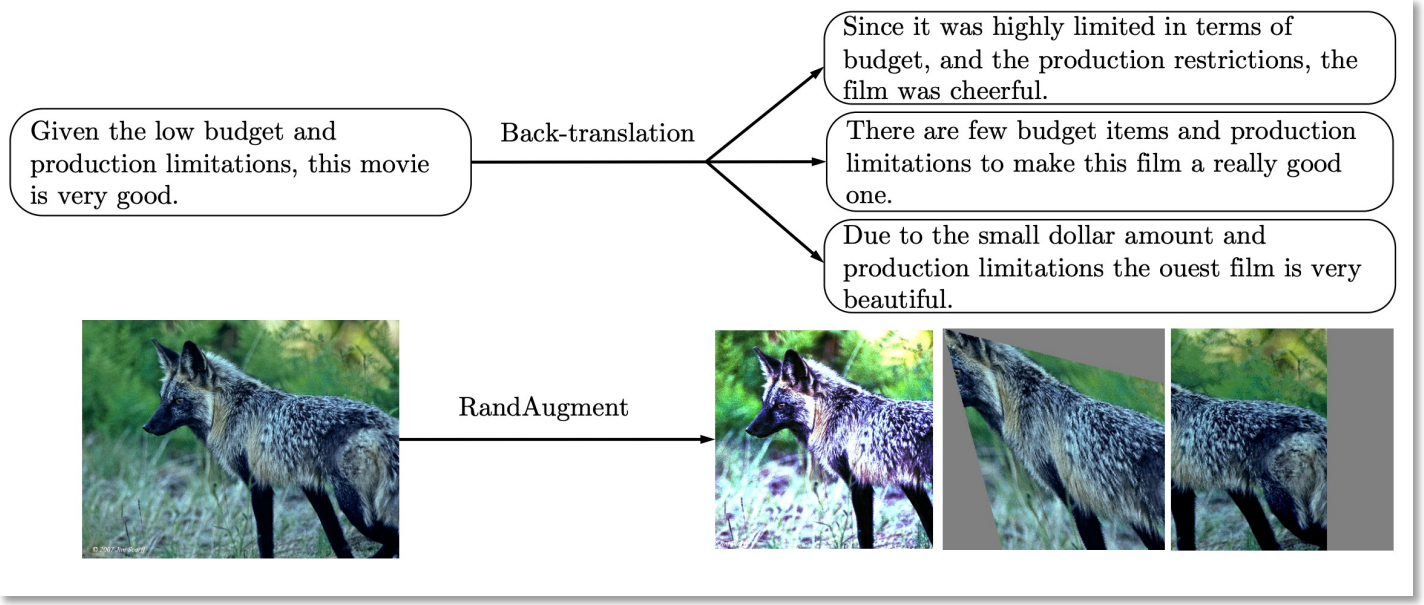
哈尔滨工业大学（深圳）

2022年11月26日

一致性学习的背景

- 对相似样本的输出预测进行正则约束
  - 提升模型的泛化性能
  - 主要应用于半监督学习
- 相似样本的定义
  - 输入/隐藏表示加噪
  - 高效的数据增强
- 优点
  - 😊模型效果提升明显
- 缺点
  - 😭模型收敛速度下降明显





(Xie et al., NeurIPS2020)

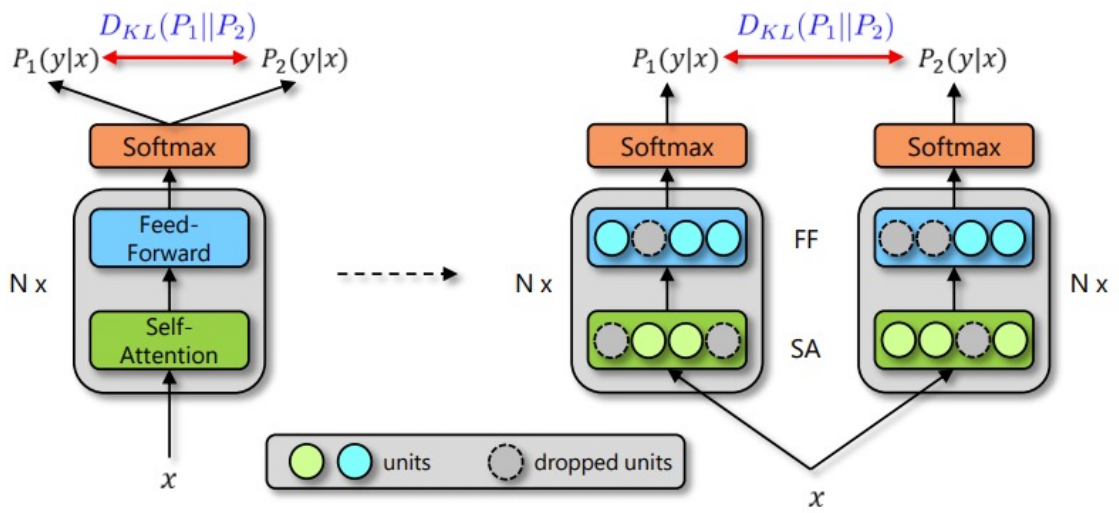
一致性学习在机器翻译中的应用

- 如何应用于**标注数据规模有限**的机器翻译任务
  - ▶ 一致性学习对相似样本的输出预测进行正则约束→输入/隐藏表示加噪（语义弱化）

Original: $\mathbf{P(y|x,\theta)}$

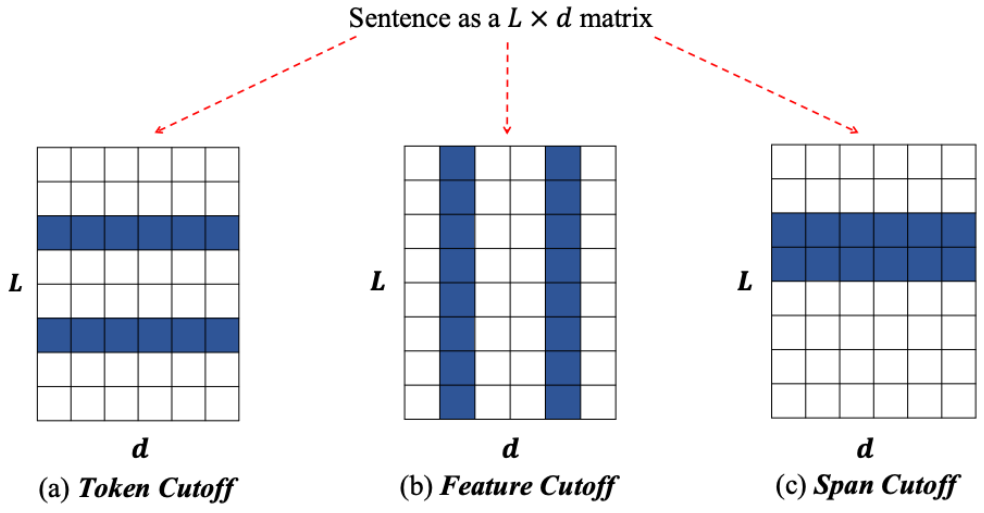
x: los clients y los asociados son enemigos .
y: the clients and the associates are enemies .

