低资源条件下的神经机器翻译方法研究

答辩人：米尔阿迪力江·麦麦提

导　师：孙茂松教授

清华大学计算机系智能技术与系统国家重点实验室

内容提要

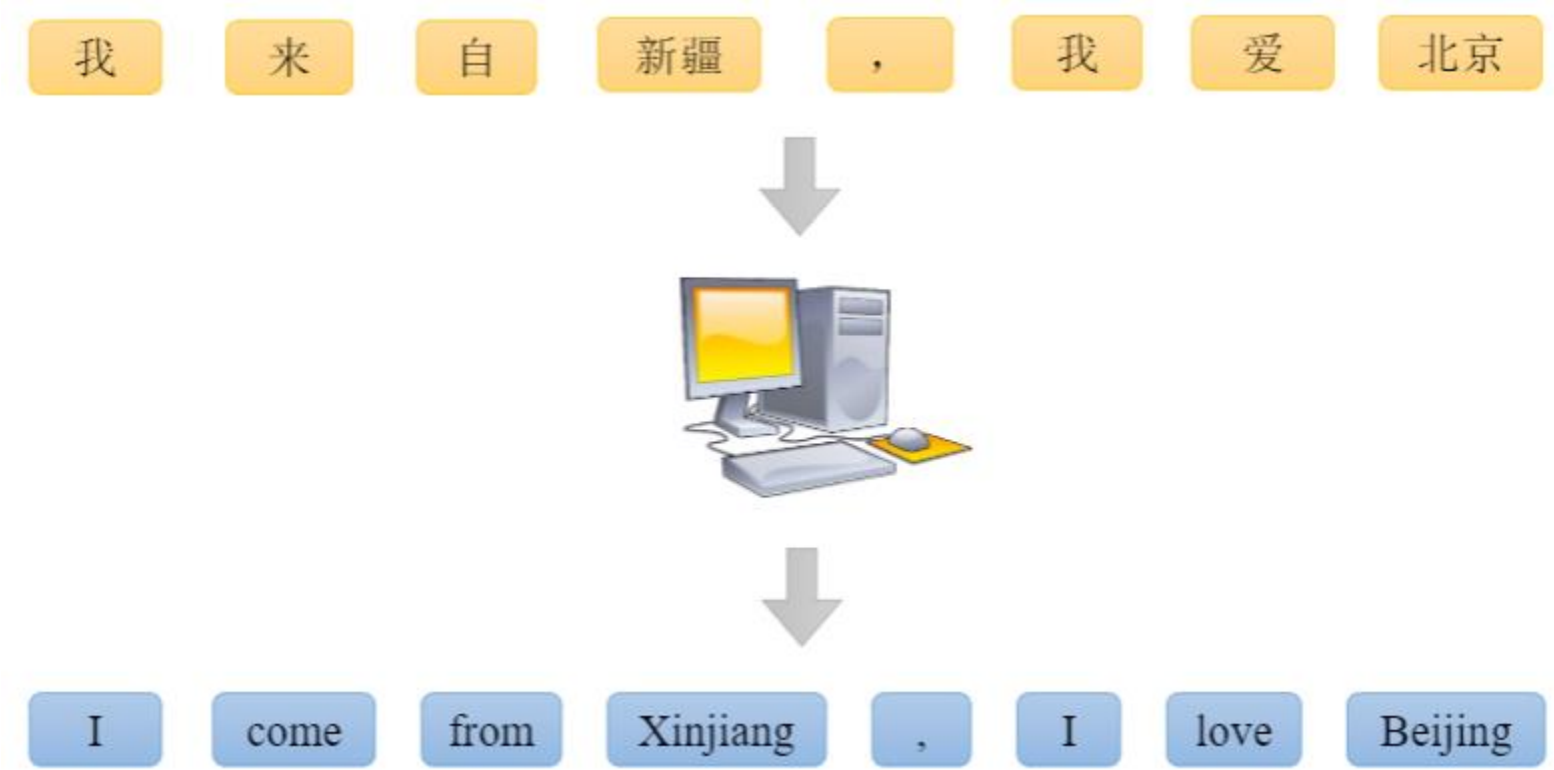
- 研究背景及意义
- 相关工作及研究现状
- 面临的问题与挑战
- 研究工作
  - 基于自监督方法的预训练词切分模型
  - 基于迁移学习的低资源语言神经机器翻译
  - 基于数据增强的低资源语言神经机器翻译
- 总结与展望

内容提要

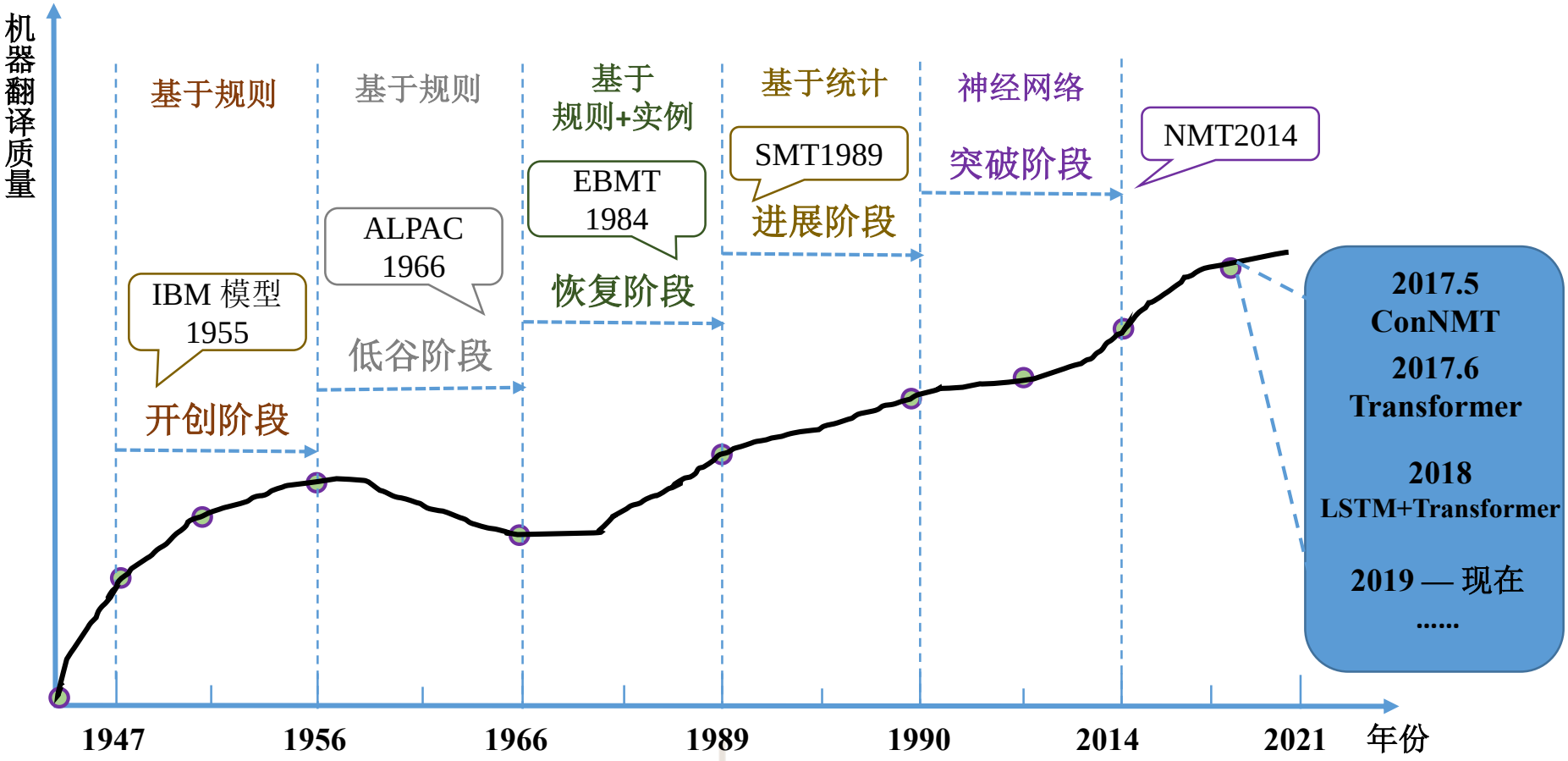
- 研究背景及意义
- 相关工作及研究现状
- 面临的问题与挑战
- 研究工作
  - 基于自监督方法的预训练词切分模型
  - 基于迁移学习的低资源语言神经机器翻译
  - 基于数据增强的低资源语言神经机器翻译
- 总结与展望

机器翻译

- 用计算机将一种序列转化为另一种序列

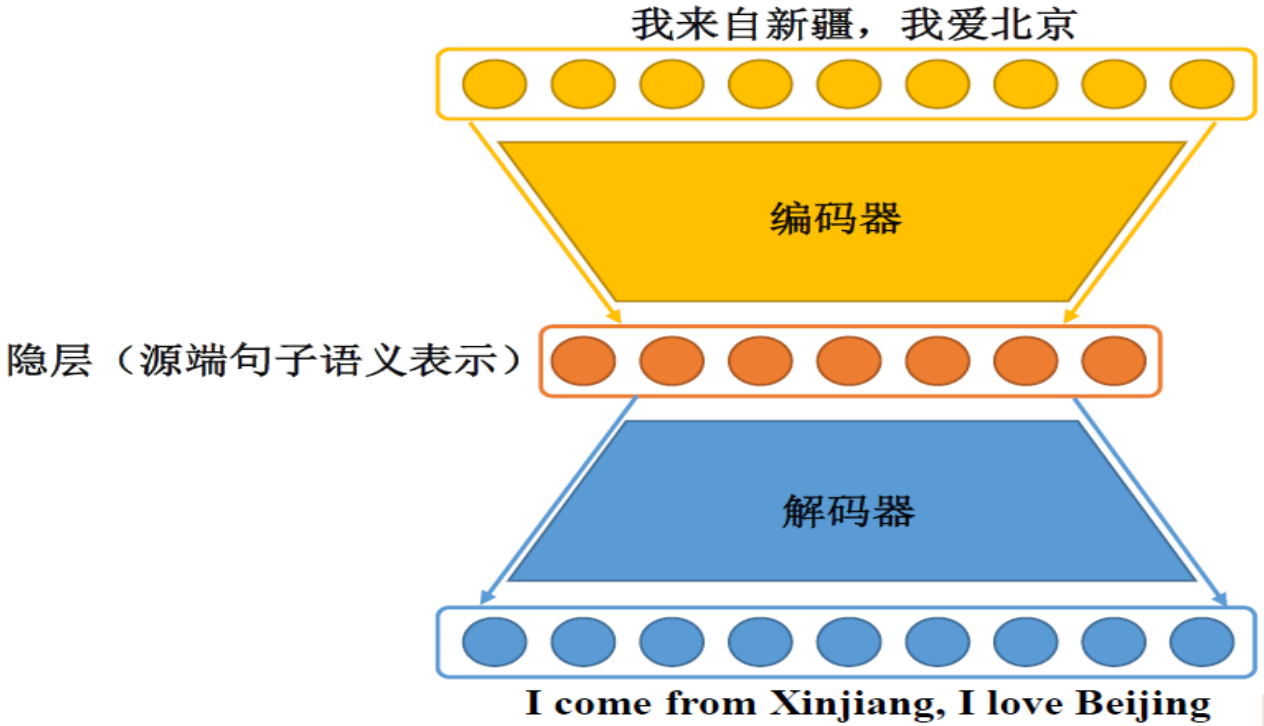


研究背景 --- 历史


机器翻译质量
基于规则
基于规则
基于
规则+实例
基于统计
神经网络
IBM 模型
1955
ALPAC
1966
EBMT
1984
SMT1989
NMT2014
开创阶段
低谷阶段
恢复阶段
进展阶段
突破阶段
2017.5
ConNMT
2017.6
Transformer
2018
LSTM+Transformer
2019 — 现在
.....
1947
1956
1966
1989
1990
2014
2021
年份