RDrop: $\mathbf{P(y|x,\theta_1)}$; $\mathbf{P(y|x,\theta_2)}$; $\mathbf{D(P||P)}$



(Liang et al., NeurIPS 2021)

Cutoff: $\mathbf{P(y|x,\theta_1)}$; $\mathbf{P(y|\tilde{x},\theta_2)}$; $\mathbf{D(P||P)}$



(Shen et al., 2020)

一致性学习在机器翻译中的应用

- **如何应用于标注数据规模有限的机器翻译任务**
  - ▶ 一致性学习对相似样本的输出预测进行正则约束→高效的数据增强（语义等价）

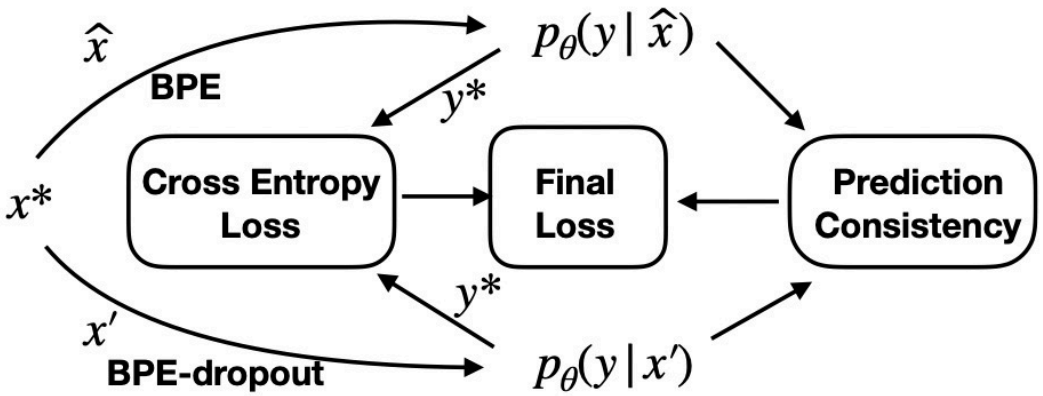
Original: $P(y|x,\theta)$

x: los clients y los asociados son enemigos .
y: the clients and the associates are enemies .

MVR: $P(y|x_1,\theta_1)$; $P(y|x_2,\theta_2)$; $D(P||P)$

x_1: los clients y los aso@@ ciados son en@@ emigos .

x_2: los c@@ lients y los aso@@ ciados son enemigos .

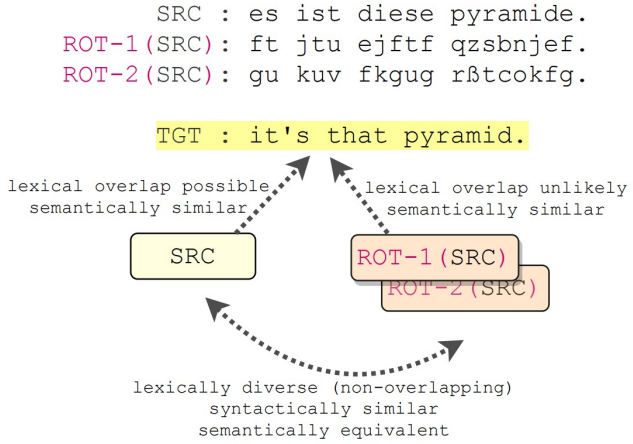


(Wang et al., NAACL 2021)

Cipher: $P(y|x_1,\theta_1)$; $P(y|x_2,\theta_2)$; $D(P||P)$

x_1: los clients y los asociados son enemigos .

x_2: mpt dmjgout z mpt btpdjbept tpo gognjhpt .

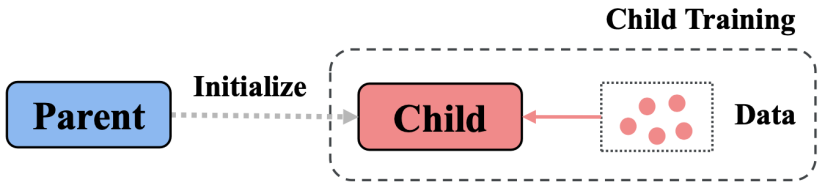


(Kambhatla et al., ACL 2022)

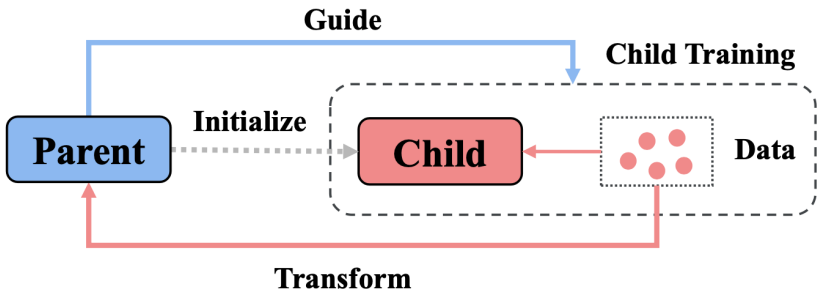
神经机器翻译的一致性表示学习

一致性学习	→	一致性表示学习
输入/隐藏表示加噪→高效的数据增强	→	原始数据与底层特征(之间)的优化
语义弱化→语义等价	→	语义增强

ConsistTL: $P(y|x,\theta_x)$; $P(y|z,\theta_z)$; $D(P||P)$



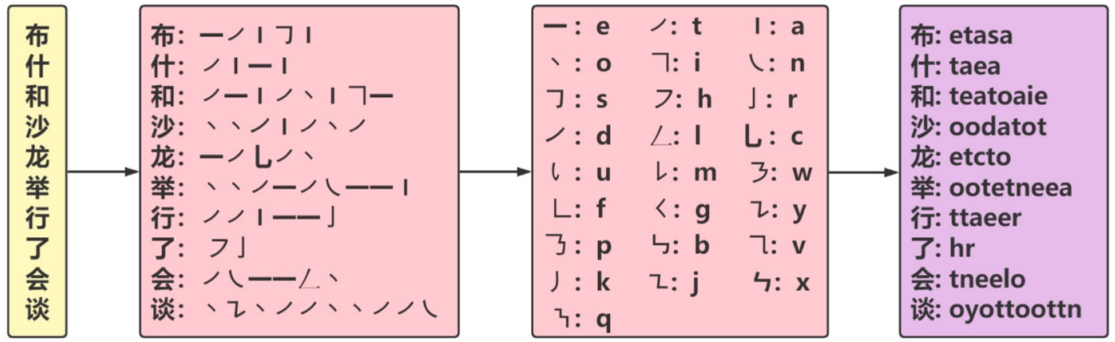
(a) Vanilla TL



(b) ConsistTL

(Li et al., EMNLP 2022)

StrokeNet: $P(y|u_1,\theta_1)$; $P(y|u_2,\theta_2)$; $D(P||P)$



x : 布什和沙龙举行了会谈

u : etasa taea teatoaie oodatot etcto ootetneea ttaeer hr tneelo oyottoottn

u_c : taoho aoto atoaiont iiloaoa tauai iiatastto aaottd rd asttci ipiaaiiaas

y : Bush held a talk with Sharon

(Wang et al., EMNLP 2022)

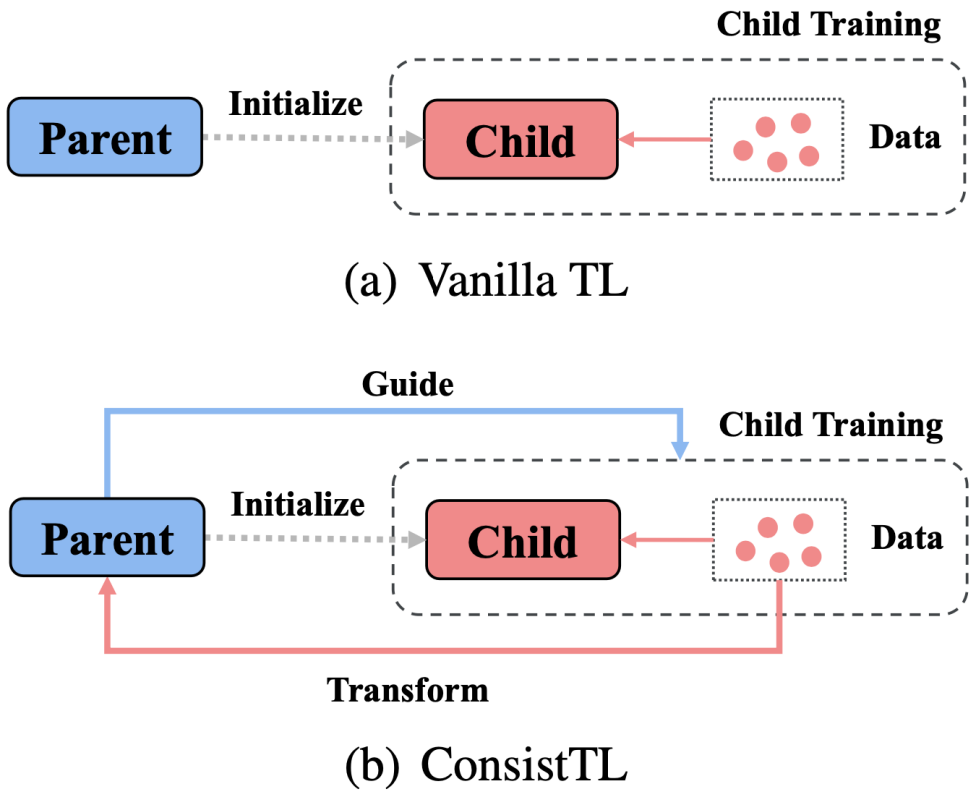
ConsisTL: Modeling Consistency in Transfer Learning for Low-resource Neural Machine Translation

Zhaocong Li[1], Xuebo Liu[2*], Derek F. Wong[1*], Lidia S. Chao[1], Min Zhang[2]
[1]NLP[2]CT Lab, University of Macau
[2]Harbin Institute of Technology, Shenzhen

EMNLP 2022

背景：基于迁移学习的低资源机器翻译

- **低资源机器翻译方法：迁移学习**
  - ▶ 将高资源翻译模型（父模型）上的知识迁移至低资源翻译模型（子模型）
    - ▶ 德语-英语→土耳其-英语
  - **优点**
    - ▶ 😊简单有效、收敛快
  - **缺点**
    - ▶ 😭初始化过程一次性迁移知识
- **ConsistTL**
  - ▶ **目标**
    - ▶ 😊父模型对子模型学习过程持续指导