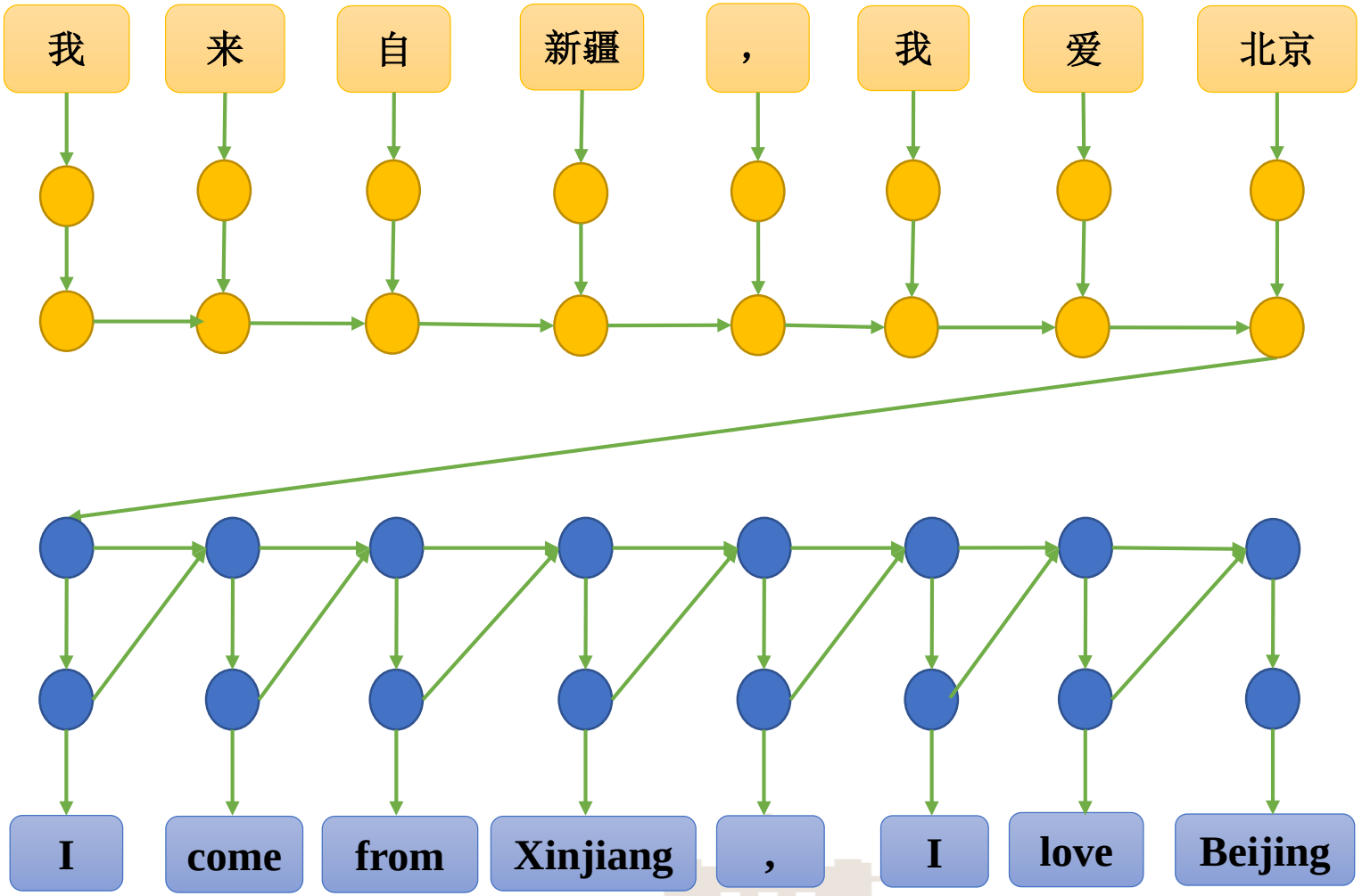
神经网络方法 --- 编码器-解码器

- 建模方式：完全利用神经网络建模，不需要特征工程。
- 表示方式：采用连续的向量表示，SMT离散地表示词语。
- 上下文与词之间的依赖：利用隐层之间的连接操作来实现。



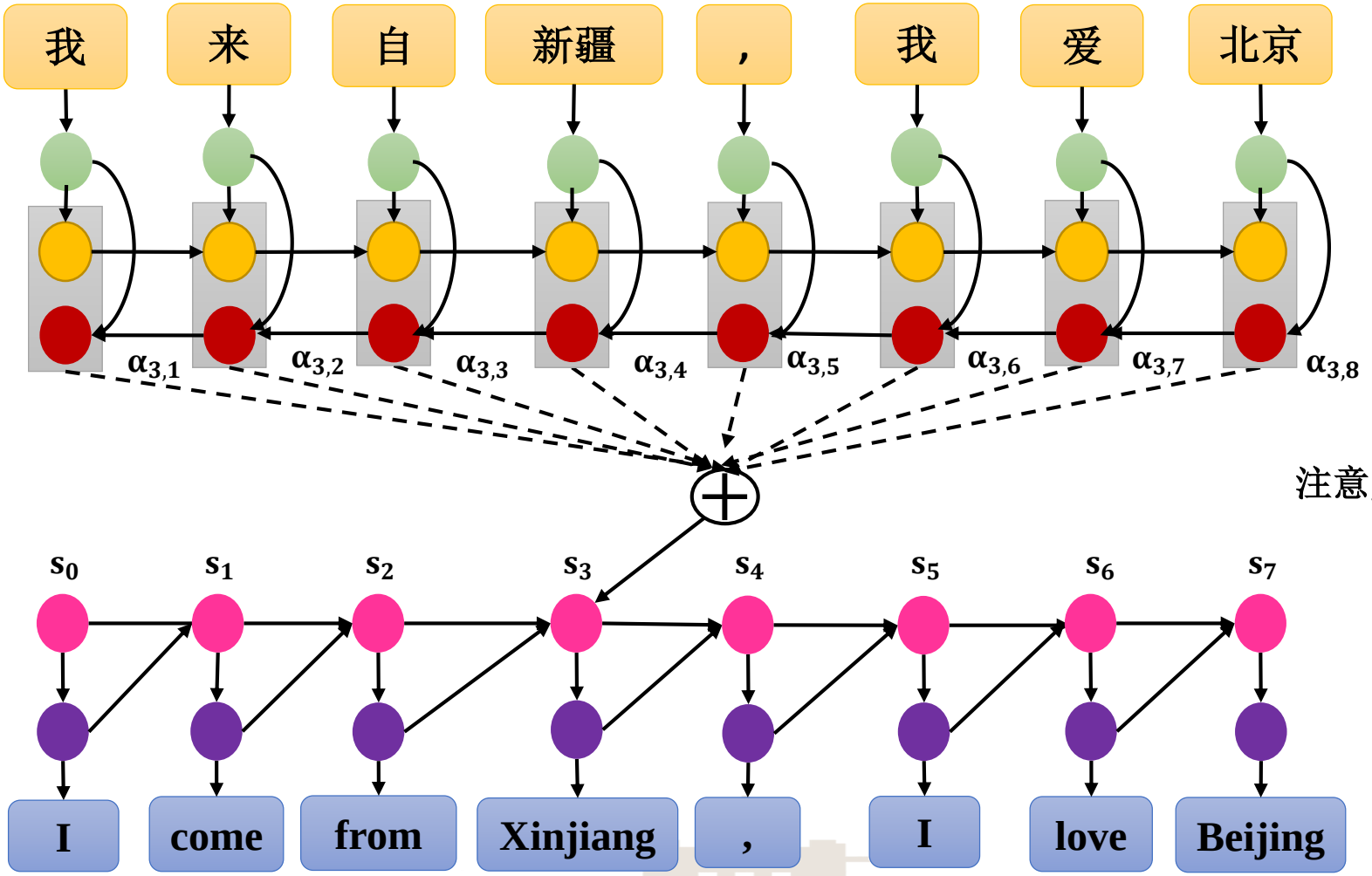
(Blunsom et al., 2013)

神经网络方法 --- 循环神经网络



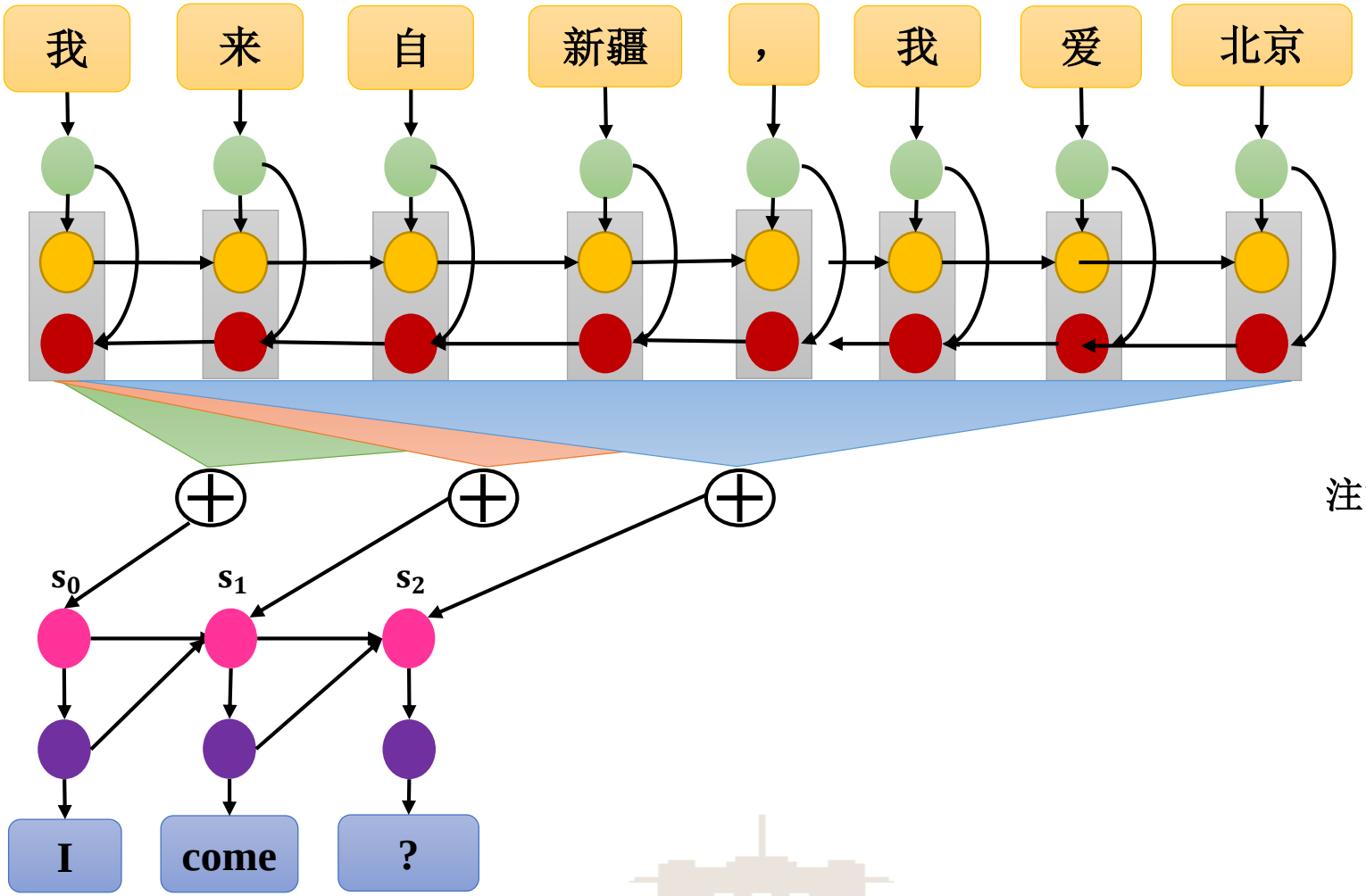
(Sutskever et al., 2014, Cho et al.,2014)

神经网络方法 --- 注意力机制



(Bahdanau et al., 2015)