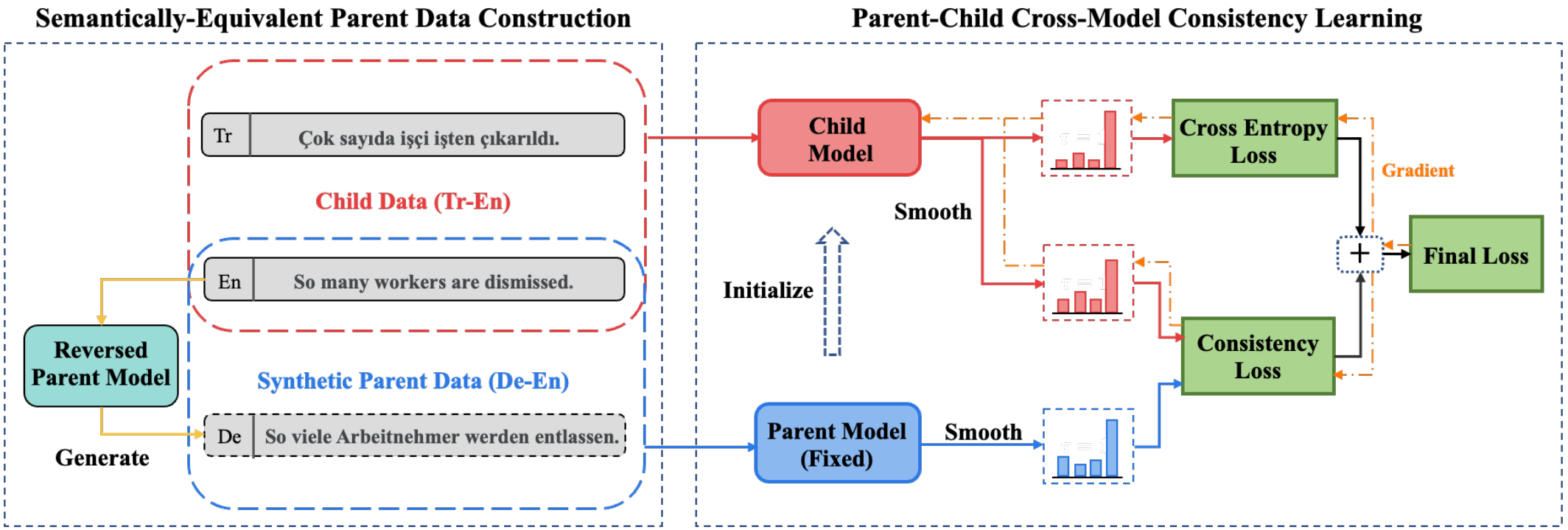


(a) Vanilla TL

(b) ConsistTL

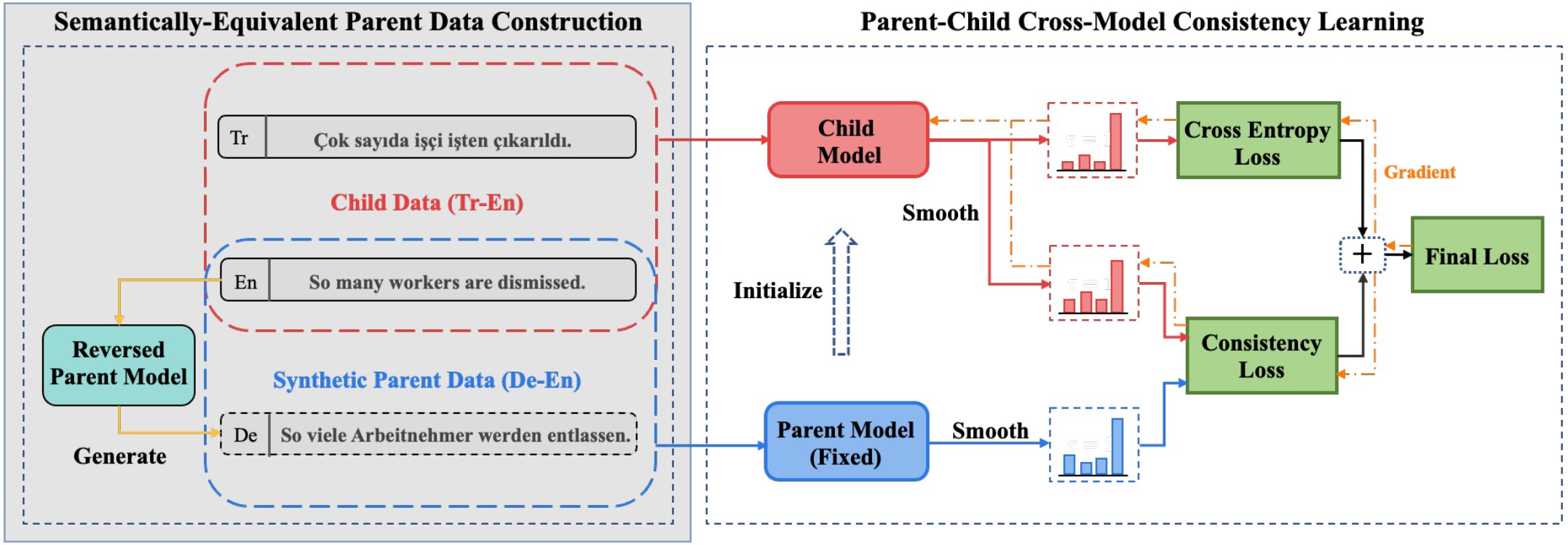
ConsistTL框架图

- 语义等价表示
  - 合成伪数据
- 父模型持续指导
  - 一致性学习



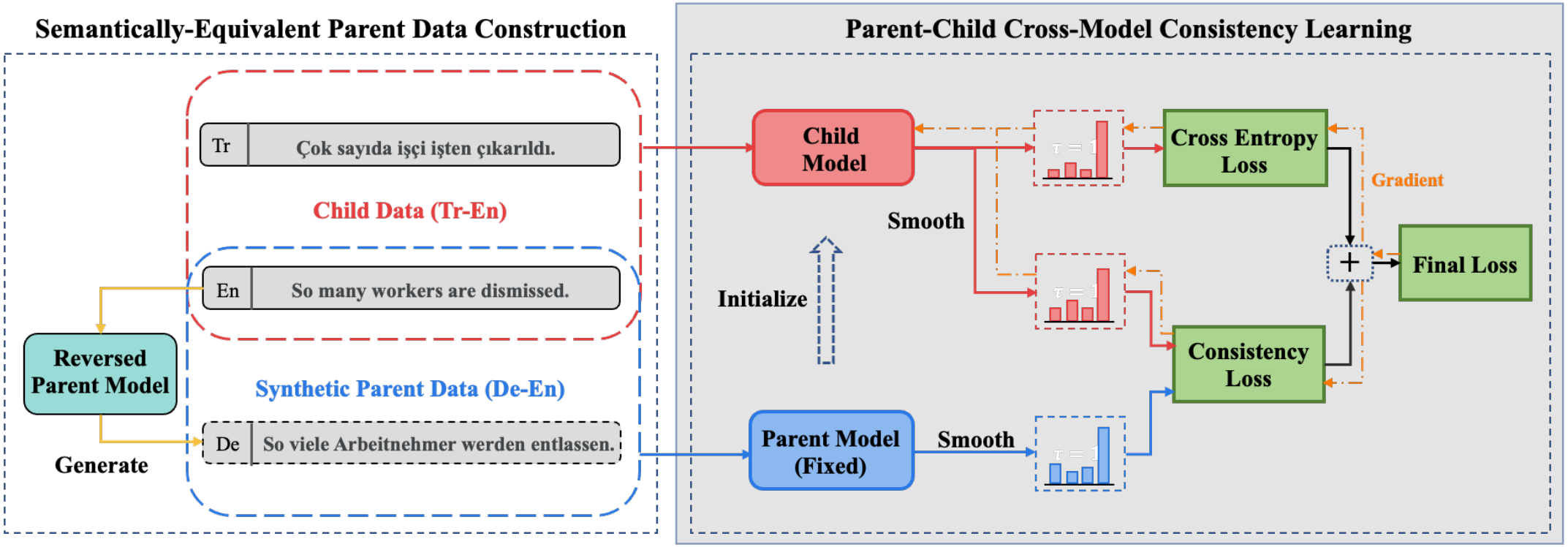
表示：跨语言语义等价样本构造

- **语义等价的父样本构建过程**
  - 目的：构建语义等价样本，为一致性建模做准备
  - 反向父模型将目标语句y进行翻译：y → z



一致性学习：父模型对子模型的持续指导

- **跨模型的一致性学习**
  - 目的：缩小父子模型分布差距，实现父模型的持续指导
  - 通过一致性损失实现： $L_d = D(P(y|x,\theta_x)||P(y|z,\theta_z))$



翻译效果提升

- ConsistTL在以下数据集上的效果

Parent

WMT17 De-En	5.8M

Child

WMT17 Tr-En	200k
Global-Voice Xx-En	8k~30k

Model	Id-En		Ca-En		Hu-En		Pl-En		Tr-En	
	BLEU	BS	BLEU	BS	BLEU	BS	BLEU	BS	BLEU	BS
Existing	12.3	-	20.0	-	5.4	-	18.0	-	18.5	-
Vanilla	1.1	13.2	1.1	15.5	0.9	0.9	1.5	18.9	17.8	51.8
+ TL	13.5	37.7	21.6	51.8	5.9	27.4	19.9	55.3	17.6	51.9
+ TM-TL	18.6	49.9	25.3	58.9	10.6	41.2	21.4	58.2	18.6	53.9
+ ConsistTL	**19.7**	**52.2**	**26.6**	**60.0**	**11.9**	**43.9**	**22.4**	**59.9**	**19.3**	**55.9**

ConsistTL 显著超过基线模型与已有TL方法

不同配置对比

- **一致性损失函数的选择**
  - JS 散度 vs. KL 散度

JS散度效果更优

Tr-En	Consistency Loss	Valid	Test
TL	-	17.8	17.6
TM-TL	-	18.8	18.6
ConsistTL	Kullback-Leibler	19.2	18.9
	Jensen–Shannon	19.8	19.3

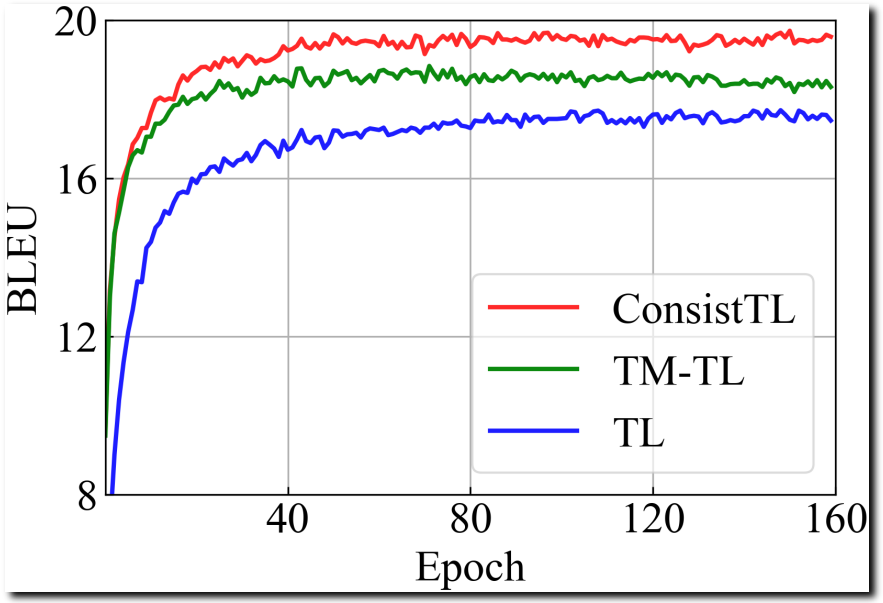
- **伪父样本生成方式的选择**
  - 束/贪婪搜索 vs. 采样

束/贪婪搜索效果更优

Tr-En	Parent Source	Valid	Test
TL	-	17.8	17.6
TM-TL	-	18.8	18.6
ConsistTL	Sampling	19.3	18.8
	Greedy Search	19.8	19.3
	Beam Search	19.8	19.3

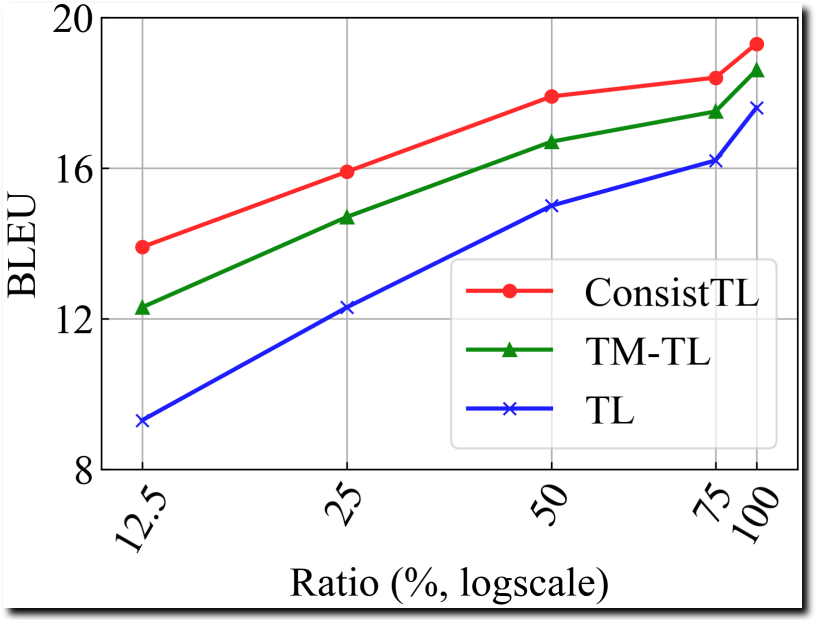
实验分析

▶ 学习效果曲线



ConsistTL 收敛快，上限高

▶ 不同规模的子集上的效果



语料规模越小，ConsistTL效果越明显

与现有的方法的互补性

▶ **Back-Translation**

Tr-En	w/ BT	Valid	Test
TM-TL (Aji et al., 2020)	Y	-	20.6
TM-TL (Ours)	N	18.8	18.6
	Y	22.3	21.6
ConsistTL	N	19.8	19.3
	Y	22.9	22.3

▶ **R-Drop**

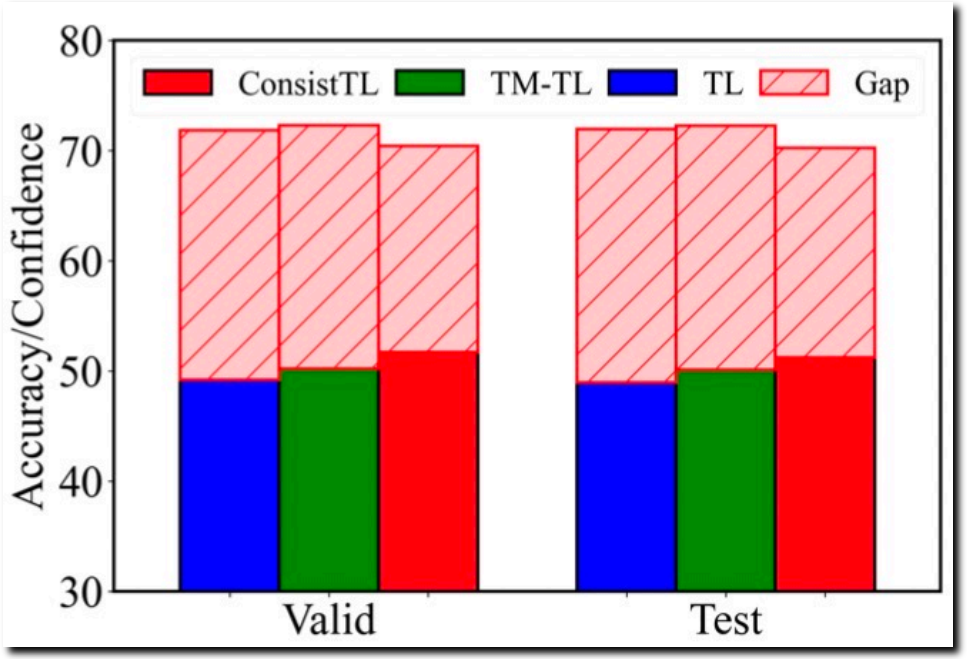
Tr-En	w/ Inner Consist.	Valid	Test
TM-TL	N	18.8	18.6
	Y	20.1	20.0
ConsistTL	N	19.8	19.3
	Y	20.8	20.6

ConsistTL 可以与领先的低资源增强方法互补

父模型预测分布的作用

- **传统迁移学习**
  - 置信度(Confidence)高
  - 翻译精度(Accuracy)低
  - 模型校准能力(Calibration)低

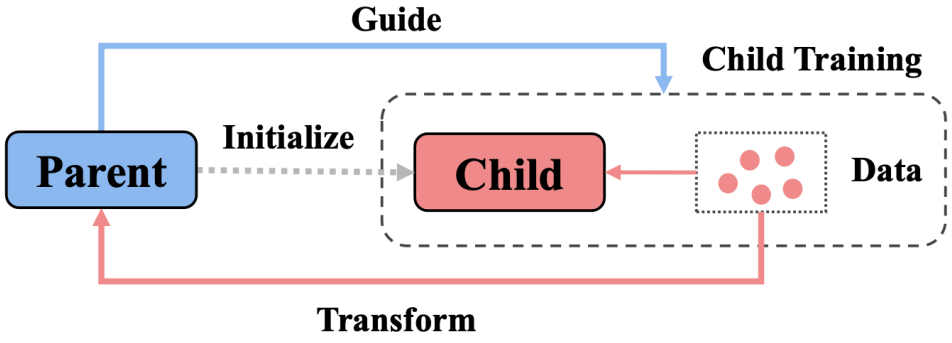
- **ConsistTL**
  - 降低置信度
  - 提升翻译精度
  - 提升模型校准能力



ConsistTL 总结

- **ConsistTL 是一种新的迁移学习方法**
  - ▶ 父模型理解子模型语义表示 → 拓宽**语义等价建模范围**
  - ▶ 父模型在训练中指导子模型学习 → 跨模型**一致性表示学习**
  - ▶ 显著提升子模型的模型校准能力 → 领域知识的**充分利用**

ConsistTL: $P(y|x,\theta_x)$; $P(y|z,\theta_z)$; $D(P||P)$



(b) ConsistTL

Breaking the Representation Bottleneck of Chinese Characters: Neural Machine Translation with Stroke Sequence Modeling