注意力机制 --- 工作流程



源端句子

词向量

RNN
编码器

注意力机制模型

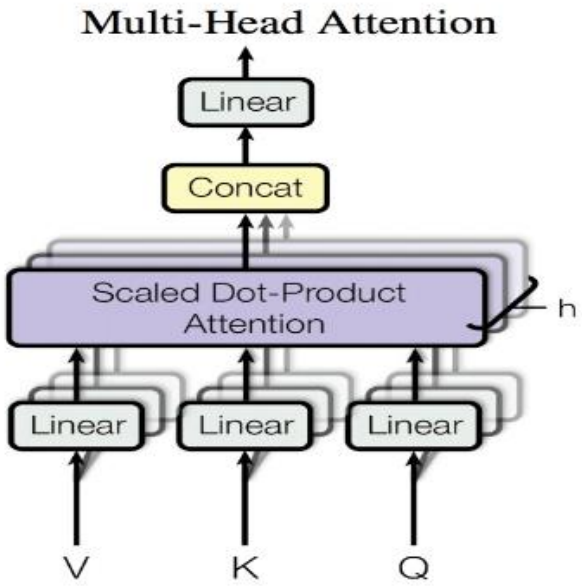
RNN
解码器

读出层

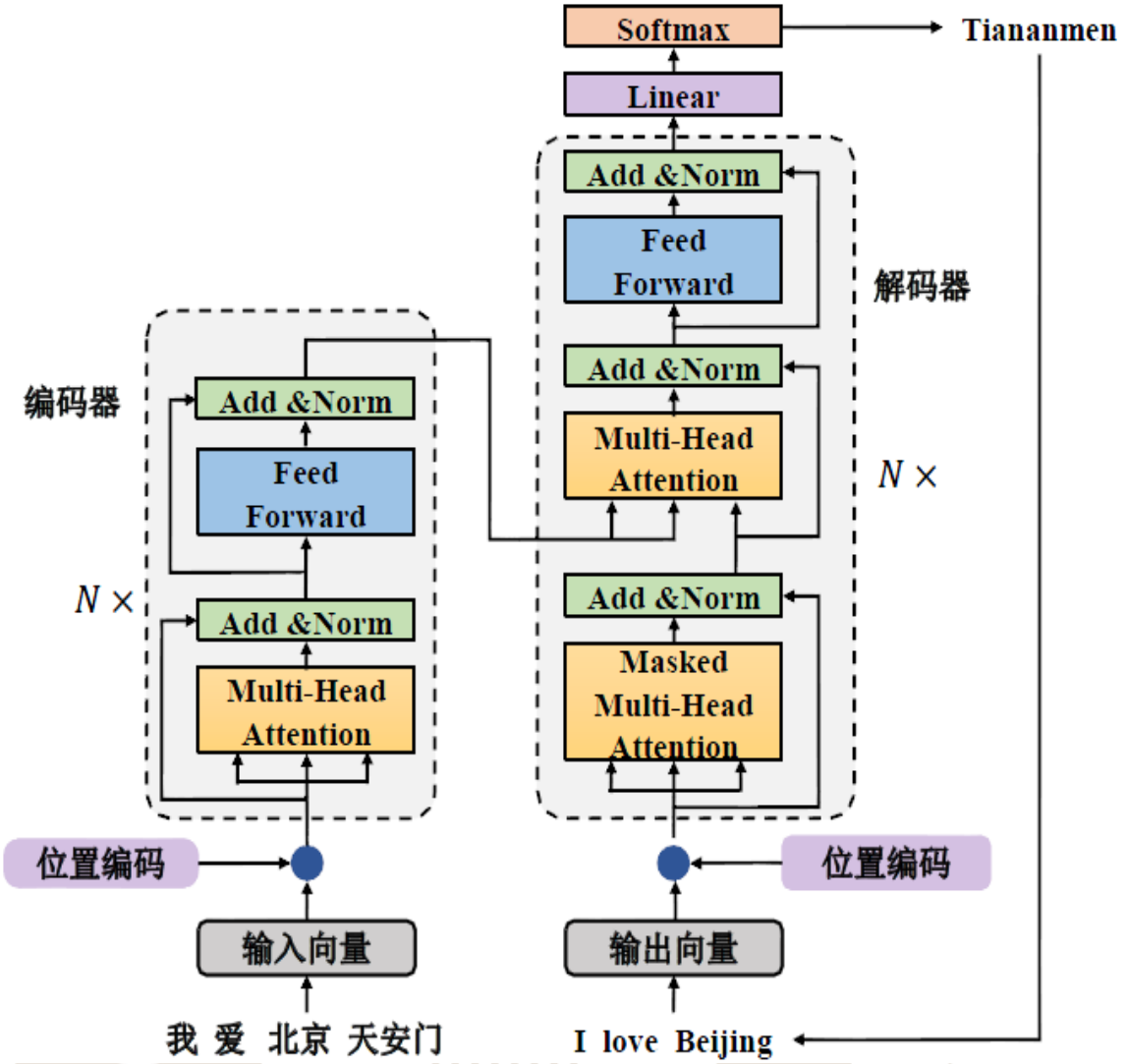
目标端句子

(Bahdanau et al., 2015)

神经网络方法 --- 自注意力机制



多头注意力机制



(Vaswani et al., 2017)

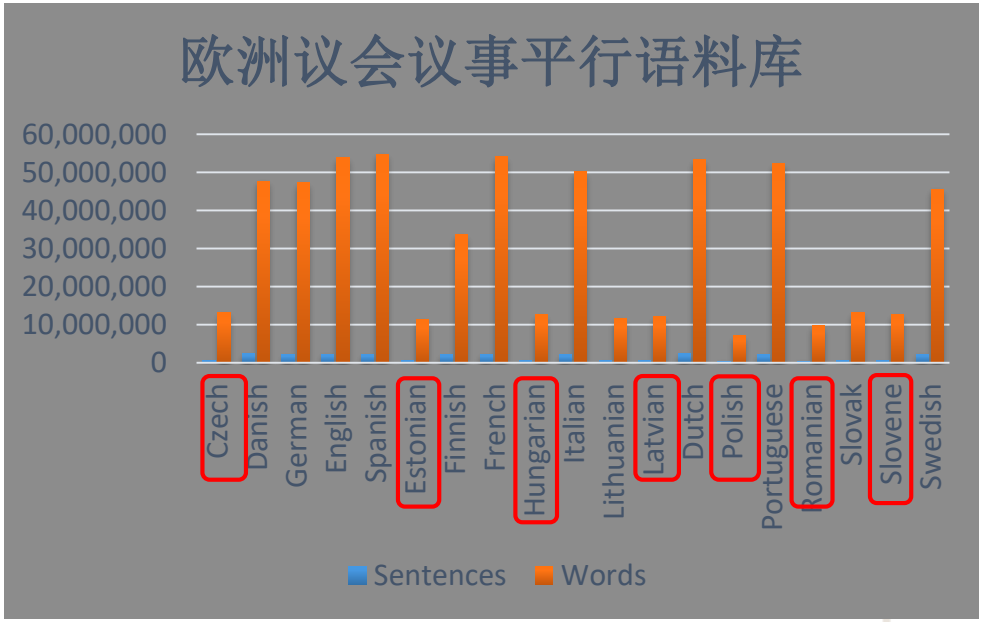
低资源NMT研究意义

- 学术价值
  - 机器翻译是以数据驱动为核心的任务，对平行语料的依赖比较强。因此在高资源语言对上性能非常不错，但在资源匮乏的语言对上性能不佳。

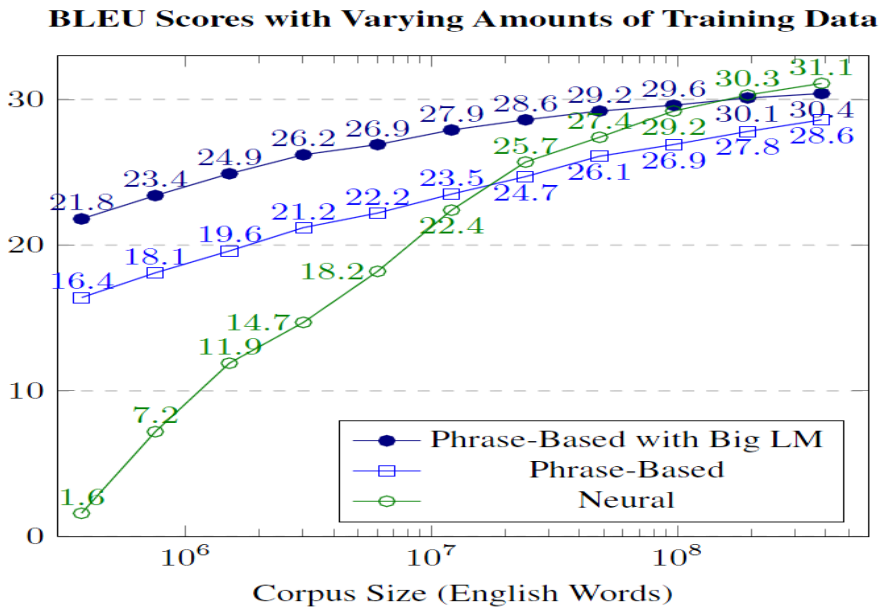
- 应用价值
  - 在“一带一路”工作中，与65个沿线国家特别是中/东/西亚国家之间的政策沟通、贸易畅通、资金融通、设施联通、民心相通与文化交流上有迫切需求。

学术价值 --- 低资源语言NMT

- 神经网络机器翻译在低资源语言场景下训练模型学习曲线更陡峭。
- 无论是SMT还是NMT都是数据驱动的方法，高度依赖大规模平行语料。



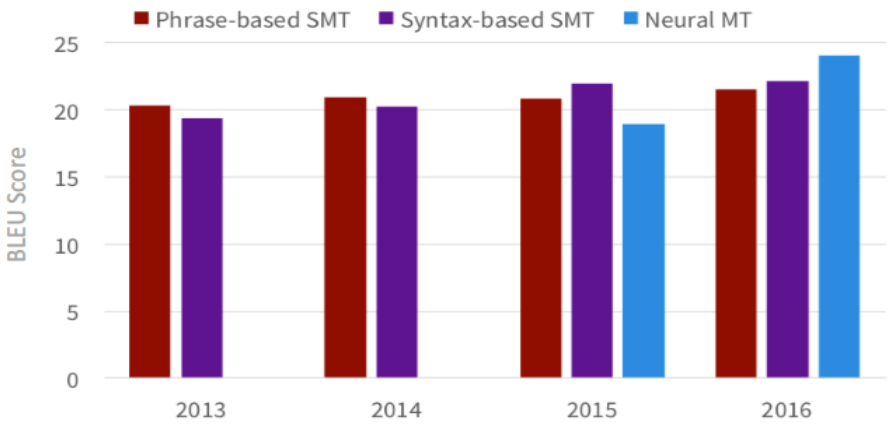
(Koehn et al., 2005)



(Koehn & Knowles., 2017)

学术价值 --- 神经机器翻译

En-De WMT2013 新闻数据集 Cased BLEU score 比较



(Sennrich et al., 2016)

图片来源： http://www.meta-net.eu/events/meta-forum-2016/slides/09_sennrich.pdf

中文分词方法对NMT的影响

分词方法	BLEU (Zh – En)
CHAR	21.16
TEACHER	23.51+2.35
CRF	23.37+2.21
CONPRUNE	23.71+2.55

(Huang et al., 2021)

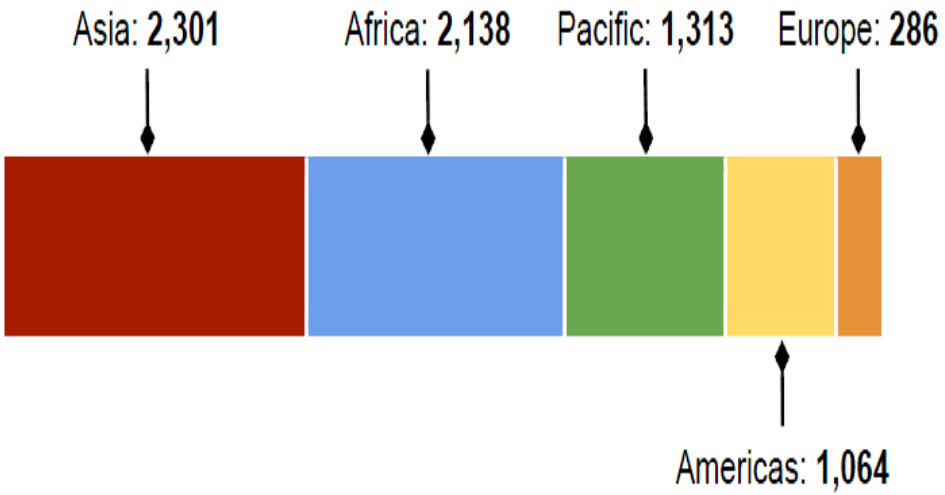
小语种场景下的NMT

Lan.	Train Size	Test Size	SMT BLEU	NMT BLEU
Hausa	1.0M	11.3K	23.7	16.8
Turkish	1.4M	11.6K	20.4	11.4
Uzbek	1.8M	11.5K	17.9	10.7
Urdu	0.2M	11.4K	17.9	5.2

(Zoph et al., 2016)

应用价值 ---“一带一路”建设

- 学术界与企业界都关心的热点。
- 建设“一带一路”工作中跨语言的顺畅沟通成为越发迫切的需求。
- “一带一路”沿线国家，特别是东盟、西亚、中亚国家，这些国家的语言几乎都是属于低资源语言。



图片来源：Washington Post Article – mentioning “Ethtnologue language of the world”,8th ed