Zhijun Wang, Xuebo Liu*, Min Zhang
Harbin Institute of Technology, Shenzhen

EMNLP 2022

中文神经机器翻译的两个瓶颈

- **中文神经机器翻译通常将汉字作为最小表示单元**

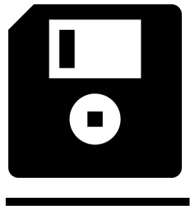
  - 两个表示瓶颈→汉字内部特征信息的缺失、中文词表规模过大

学习瓶颈

扎　拉　打　抬

都有用手做动作的涵义

😭模型难以学习到汉字丰富的内部特征信息（如部首等）



参数瓶颈

大规模多语预训练模型

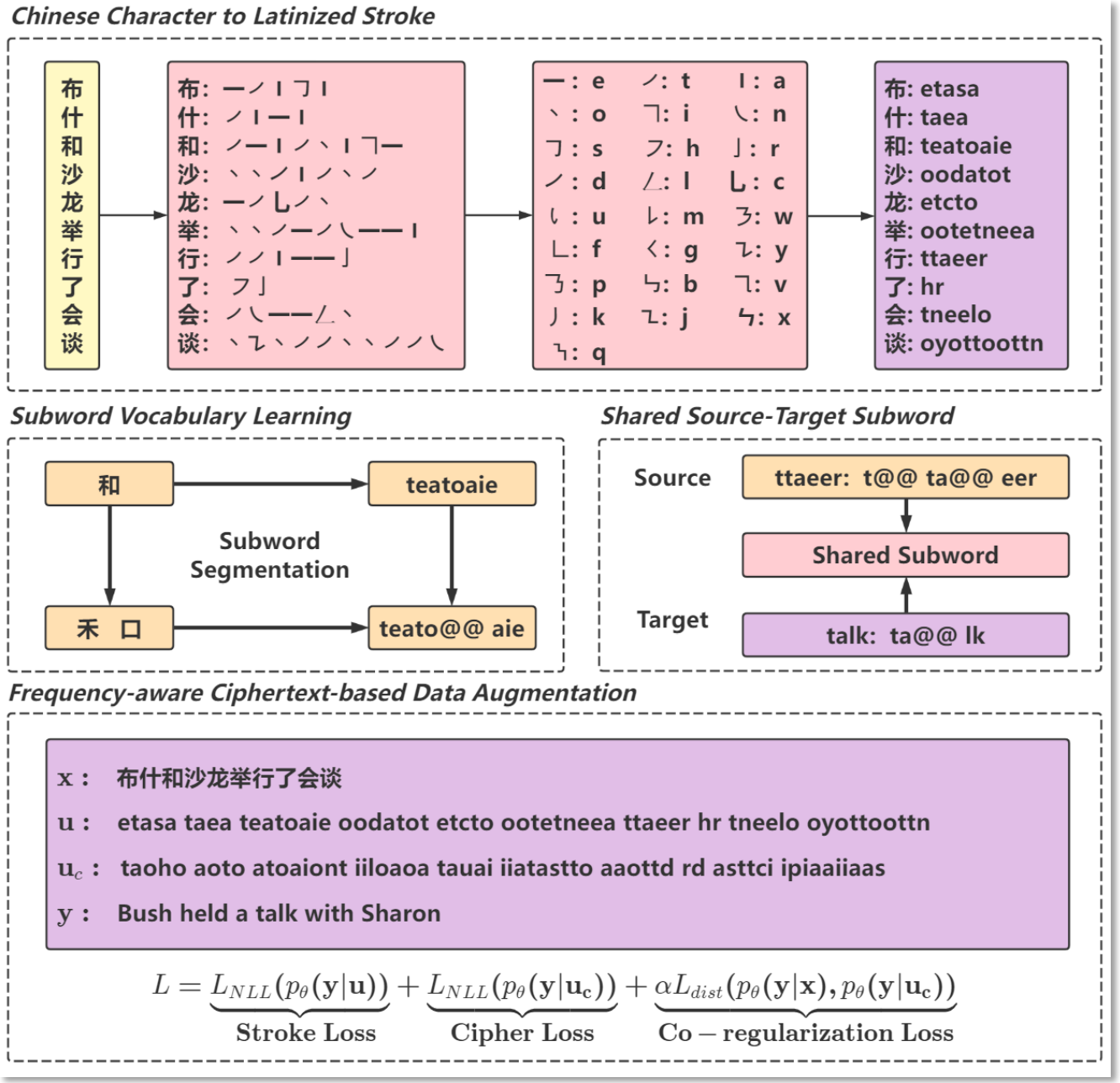
mBART-CC50的词表

所有词的总数	250K
中文词的数目	20K

😭每个中文字词都需要用唯一的向量表示

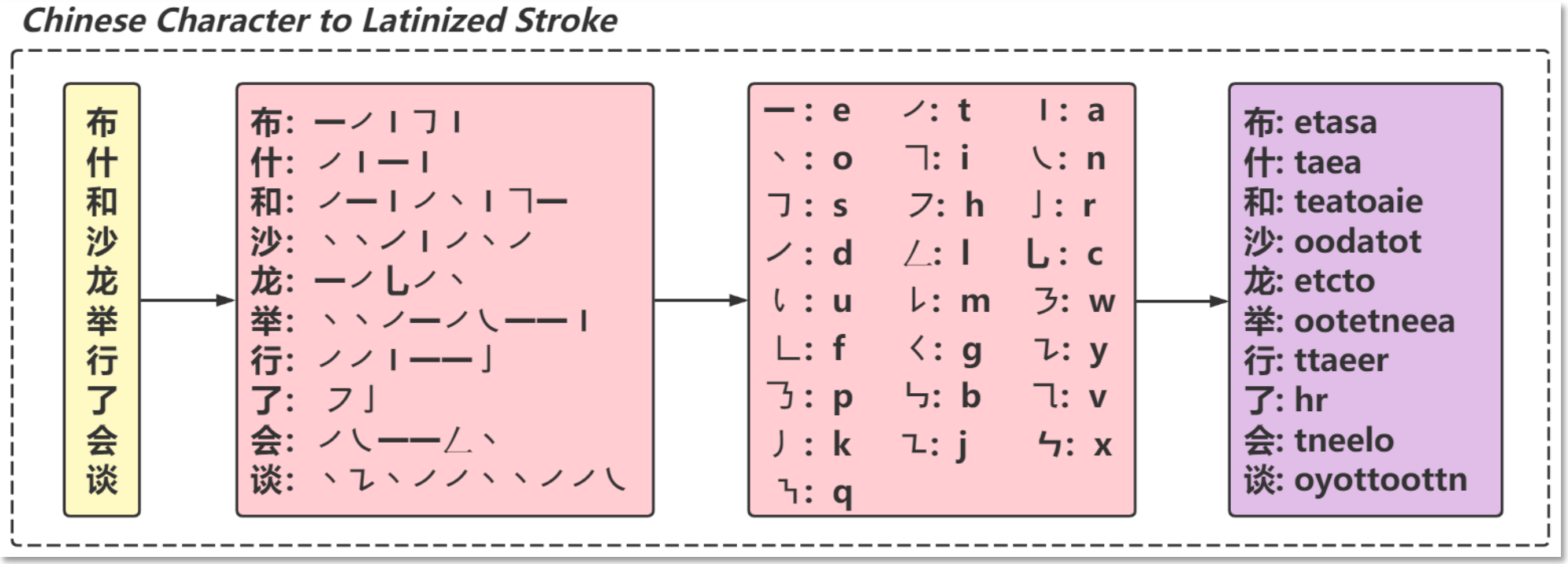
StrokeNet整体框架图

- **拉丁化笔画序列建模**
  - 频率映射方法
- **打破学习瓶颈**
  - 子词学习
  - 频率感知密文数据增强
- **打破参数瓶颈**
  - 共享的源端目标端表示



拉丁化笔画序列建模

- 将**汉字**转换成**拉丁化的笔画序列**
  - ▶ 将汉字映射为唯一对应的笔画序列
  - ▶ 将每种笔画映射为唯一对应的小写拉丁字母



拉丁化笔画序列建模

- **频率映射方法**

  - 构建从25种笔画到26种小写拉丁字母的一一映射(`z` 不使用)