图片来源：http://www.mrcjcn.com/n/224527.html

内容提要

- ✓研究背景及意义
- ●相关工作及研究现状
- ●面临的问题与挑战
- ●研究工作
  - ●基于自监督方法的预训练词切分模型
  - ●基于迁移学习的低资源语言神经机器翻译
  - ●基于数据增强的低资源语言神经机器翻译
- ●总结与展望

相关工作及研究现状

- 无监督预处理
- 利用高资源语言
- 高质量数据生成
- 半监督/无监督学习
- 零样本/少样本学习
- 对偶学习
- 元学习
- 预训练

研究现状 --- 词切分

- 词是一个最基本的语义单位
- 标注一致性：

操作系统(operating system)

操作(operating) /系统(system)

- 词的边界：

①犯罪(crime)/**案**(case)

②走私**案**(smuggling case)

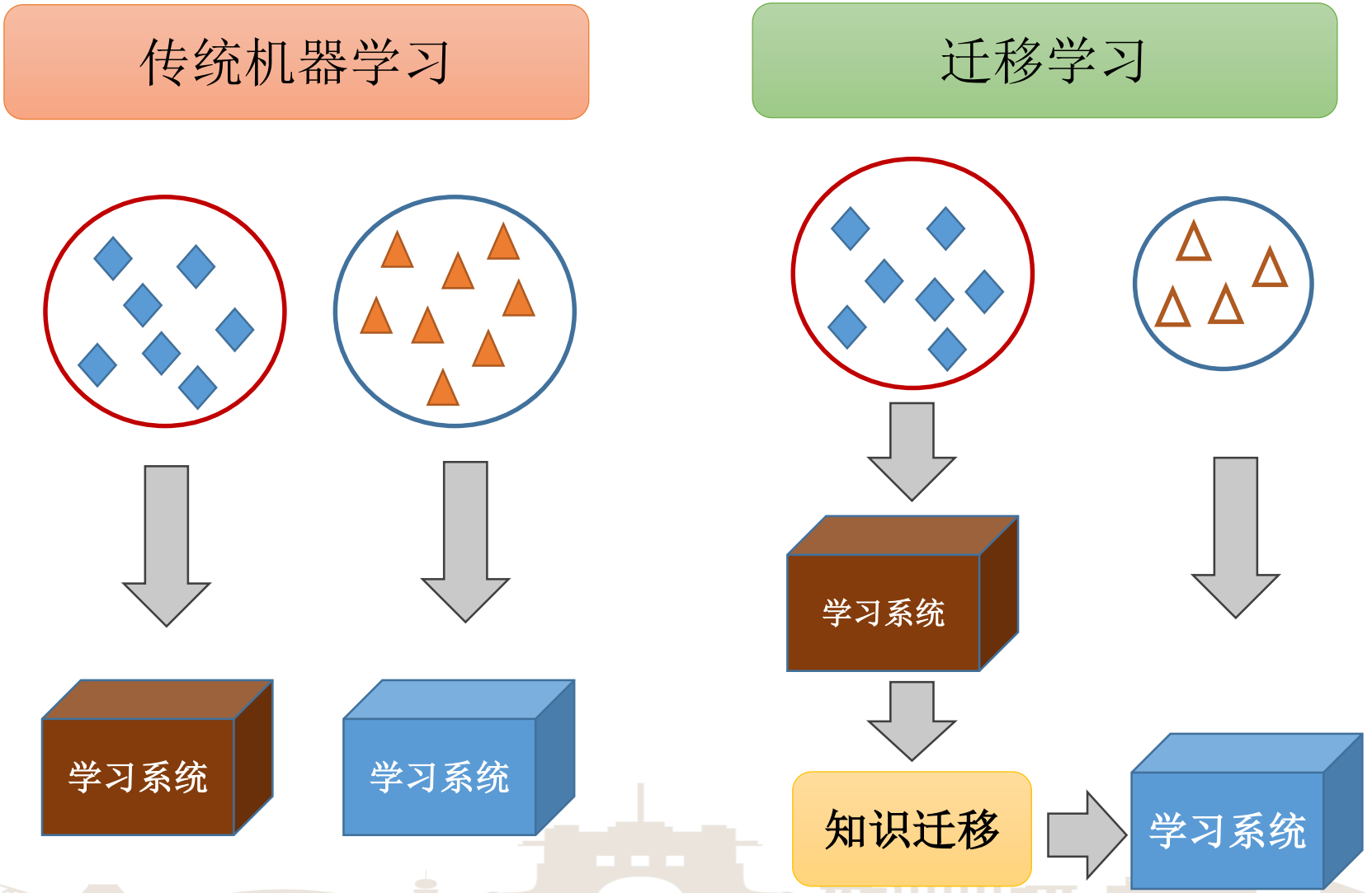
③话语**权**(discourse power)

Method	Zhang	Xiao	Fan	attend	a tournament	
PKU	张	小凡		参加	比武	大会
MSRA	张小凡			参加	比武大会	
Zhuxian	张小凡			参加	比武	大会

相关工作 --- 词切分

- 基于监督学习
  - 统计的方法 (Zhao et al., 2010)
  - 神经网络方法 (Chen et al., 2015; Cai et al., 2017)
- 基于外部资源
  - 统计特征 (Wang et al., 2019)
  - 领域词典 (Zhang et al., 2018)
  - 预训练 (Zhou et al., 2017)
- 基于预训练语言模型
  - BERT, RoBERTa (Tian et al., 2020; Huang et al., 2020)

研究现状 --- 迁移学习



相关工作 --- 高资源语言

- 基于多语言编码器的注意力机制

  (Marton et al., 2009; Nakov, et al., 2012; Dong et al., 2015; Zoph and Knight et al., 2016; Thanh-Le et al., 2016 ; Schwenk et al., 2017 ; Johnson et al., 2016 ; Get al., 2018; Dabre et al., 2019; Wang et al.,2020)

- 基于迁移学习的方法

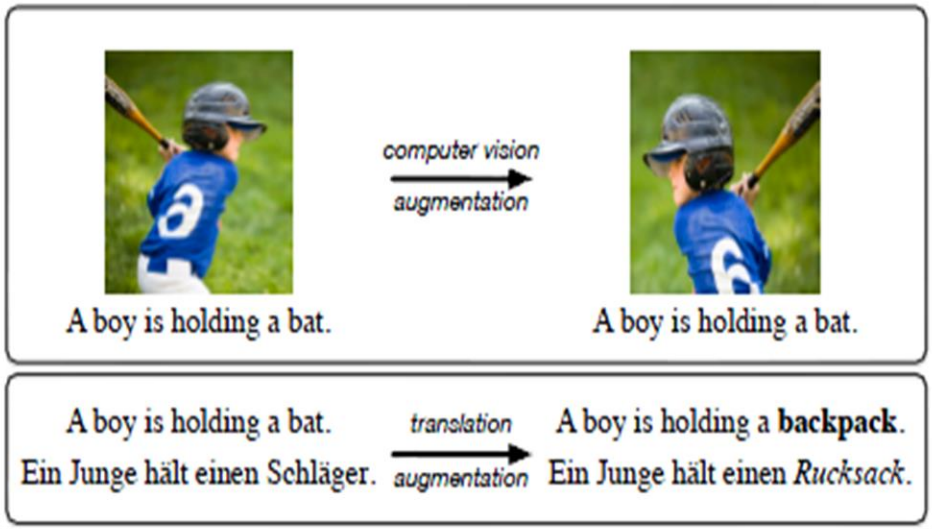
  (Wang, et al., 2012; Zoph et al., 2016 ; Zoph et al., 2017b ; Nguyen et al., 2017 ; Chu et al., 2017 ; Passban et al., 2017 ; Dabre et al., 2017; Wang al., 2018 ; Kim et al., 2019; Baijun et al., 2020; Artex et al., 2020)

- 基于元学习的方法

  (Gu et al., 2018; Li et al.,2020)

研究现状 --- 数据增强

- 数据增强是可以用来使训练数据规模扩大的方法。
- CV领域的增强：通过诸如随机旋转，调整大小，镜像和裁剪等转换来增强数据。(Alex et al., 2012, Ekin et al., 2018), GAN (Antreas Antoniou et al., 2017), ……



相关工作 --- 数据增强

- 基于单语数据的增强方法

  (Koehn et al., 2002; Quirk et al., 2004; Ueffing, et al., 2006; Marta, et al., 2006; Wubben et al., 2012; Gulchere et al., 2015; Sennrich et al., 2016b ; Cheng et al., 2016b ; Zhang et al., 2018; Zhou et al., 2019; Edunov et al., 2020; Marie et al., 2020)

- 基于词级别替换的增强方法

  (Francis et al., 2009; Fadaee et al., 2017 ; Huang et al., 2016; Sennrich et al., 2016c; Ribeiro et al., 2018; Wang et al., 2018; Xia et al., 2019; Ngeyuen et al., 2020; Baziotis et al.,2020)

内容提要

- ✓研究背景及意义
- ✓相关工作及研究现状
- ●面临的问题与挑战
- ●研究工作
  - ●基于自监督方法的预训练词切分模型
  - ●基于迁移学习的低资源语言神经机器翻译
  - ●基于数据增强的低资源语言神经机器翻译
- ●总结与展望

低资源语言NMT面临的挑战

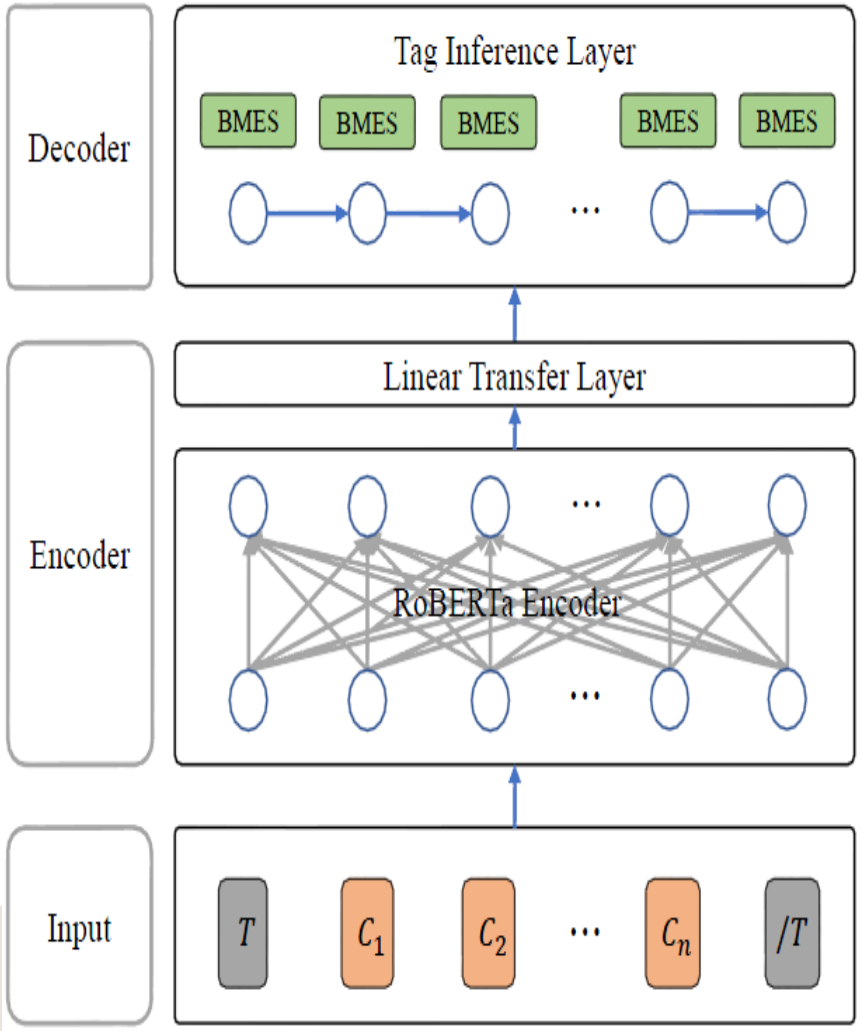
- 如何实现**高效**、**准确**的词切分问题
- 如何高效利用**高资源语言**问题
- 如何实现高质量的**数据增强**问题

挑战1：词切分模型不稳定

- Huang et al. (2020) 采用了预训练模型（RoBERTa）的方法，同时考虑了CWS中的OOV问题
- Tian et al. (2020) 同样利用了比原始BERT性能好的模型去探索了CWS的方法。

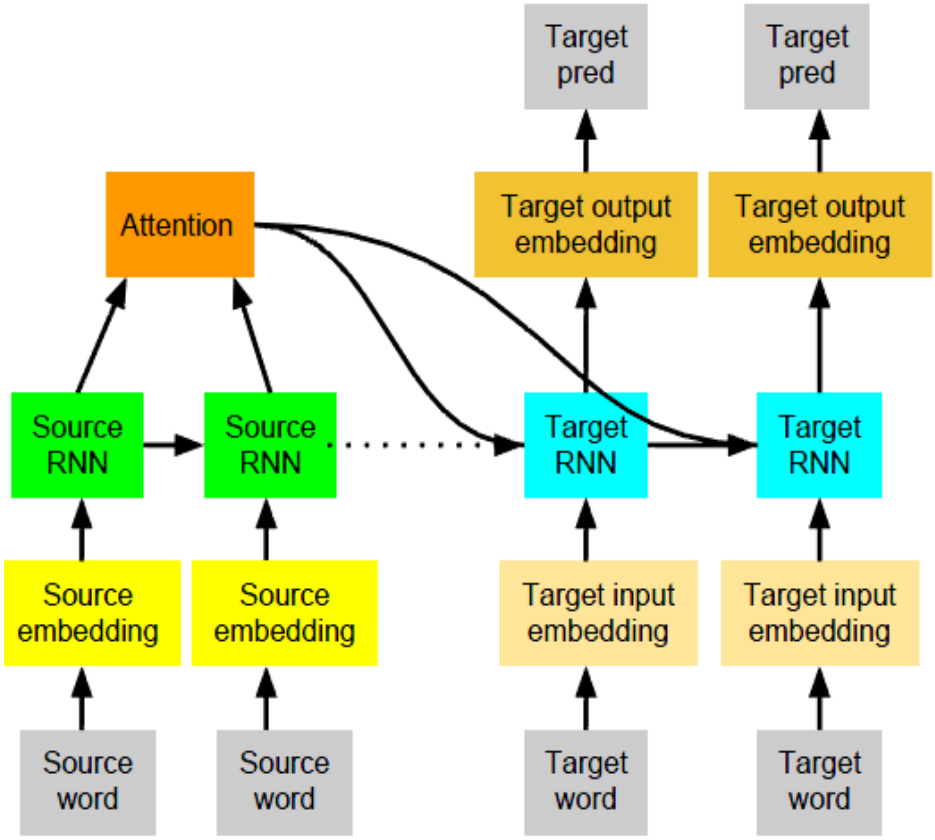
- 存在的问题：
  - 网络架构复杂
  - 计算复杂度高（RoBERTa）
  - 鲁棒性差，很难得稳定提升

（特别是在不同数据集和不同领域上）



挑战2：无法充分利用高资源语言

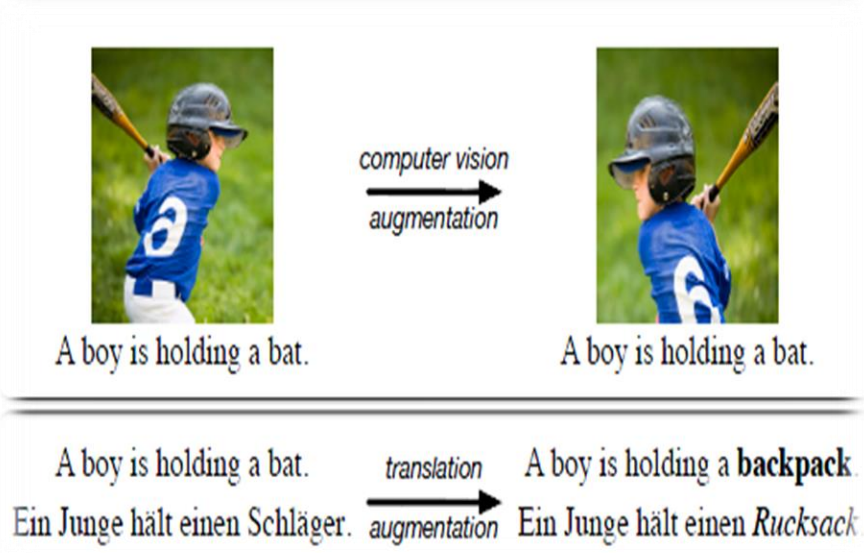
- Zoph et al. (2016) 使用迁移学习的方法，利用高资源语言来指导低资源语言
- Passban et al. (2017) 在以上工作的基础上，利用不同领域不同规模的同一种高资源语言



- 存在的问题：
  - 两种方法都难以高效利用**多个高资源**语言
  - 忽略字符级别的相似度

挑战3：数据增强质量不佳

- Fadaee et al. (2017) 用基于词级别替换的数据增强方法来扩充语料规模
- Cheng et al. (2016b) 用单语语料回译 (BT) 来扩充语料规模



Semantic: **John waters the** [**Plant**/**Bike**]

Syntax: **I have three** [**bags**/**pencil**]

- 存在的问题：
  - 依赖已有的翻译
  - 难以有效解决数据增强以后产生的**语义**和**句法**错误

original pair
$S: s_1, ..., s_i, ..., s_n$
$T: t_1, ..., t_j, ..., t_m$

augmented pair
$S': s_1, ..., s'_i, ..., s_n$
$T': t_1, ..., t'_j, ..., t_m$

内容提要

- ✓研究背景及意义
- ✓相关工作及研究现状
- ✓面临的问题与挑战
- ●研究工作
  - ●基于自监督方法的预训练词切分模型
  - ●基于迁移学习的低资源语言神经机器翻译
  - ●基于数据增强的低资源语言神经机器翻译
- ●总结与展望

研究工作

研究课题	低资源神经机器翻译				
主要方向	预处理	小语种翻译			
核心思想	鲁棒性 （预训练策略）	不产生数据 （迁移学习）		产生数据 （数据增强）	
研究挑战	如何实现高效、准确的词切分	如何充分并高效利用高资源语言		如何实现高质量（语法错误少）数据增强	
具体工作	自监督训练	多轮迁移	混合迁移	约束采样	复述表和词性

内容提要

- ✓研究背景及意义
- ✓相关工作及研究现状
- ✓面临的问题与挑战
- ●研究工作
  - ●基于自监督方法的预训练词切分模型
  - ●基于迁移学习的低资源语言神经机器翻译
  - ●基于数据增强的低资源语言神经机器翻译
- ●总结与展望

基于自监督的预训练词切分方法

- 动机：
  - 词切分无论是在机器翻译还是在对话、问答、信息检索、语音信息处理等不同任务上都起了**非常重要**的作用。
  - **分词模型**的性能对这些任务上不同模型最后的泛化能力的**影响**非常大。
  - 前人的工作中，模型**架构设计复杂**、依赖性强。
  - 计算复杂度高，一般实验室环境下**很难训练大规模**的模型。
  - 不同数据集上的提升不稳定，**鲁棒性**不佳。

词切分背景

- 输入序列（按字符级别）：

$$X = \{x_1, \cdots, x_n\}$$

$$Y^* = \{y_1^*, \cdots, y_n^*\}$$

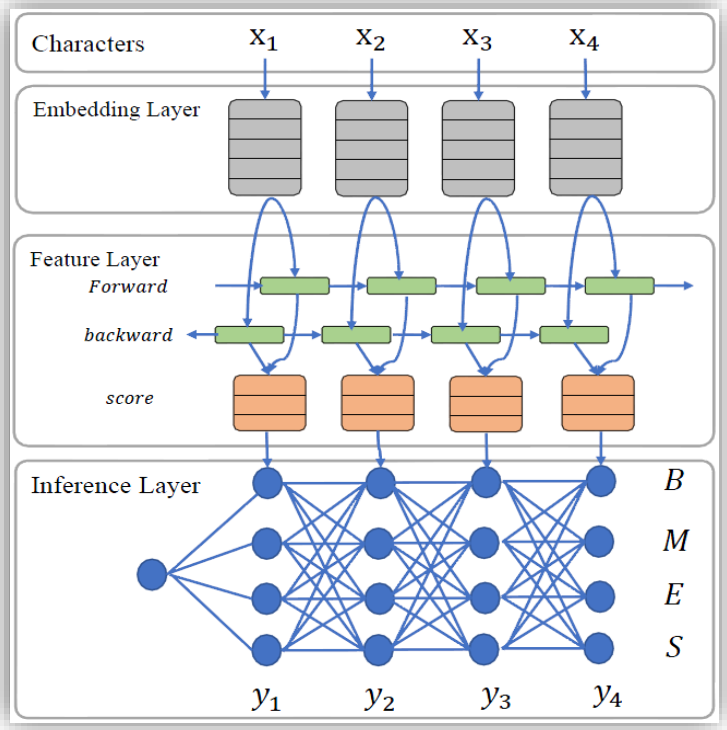
$$Y^* = \arg\max_{Y \in \mathcal{L}^n} p(Y|X)$$

$$\mathcal{L} = \{B, M, E, S\}$$

- 向量表示：
  - x_i可以映射到$\mathbf{e}_{x_i} \in \mathbb{R}^{d_e}$

- 特征获取
  - 典型的神经网络 (LSTM)

$$\mathbf{h}_i = \overrightarrow{\mathbf{h}}_i \oplus \overleftarrow{\mathbf{h}}_i$$
$$= \text{Bi-LSTM}(\mathbf{e}_{x_i}, \overrightarrow{\mathbf{h}}_{i-1}, \overleftarrow{\mathbf{h}}_{i+1}, \theta)$$



通用中文分词模型框架

- 输出层：
  - 用CRF预测4个标签

$$p(Y|X) = \frac{\Psi(Y|X)}{\sum_{Y' \in \mathcal{L}^n} \Psi(Y'|X)}$$

(Chen et al., 2017)

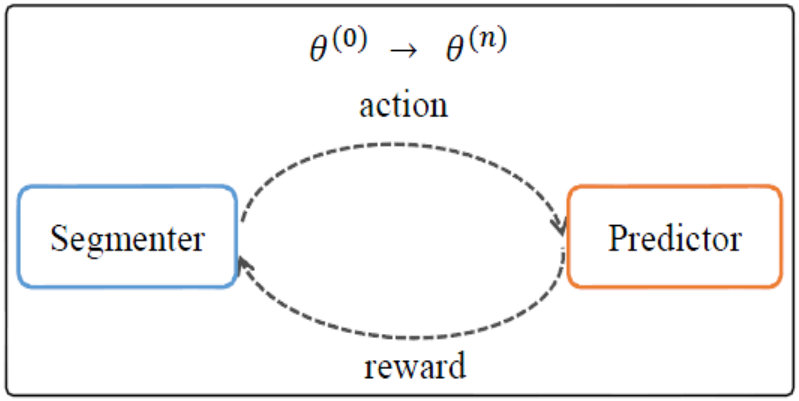
方法

- 输入序列

$$q(\mathbf{y}|\mathbf{x}) = \mathbb{E}_{\mathbf{x}_m|\mathbf{x}_o^{(s)},\mathbf{y};\gamma}\left[\Delta\left(\mathbf{x}_m, \mathbf{x}_m^{(s)}\right)\right]$$
$$= \sum_{\mathbf{x}_m \in M(\mathbf{x},\mathbf{y})} P\left(\mathbf{x}_m|\mathbf{x}_o^{(s)};\gamma\right)\Delta\left(\mathbf{x}_m, \mathbf{x}_m^{(s)}\right)$$

- 输入序列$\mathbf{x}$，标签序列$\mathbf{y}$
- $M(\mathbf{x},\mathbf{y})$是当分词结果为$\mathbf{y}$时$\mathbf{x}$的所有合法的掩码序列
- MLM的预测结果$\mathbf{x}_m$，掩码部分的标准答案$\mathbf{x}_m^{(s)}$，未被掩码的部分$\mathbf{x}_o^{(s)}$

$$\Delta\left(\mathbf{x}_m, \mathbf{x}_m^{(s)}\right) = 1 - sim\left(\mathbf{x}_m, \mathbf{x}_m^{(s)}\right)$$



自监督分词模型架构

Segged Seq.	小明 喜欢 吃 巧克力 。
Masked Input	[M] [M] 喜 欢 吃 巧 克 力 。 小 明 [M] [M] 吃 巧 克 力 。 小 明 喜 欢 [M] 巧 克 力 。 小 明 喜 欢 吃 [M] [M] 力 。 小 明 喜 欢 吃 巧 [M] [M] 。 小 明 喜 欢 吃 巧 克 力 [M]