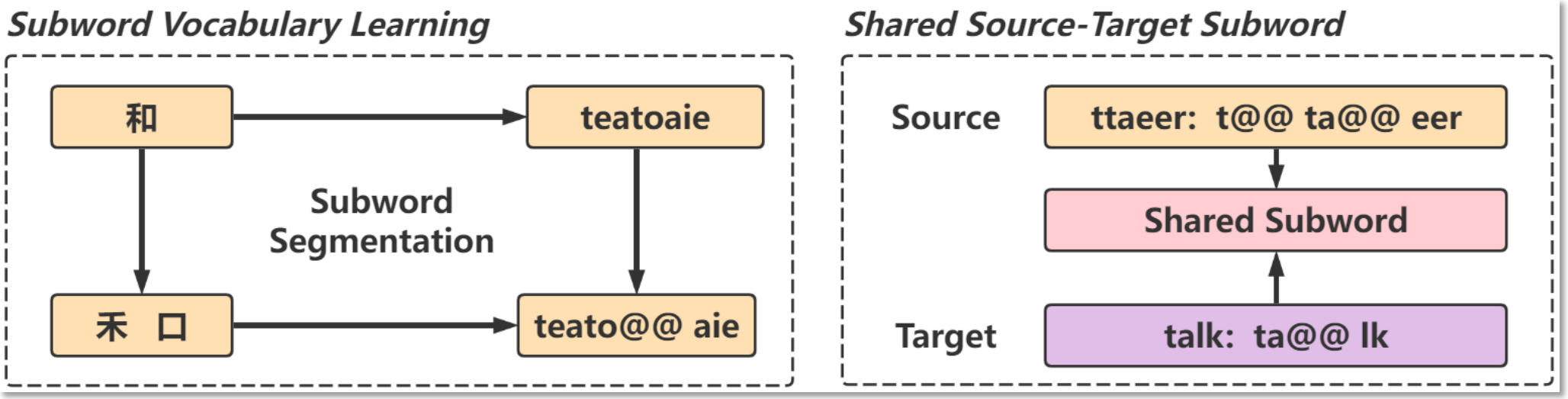
Stroke	一	丿	丨	丶	…	乙	乚	乛	
Freq	27.9	16.9	16.6	13.2		0.068	0.028	0.00012	
Letter	e	t	a	o		j	x	q	z
Freq	12.7	9.1	8.2	7.5		0.153	0.15	0.095	0.074

小写拉丁字母的频率来源于Wikipedia.

笔画的频率来源于对WMT 2017 Zh-En数据集的统计

基于拉丁化笔画序列的中文NMT

- 将中文的拉丁化笔画序列应用到神经机器翻译任务中(NMT)→StrokeNet
  - ▶ 打破学习瓶颈→使用子词学习来提取汉字内部特征信息
  - ▶ 打破参数瓶颈→学习拉丁化笔画序列与英文的共享子词+权重共享



基于拉丁化笔画序列的中文NMT

- 将中文的拉丁化笔画序列应用到神经机器翻译任务中(NMT)→ StrokeNet
  - ▶ 打破学习瓶颈→频率感知的密文数据增强方法
  - ▶ 明文和密文的一致性表示学习

Frequency-aware Ciphertext-based Data Augmentation

x： 布什和沙龙举行了会谈

u： etasa taea teatoaie oodatot etcto ootetneea ttaeer hr tneelo oyottoottn

u_c： taoho aoto atoaiont iiloaoa tauai iiatastto aaottd rd asttci ipiaaiiaas

y： Bush held a talk with Sharon

$$L = \underbrace{L_{NLL}(p_\theta(\mathbf{y}|\mathbf{u}))}_{\text{Stroke Loss}} + \underbrace{L_{NLL}(p_\theta(\mathbf{y}|\mathbf{u_c}))}_{\text{Cipher Loss}} + \underbrace{\alpha L_{dist}(p_\theta(\mathbf{y}|\mathbf{x}), p_\theta(\mathbf{y}|\mathbf{u_c}))}_{\text{Co-regularization Loss}}$$

翻译效果提升

- StrokeNet 在以下数据集上的效果：

WMT17 Zh-En	20M
NIST Zh-En	1.25M
IWSLT17 Ja-En	223K

Model	IWSLT Ja-En	WMT17 Zh-EN
Shared-Private (Liu et al., 2019)	12.4	-
Norm (Liu et al., 2020a)	-	25.3
Prior (Chen et al., 2021b)	-	25.5
Vanilla	12.0	24.4
StrokeNet	**13.1**	**26.5**

Model	Parameters		Performance (BLEU)					
	Total	Embedding	Valid	MT02	MT03	MT04	MT08	ALL
Shared-Private (Liu et al., 2019)	63M	19M	42.6	43.7	42.0	44.5	33.8	41.6
AdvGen (Cheng et al., 2019)	95M	39M	47.0	47.0	46.5	47.4	37.4	-
AdvAug (Cheng et al., 2020)	95M	39M	49.2	49.0	48.0	48.9	39.6	-
Manifold (Chen et al., 2021a)	83M	46M	49.4	49.6	**50.3**	49.5	39.2	-
Vanilla	80M	36M	47.4	47.7	47.5	47.7	38.1	45.5
StrokeNet	**59M**	**15M**	**49.7**	**50.6**	49.8	**50.1**	**40.7**	**48.0**

在WMT 2017 Zh-En数据集上取得了SOTA(26.5 BLEU)结果(不使用额外数据)

参数规模减小

- StrokeNet 在以下数据集上的效果：

▶ 不同语言对参数比较

Model		NIST Zh-En	WMT17 Zh-En	IWSLT Ja-En
Vanilla	Src	40K	50K	33K
	Trg	30K	37K	30K
StrokeNet		29K	50K	28K

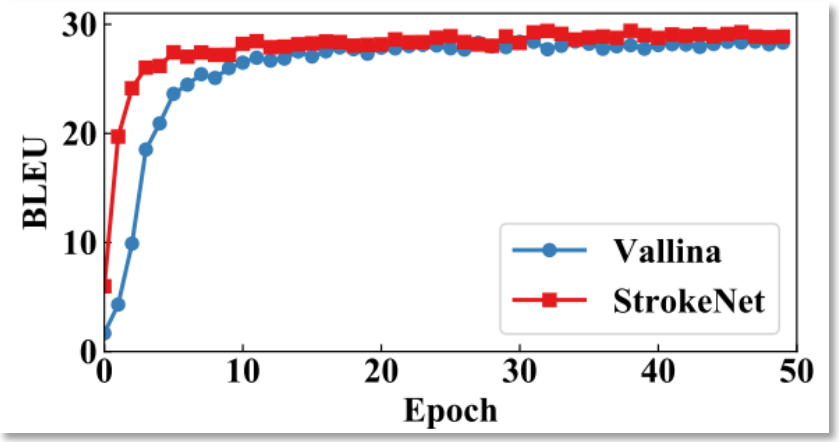
▶ 不同模型参数比较

Model	Parameters	
	Total	Embedding
Shared-Private (Liu et al., 2019)	63M	19M
AdvGen (Cheng et al., 2019)	95M	39M
AdvAug (Cheng et al., 2020)	95M	39M
Manifold (Chen et al., 2021a)	83M	46M
Vanilla	80M	36M
StrokeNet	**59M**	**15M**