Mask count = 2时的所有合法掩码序列

方法

- 训练过程类似于MRT (Shen et al., 2016)

$$J(\theta) = \sum_{\mathbf{x} \in \mathbf{X}} \mathbb{E}_{\mathbf{y}|\mathbf{x};\theta}[q(\mathbf{y}|\mathbf{x})] = \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in Y(\mathbf{x})} P(\mathbf{y}|\mathbf{x};\theta) q(\mathbf{y}|\mathbf{x})$$

  - $Y(\mathbf{x})$是所有可能的分词结果构成的集合；直接计算cost太大，因此从$Y(\mathbf{x})$中采样出一个子集$S(\mathbf{x})$

- 在$S(\mathbf{x})$上定义一个新的概率分布

$$Q(\mathbf{y}|\mathbf{x};\theta,\alpha) = \frac{P(\mathbf{y}|\mathbf{x};\theta)^{\alpha}}{\sum_{\mathbf{y}' \in S(\mathbf{x})} P(\mathbf{y}'|\mathbf{x};\theta)^{\alpha}}$$

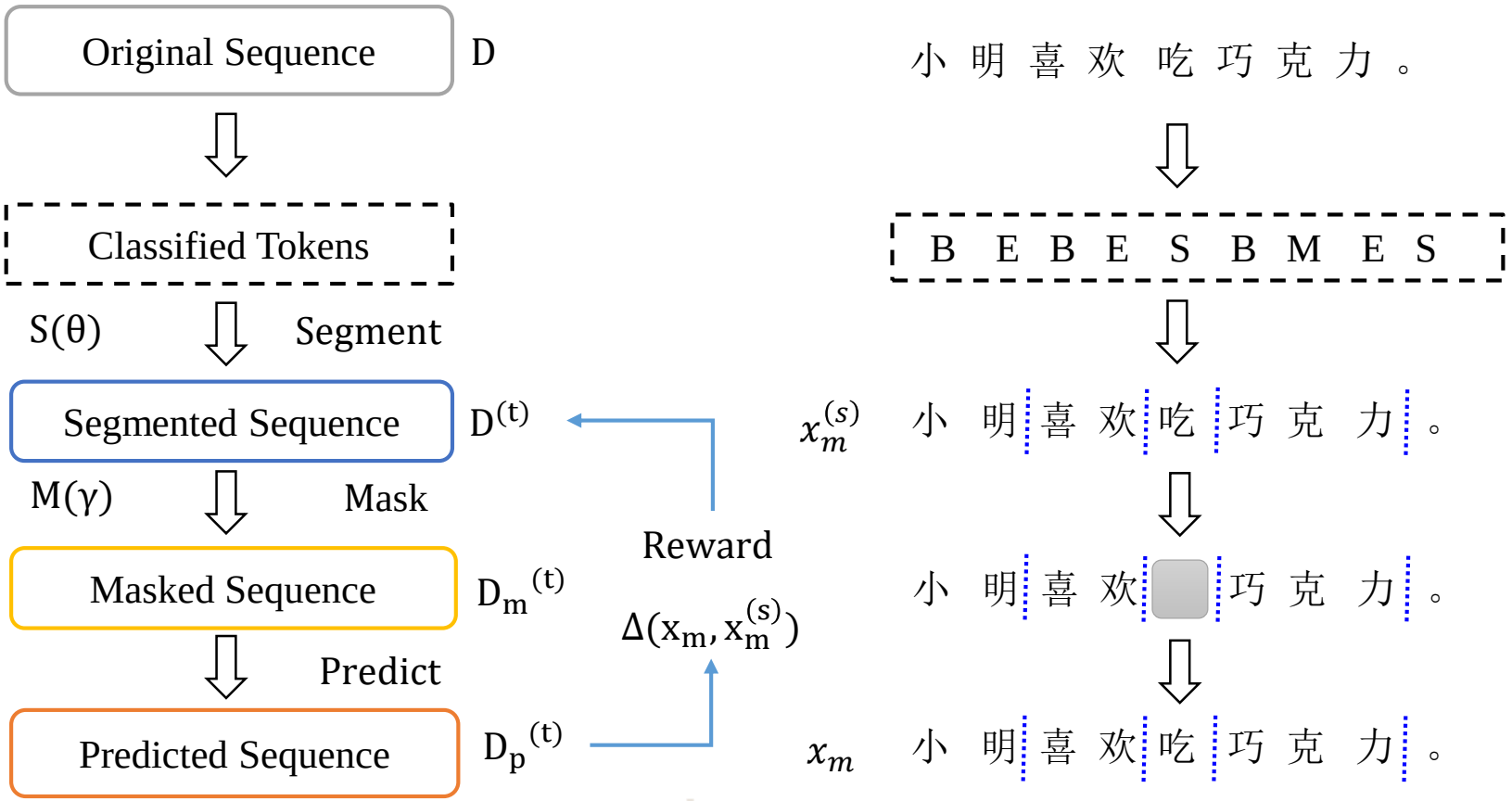
- 在Q上计算优化目标的近似值

$$J(\theta) = \sum_{\mathbf{x} \in \mathbf{X}} \mathbb{E}_{\mathbf{y}|\mathbf{x};\theta,\alpha}[q(\mathbf{y}|\mathbf{x})] = \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{y} \in S(\mathbf{x})} Q(\mathbf{y}|\mathbf{x};\theta,\alpha) q(\mathbf{y}|\mathbf{x})$$

- 加正则项来降低Q的分母减小带来的负面影响

$$J(\theta) = \sum_{\mathbf{x} \in \mathbf{X}} \left(\sum_{\mathbf{y} \in S(\mathbf{x})} Q(\mathbf{y}|\mathbf{x};\theta,\alpha) q(\mathbf{y}|\mathbf{x}) - \lambda \sum_{\mathbf{y}' \in S(\mathbf{x})} P(\mathbf{y}'|\mathbf{x};\theta)^{\alpha} \right)$$

模型架构


Original Sequence
D
Classified Tokens
S(θ)
Segment
Segmented Sequence
D(t)
M(γ)
Mask
Masked Sequence
Dm(t)
Predict
Predicted Sequence
Dp(t)
Reward
Δ(xm, xm(s))
小 明 喜 欢 吃 巧 克 力 。
B E B E S B M E S
xm(s)
小 明 喜 欢 吃 巧 克 力 。
小 明 喜 欢 巧 克 力 。
xm
小 明 喜 欢 吃 巧 克 力 。

实验设置

- 数据集
  - SIGHAN05: MSRA,PKU,AS,CITYU
  - SIGHAN08: CTB, SXU
  - SIGHAN10: Finance, Literature, Medicine
  - OTHER 公开数据集: CNC, UDC, ZX

模型参数

Parameter	BERT
Hidden Layer	768
Number of Layers	12
Number of Heads	12
Learning Rate	2e-5
Batch Size	64
Dropout	0.1
Epochs	10

数据属性

Corpora	Train	Dev.	Test	Word			Char		
				Type	Token.	Avglen.	Type	Token.	Avglen.
MSRA	84.80K	2.0K	4.0K	90.10K	2.50M	27.24	5.20K	4.01M	46.62
PKU	19.06K	2.0K	1.9K	58.20K	1.21M	57.82	4.70K	1.83M	95.85
AS	0.7M	2.0K	14.4K	0.14M	5.60M	7.7	6.11K	8.37M	11.80
CITYU	53.02K	2.0K	1.5K	70.76K	1.50M	27.45	4.92K	2.40M	45.33
CTB	24.42K	1.9K	2.0K	47.60K	0.80M	27.67	4.44K	1.30M	45.50
SXU	15.62K	1.5K	3.7K	35.92K	0.64M	30.90	4.28K	1.04M	50.50
CNC	0.21M	25.9K	25.9K	0.14M	7.30M	28.19	6.86K	10.08M	43.28
UDC	4.0K	0.5K	0.5K	20.13K	0.12M	24.67	3.60K	0.20M	39.14
ZX	2.37K	0.8K	1.4K	9.14K	0.12M	26.87	2.61K	0.17M	38.05

实验结果

单标准学习模式 F1-score (%)

Methods	SIGHAN05				SIGHAN08		OTHER		
	MSRA	PKU	AS	CITYU	CTB	SXU	CNC	UDC	ZX
Chen et al. (2017)	95.84	93.30	94.20	94.07	95.30	95.17	-	-	-
Zhou et al. (2017)	97.80	96.00	-	-	96.20	-	-	-	-
Yang et al. (2017)	97.50	96.30	95.70	96.90	96.20	-	-	-	-
He et al. (2018)	97.29	95.22	94.90	94.51	95.21	95.78	97.11	93.98	95.57
Gong et al. (2019)	96.46	95.74	94.51	93.71	97.09	95.57	-	-	-
LSTM+BEAM	97.10	95.80	95.30	95.60	96.10	95.95	96.10	96.20	96.30
LSTM+CRF	98.10	96.10	96.00	96.80	96.30	96.55	96.61	96.00	96.40
BERT	96.91	95.34	96.47	97.10	97.27	96.40	96.66	97.23	96.49
SELFATT+SOFT	97.60	95.50	95.70	96.40	97.28	96.60	96.88	97.12	96.50
BERT+LTL	97.53	96.23	97.03	97.63	97.34	96.65	96.89	97.51	96.72
Ours	**98.12**	**96.24**	**97.30**	**97.83**	**97.45**	**96.97**	**97.25**	**97.74**	**96.82**

实验结果

多标准学习模式 F1-score (%)

Methods	SIGHAN05				SIGHAN08		OTHER		
	MSRA	PKU	AS	CITYU	CTB	SXU	CNC	UDC	ZX
Chen et al. (2017)	96.04	94.32	94.64	95.55	96.18	96.04	-	-	-
He et al. (2018)	97.35	95.78	95.47	95.60	95.84	96.49	97.00	94.44	95.72
Gong et al. (2019)	97.78	96.15	95.22	96.22	97.26	97.25	-	-	-
BERT	97.22	96.06	97.07	97.39	97.36	96.81	96.71	97.48	96.60
BERT+LTL	96.67	96.30	97.16	97.72	97.38	96.90	97.10	97.61	96.81
Ours	**98.19**	**96.32**	**97.43**	**97.80**	**97.66**	**97.03**	**97.34**	**98.25**	**97.08**

带噪声数据上的单标准学习模式 F1-score (%)

Methods	SIGHAN05				SIGHAN08		OTHER		
	MSRA	PKU	AS	CITYU	CTB	SXU	CNC	UDC	ZX
LSTM+BEAM	96.86	95.70	95.17	95.35	95.89	95.83	95.89	96.07	96.18
LSTM+CRF	97.89	95.89	95.88	96.67	96.19	96.47	96.49	95.85	96.25
BERT	96.78	95.20	96.28	97.01	97.14	96.24	96.51	97.11	96.30
SELFATT+SOFT	97.47	95.40	95.57	96.29	97.16	96.49	96.61	97.08	96.33
BERT+LTL	97.42	96.15	96.76	97.52	97.27	96.55	96.69	97.40	96.53
Ours	**97.93**	**96.18**	**97.12**	**97.68**	**97.32**	**96.83**	**97.12**	**97.63**	**96.67**

实验结果

不同领域上的性能比较F1-score (%)

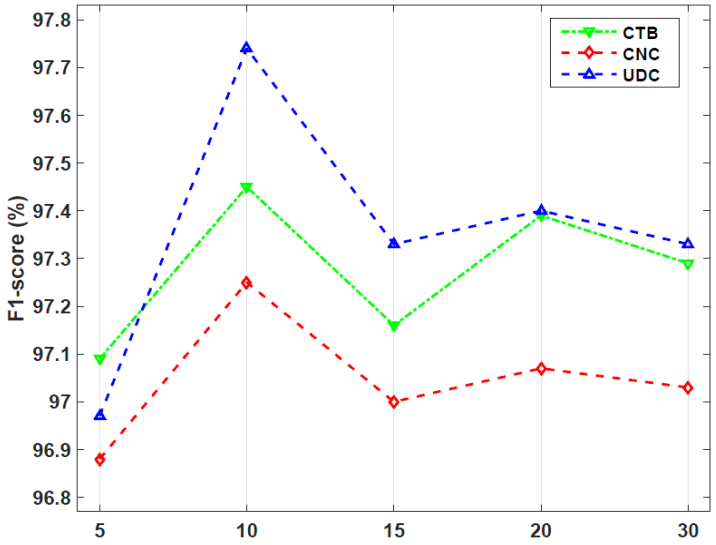
Methods	Fin.	Lit.	Med.
Chen et al. (2017)	95.20	92.89	92.16
Cai et al. (2017)	95.38	92.90	92.10
Huang et al. (2017)	95.81	94.33	92.26
Zhao et al. (2018)	95.84	93.23	93.73
Zhang et al. (2018)	96.06	94.76	94.18
BERT	95.87	95.57	94.66
BERT+LTL	95.96	95.88	94.87
Ours	95.93	**95.96**	**95.08**

预训练模型的影响

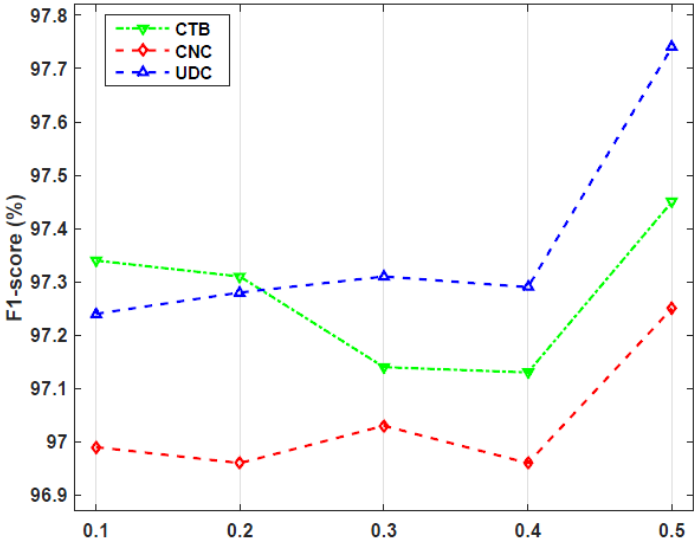
Corpora	PTM	P.	R.	F.
MSRA	×	97.06	97.61	97.34
	√	98.18	98.06	**98.12**
AS	×	96.05	96.78	96.41
	√	96.30	98.33	**97.30**
CTB	×	95.97	96.23	96.10
	√	97.49	97.41	**97.45**
CNC	×	96.08	95.42	95.75
	√	97.41	97.08	**97.25**

实验结果

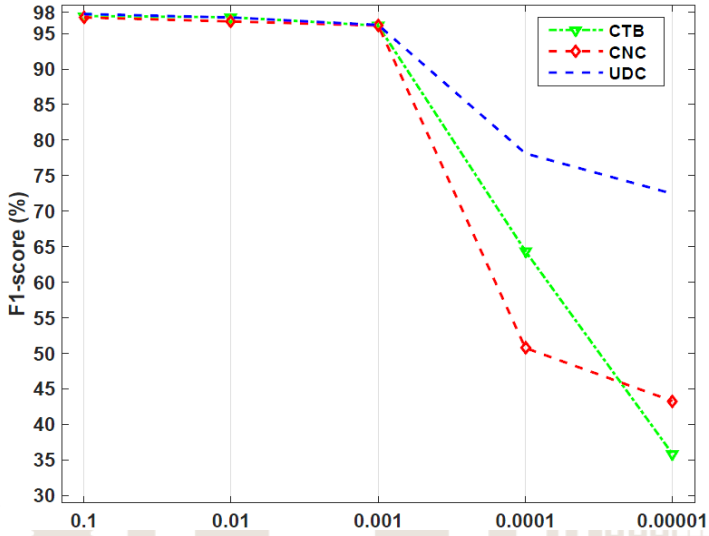
$S(\mathbf{x})$的大小的影响



参数α的影响



参数λ的影响



实验结果

低资源数据集特性介绍

Languages	Train	Dev	Test	Source		Target	
				Voc.	Word	Voc.	Word
Zh – Az	20.1K	0.5K	0.5K	11.9K	0.6M	25.1K	0.6M
Zh – Tr	101.6K	1.0K	1.0K	12.8K	2.9M	29.2K	2.7M
Zh – Ur	78.0K	1.0K	1.0K	12.7K	2.4M	17.6K	2.6M
Zh – Ug	46.3K	1.0K	1.0K	42.1K	11.2M	73.5K	1.1M

自监督词切分模型对低资源神经机器翻译的影响

处理方法	Zh – Az	Zh – Tr	Zh – Ur	Zh – Ug
字符处理	32.06	15.32	23.90	22.40
随机切分	27.98	12.27	19.99	18.46
子词切分	30.16	14.77	22.87	20.12
我们方法	**33.07**	**15.89**	**24.59**	**23.68**

小结

- 我们提出了**一种自监督的中文分词**方法，使用MLM的预测结果辅助分词模型的训练；
- 我们提出了**MRT的一种改进版本**，即带正则项的MRT，用于提高自监督分词模型的性能；
- 实验结果表明，我们的方法优于以前的方法，并且对带有噪声的数据具有更好的**鲁棒性**。
- 同时，**本方法**对中文与**资源匮乏的语言**之间的神经机器翻译**有帮助**。

内容提要

- ✓研究背景及意义
- ✓相关工作及研究现状
- ✓面临的问题与挑战
- ●研究工作
  - ✓基于自监督方法的预训练词切分模型
  - ●基于迁移学习的低资源语言神经机器翻译
  - ●基于数据增强的低资源语言神经机器翻译
- ●总结与展望

基于迁移学习的低资源语言NMT

- 动机：
  - 前人方法**无法充分**利用多个高度接近的高资源语言。
  - 从父模型迁移到子模型时**忽略字符**级别的相似度。
  - 预处理方面**忽略了**父语言和子语言之间原始文本的**统一化**表示。
  - 初始化子模型之前训练父模型时**无法得到**与子模型**有关的信息。**

迁移学习背景

- 我们将 $L_3 \rightarrow L_2$ 和 $L_1 \rightarrow L_2$ 作为父模型和子模型的语言对。 L_3 和 L_1 分别表示父模型和子模型的源语言，L_2是它们的目标语言。

- $\theta_{L_3 \rightarrow L_2} = \{< e_{L_3}, W, e_{L_2} >\}$，其中 e_{L_3} 和 e_{L_2} 是父模型的源语言和目标语言的词向量，W 是模型参数。

- $\hat{\theta}_{L_3 \rightarrow L_2} = \underset{\theta_{L_3 \rightarrow L_2}}{\operatorname{argmax}} \{L(D_{L_3}, \theta_{L_3 \rightarrow L_2})\}$ 训练父模型 $M_{L_3 \rightarrow L_2}$

- 然后利用父模型 $M_{L_3 \rightarrow L_2}$ 去初始化 $M_{L_1 \rightarrow L_2}$：

- $\theta_{L_1 \rightarrow L_2} = f(\hat{\theta}_{L_3 \rightarrow L_2})$，其中 f 是初始化函数。

方法

●最初的迁移学习在低资源NMT：

$$\theta_{L_3} = \{\langle e_{L_3}, W, e_{L_3} \rangle\}$$

$$\hat{\theta}_{L_3 \to L_2} = \underset{\theta_{L_3 \to L_2}}{\operatorname{argmax}} \{L(D_{L_3 \to L_2}, \theta_{L_3 \to L_2})\}$$

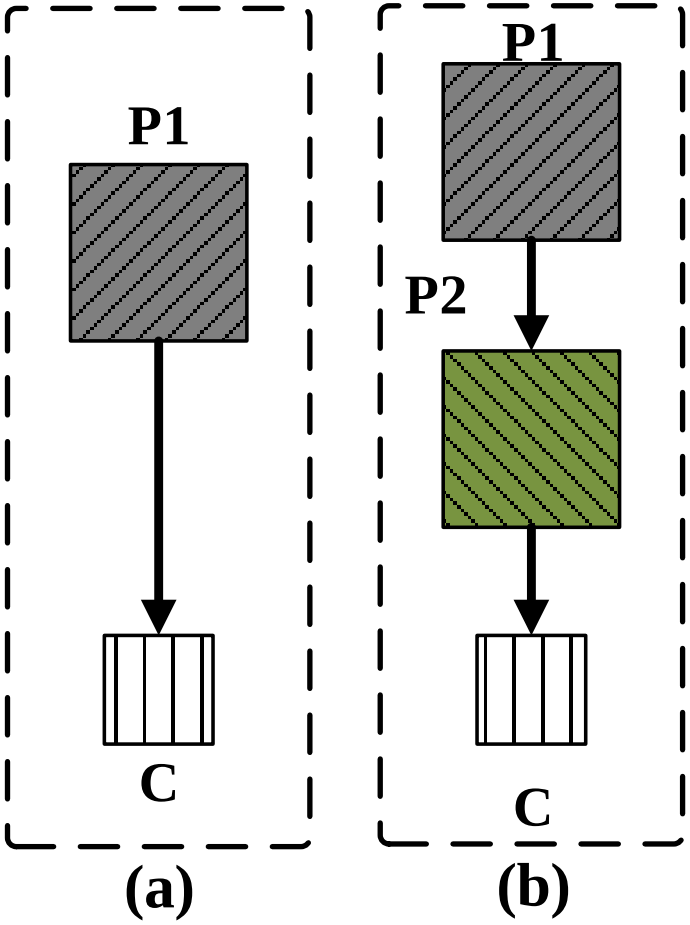
$$\theta_{L_1 \to L_2} = f(\hat{\theta}_{L_3 \to L_2})$$

●改进的迁移学习在低资源NMT：

$$\theta_{L_3' \to L_2} = f(\hat{\theta}_{L_3 \to L_2})$$

$$\hat{\theta}_{L_3' \to L_2} = \underset{L_3 \to L_2}{\operatorname{argmax}} \{L(D_{L_3' \to L_2}, \theta_{L_3' \to L_2})\}$$

$$\theta_{L_1 \to L_2} = f(\hat{\theta}_{L_3' \to L_2})$$



(Zoph et al., 2017)　(Passban et al., 2017)

方法1

$$\theta_{L_4 \to L_2} = f(\hat{\theta}_{L_3 \to L_2})$$

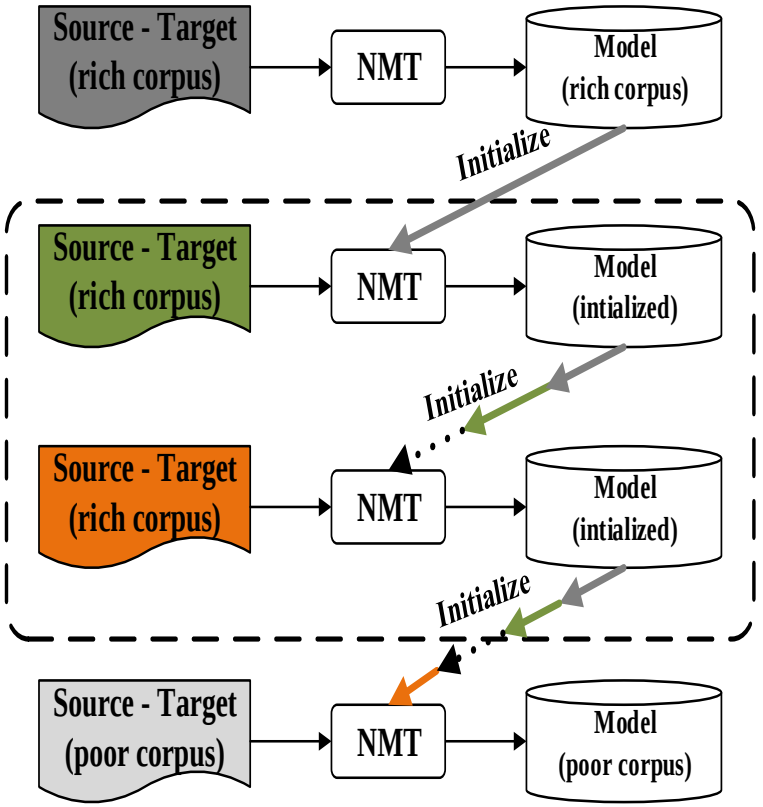
$$\hat{\theta}_{L_4 \to L_2} = \underset{\theta_{L_4 \to L_2}}{\operatorname{argmax}} \{L(D_{L_4 \to L_2}, \theta_{L_4 \to L_2})\}$$