参数高效: 词嵌入向量参数规模显著减小

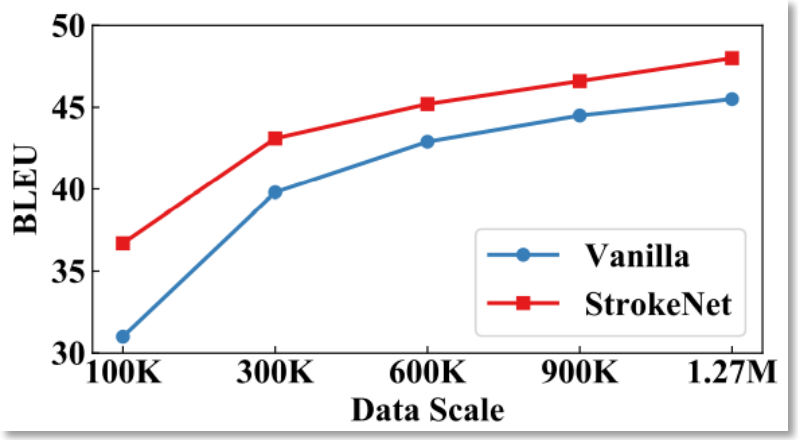
实验分析

▶ 学习效果曲线



StrokeNet 学习更好更快

▶ 不同规模的子集上的效果



StrokeNet在不同数据规模下都有效

▶ 消融实验

- ▶ 频率映射方法表现优于随机映射
- ▶ 频率感知的密文数据增强方法起重要作用
- ▶ 共享子词技术很好地融入了 StrokeNet

Model	Embedding	BLEU
Vanilla	36M	45.5
StrokeNet with Random Mapping	15M	47.5
StrokeNet with Frequency Mapping	15M	48.0
StrokeNet without Cipher DA	15M	45.9
StrokeNet without Shared Voc.	21M	47.7

StrokeNet总结

- StrokeNet 是一种全新的中文表示方法
  - 使用汉字的拉丁化笔画序列 → 中文底层特征表示的优化
  - 打破了中文NMT的表示学习瓶颈 → 明文和密文的一致性表示学习
  - 打破了中文NMT的参数瓶颈 → 源端和目标端的共享表示学习
  - 中文语言符号结构的探索 → 对非字母语言的启示

https://github.com/zjwang21/StrokeNet

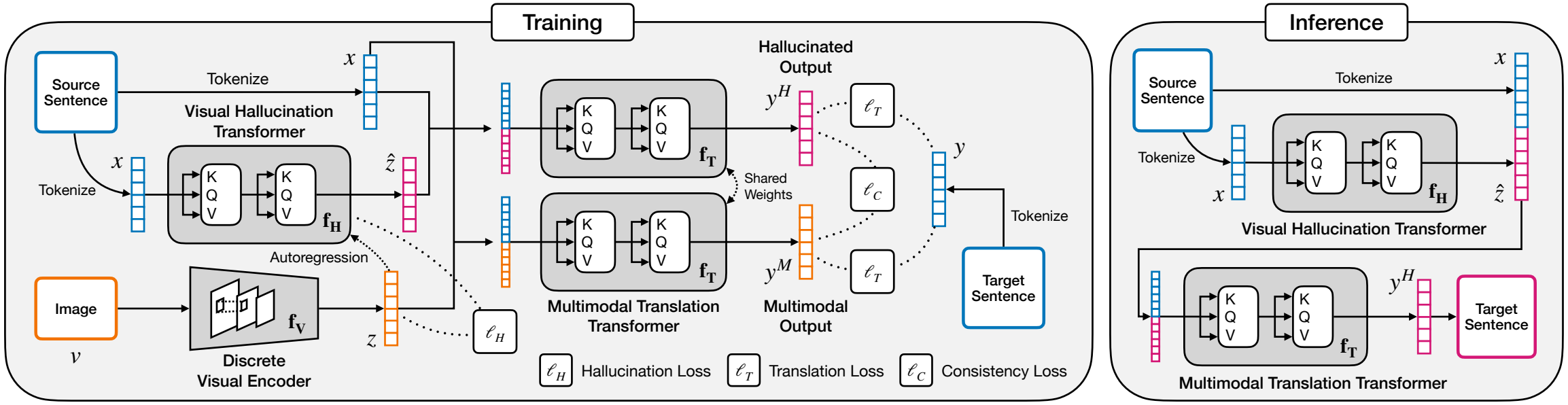
一致性表示学习展望与未来

一致性学习 → 一致性表示学习

输入/隐藏表示加噪→高效的数据增强 → 原始数据与底层特征(之间)的优化

语义弱化→语义等价 → 语义增强→跨模态语义增强→模态融合

VALHALLA: $P(y|x,z,\theta_1)$; $P(y|x,\hat{z},\theta_2)$; $D(P||P)$



(Li et al., CVPR 2022)

一致性表示学习总结

一致性学习 → **一致性表示学习**

输入/隐藏表示加噪→高效的数据增强 → 原始数据与底层特征(之间)的优化

语义弱化→语义等价 → 语义增强→跨模态语义增强

RDrop: $P(y|x,\theta_1)$; $P(y|x,\theta_2)$; $D(P||P)$

(Liang et al., NeurIPS 2021)

Cutoff: $P(y|x,\theta_1)$; $P(y|\tilde{x},\theta_2)$; $D(P||P)$

(Shen et al., 2020)

MVR: $P(y|x_1,\theta_1)$; $P(y|x_2,\theta_2)$; $D(P||P)$

(Wang et al., NAACL 2021)

Cipher: $P(y|x_1,\theta_1)$; $P(y|x_2,\theta_2)$; $D(P||P)$

(Kambhatla et al., ACL 2022)

ConsistTL: $P(y|x,\theta_x)$; $P(y|z,\theta_z)$; $D(P||P)$

(Li et al., EMNLP 2022)

StrokeNet: $P(y|u_1,\theta_1)$; $P(y|u_2,\theta_2)$; $D(P||P)$

(Wang et al., EMNLP 2022)

VALHALLA: $P(y|x,z,\theta_1)$; $P(y|x,\hat{z},\theta_2)$; $D(P||P)$

(Li et al., CVPR 2022)

参考文献

[1] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, Quoc V. Le. Unsupervised Data Augmentation for Consistency Training. NeurIPS 2020.

[2] Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu. R-Drop: Regularized Dropout for Neural Networks. NeurIPS 2021.

[3] Dinghan Shen, Mingzhi Zheng, Yelong Shen, Yanru Qu, Weizhu Chen. A Simple but Tough-to-Beat Data Augmentation Approach for Natural Language Understanding and Generation. arXiv 2020.

[4] Xinyi Wang, Sebastian Ruder, Graham Neubig. Multi-view Subword Regularization. NAACL 2021.

[5] Nishant Kambhatla, Logan Born, Anoop Sarkar. CipherDAug: Ciphertext based Data Augmentation for Neural Machine Translation. ACL 2022.

[6] Zhaocong Li, Xuebo Liu, Derek F. Wong, Lidia S. Chao and Min Zhang. ConsistTL: Modeling Consistency in Transfer Learning for Low-Resource Neural Machine Translation. EMNLP 2022.

[7] Zhijun Wang, Xuebo Liu, Min Zhang. Breaking the Representation Bottleneck of Chinese Characters: Neural Machine Translation with Stroke Sequence Modeling. EMNLP 2022.

[8] Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Chen, Rogerio Feris, David Cox, Nuno Vasconcelos. VALHALLA: Visual Hallucination for Machine Translation. CVPR 2022.