$$\vdots \quad = \quad \vdots$$

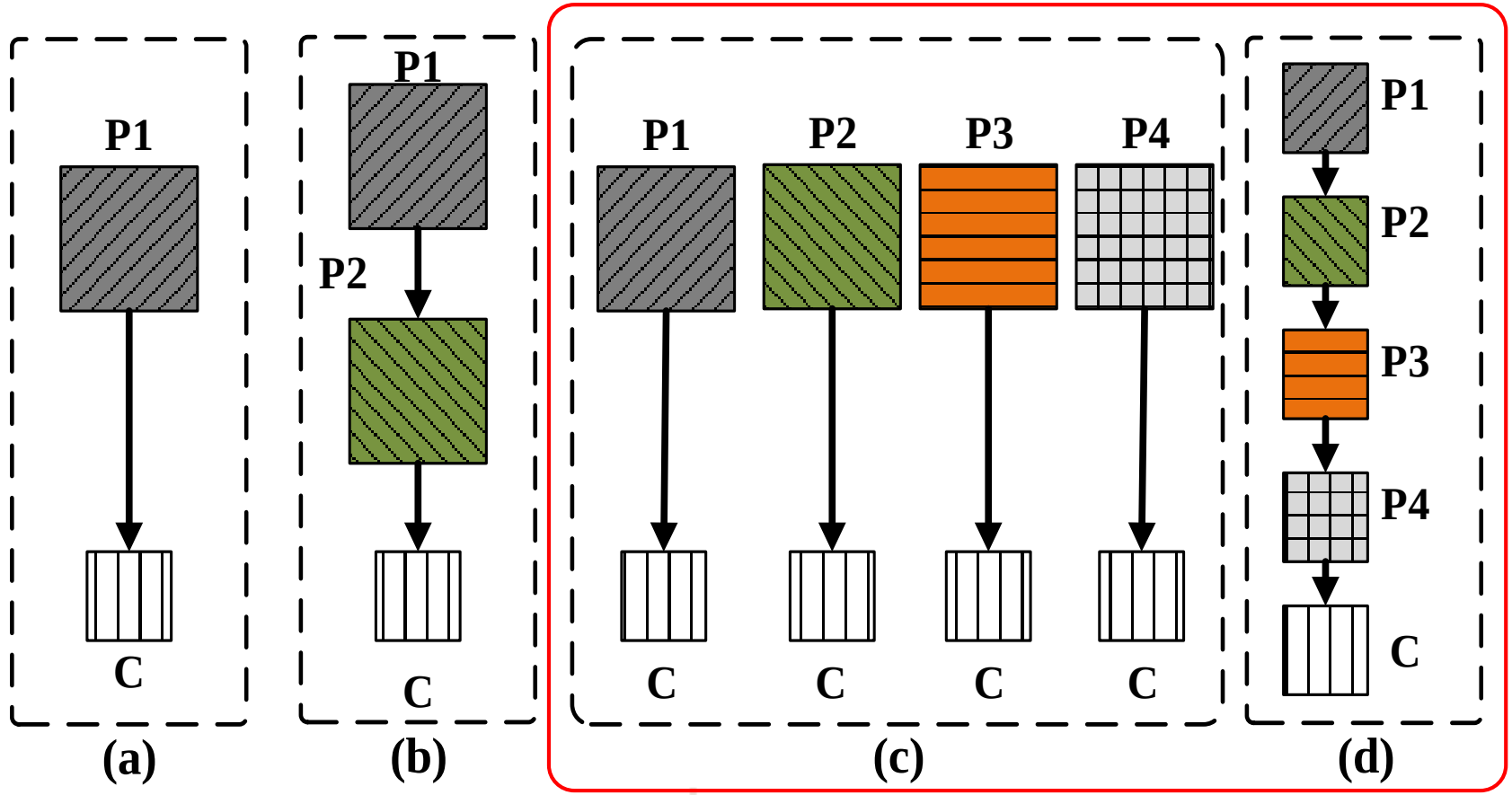
$$\theta_{L_{k+1} \to L_2} = f(\hat{\theta}_{L_k \to L_2})$$

$$\hat{\theta}_{L_{k+1} \to L_2} = \underset{\theta_{L_{k+1} \to L_2}}{\operatorname{argmax}} \{L(D_{L_{k+1} \to L_2}, \theta_{L_{k+1} \to L_2})\}$$

$$\theta_{L_1 \to L_2} = f(\hat{\theta}_{L_{k+1} \to L_2})$$



模型架构1



(Zoph et al., 2017) (Passban et al., 2017) **Max** **Multi**

实验设置

- 同构的数据集
  - Chinese LDC 数据集
  - OpenSubtitles2016
  - Tanzil corpora

数据特性介绍

Languages	Train	Dev	Test	Source		Target	
				Voc.	Word	Voc.	Word
Ar – Zh	5.1M	2.0K	2.0K	1.0M	32.2M	0.5M	37.4M
Fa – Zh	1.4M	2.0K	1.0K	0.2M	10.4M	0.2M	10.0M
Ur – Zh	78.0K	1.0K		17.6K	2.6M	12.7K	2.4M
Fi – Zh	2.8M	2.0K		0.7M	18.4M	0.3M	23.1M
Hu – Zh	4.1M	2.0K		1.0M	30.4M	0.5M	32.5M
Tr – Zh	4.4M	2.0K		0.7M	30.6M	0.5M	35.9M
Ug – Zh	**46.3K**	1.0K		73.5K	1.1M	42.1K	11.2M

语言属性分布

Languages		Family	Group	Branch	Order	Unit	Inflection
Arabic	Ar	Hamito-Semitic	Semitic	South	VSO	Word	High
Farsi	Fa	Indo-European	Indic	West	SOV		Moderate
Urdu	Ur		Iranian	Iranian			
Finnish	Fi	Uralic	Fino -Ugric	Finnish	SVO		
Hungarian	Hu			Ugric			
Turkish	Tr	Altatic	Turkic	Oghuz	SOV		
Uyghur	Ug			Turkic			
Chinese	Zh	Sino-Tibetan	Chinese	Chinese	SVO	Character	Light

实验结果

共享单词比例

	Ar	Fa	Ur	Fi	Hu	Tr	Ug
Ar		11.49%	8.31%	0.52%	0.43%	0.73%	0.77%
Fa	2.34%		8.29%	0.27%	0.30%	0.32%	0.57%
Ur	0.15%	0.75%		0.01%	0.01%	0.03%	0.11%
Fi	0.36%	0.94%	0.53%		2.74%	3.80%	0.50%
Hu	0.45%	1.46%	0.70%	3.85%		5.07%	0.75%
Tr	0.57%	1.22%	1.14%	4.22%	4.01%		2.47%
Ug	0.06%	0.21%	0.47%	0.05%	0.06%	0.24%	

不同规模的父模型影响

Method	Parent	Child	BLEU
Transformer	N/A	Ug – Zh	28.28
MRTL (R=1)	Tr (0.5M) – Zh		29.89
	Tr (2.4M) – Zh		30.88
	Tr (4.4M) – Zh		32.74

实验结果

统一化转换例子

Language	Original	Latin	Chinese	English	Unified
Ar	مدرسة	maktab	学校	School	mektep
Ug	مەكتەپ	mektep			
Tr	okul	mektep			
Fa	باغ وحش	bağça	果园	Orchard	bağça
Ug	باغچا	bağça			
Tr	bahçesi	bahçe			

统一化转换的影响

Method	Round	Parent	Child	BLEU
Transformer	R=0	N/A	Ug – Zh	28.28
MRTL (Original)	R=1	Ur – Zh		10.29
		Fa – Zh		**28.83**
		Ar – Zh		30.64
MRTL (**Unified**)		Ur – Zh		10.93
		Fa – Zh		**29.96**
		Ar – Zh		31.64

每一个父模型（**Max**）的影响

Method	Parent	Child	BLEU
Transformer	N/A	Ug – Zh	28.28
MRTL (R=1)	Ur – Zh		10.93
	Fa – Zh		29.63
	Fi – Zh		30.85
	Tr (2.4M) – Zh		30.88
	Ar – Zh		31.64
	Hu – Zh		32.41
	Tr – Zh		32.74

实验结果

不同语系的影响

MRTL	Parent	Family	Domain	Size	Child	BLEU
R=0	N/A	Altaic	CLDC	46.3K	Ug – Zh	28.28
R=1	Hu – Zh	Uralic	Open Subtitle	4.1M		32.41
	Tr – Zh	Altaic				32.58

不同领域的影响

MRTL	Parent	Family	Domain	Size	Child	BLEU
R=0	N/A	Altaic	CLDC	46.3K	Ug – Zh	28.28
R=1	Ur – Zh	Indo-European	Tanzil	78.0K		10.93
	Fa – Zh		Open Subtitle			24.27

不同规模语料的影响

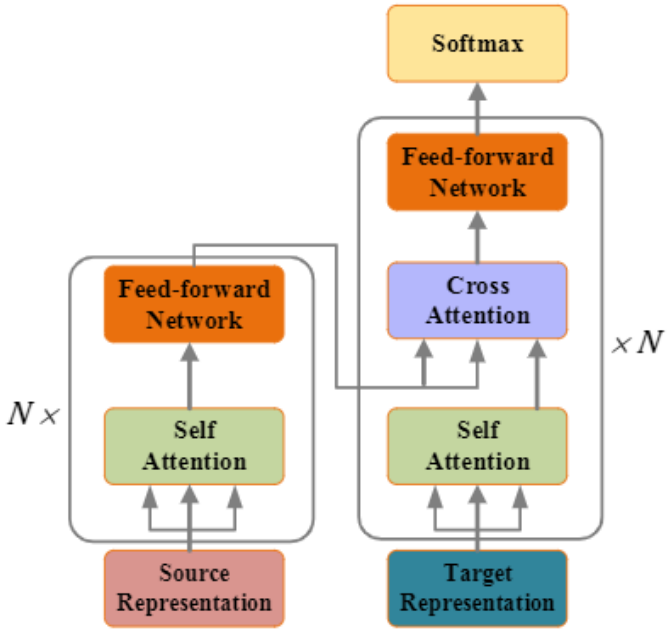
MRTL	Parent	Family	Domain	Size	Child	BLEU
R=0	N/A	Altaic	CLDC	46.3K	Ug – Zh	28.28
R=1	Fi – Zh	Uralic	Open Subtitle	2.8M		30.85
	Hu – Zh			4.1M		32.41

实验结果

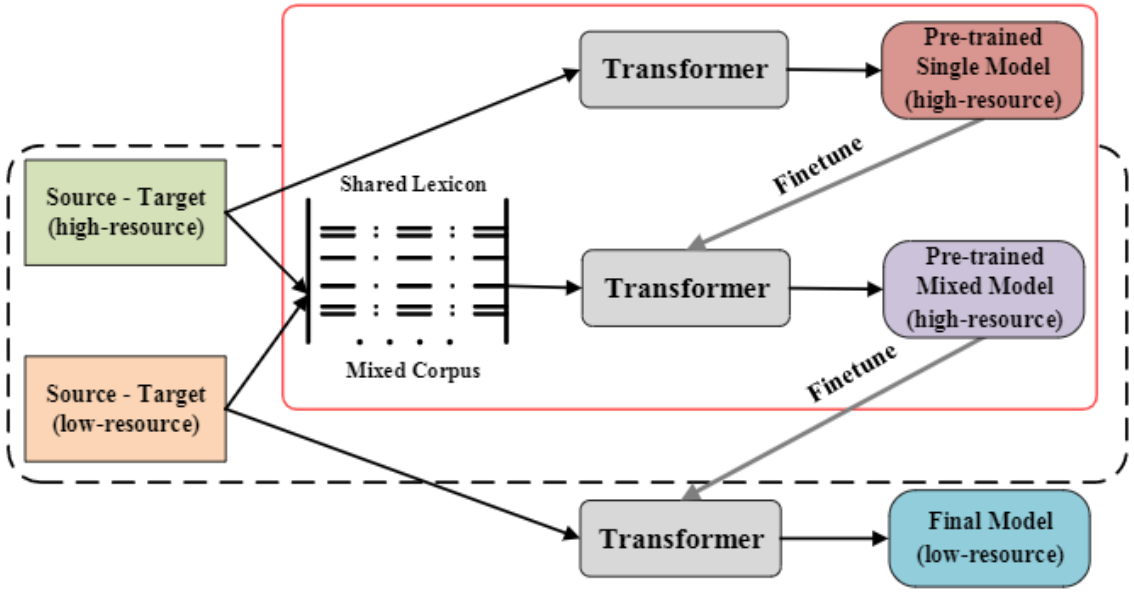
MRTL的影响

Method	Round	Parent	Child	BLEU
Transformer	R=0	N/A	Ug – Zh	28.28
Many-to-One				32.43
MRTL	R=1	Tr (4.4M) – Zh		32.03
	R=2	Tr (4.4M), (2.4M) – Zh		32.54
	R=3	Tr (4.4M), (2.4M), Fi – Zh		33.54
	R=4	Tr (4.4M), (2.4M), Fi, Hu – Zh		33.66
		Ar (Unified), Tr (4.4M), Hu, Fi – Zh		33.73
		Tr (4.4M), Ar (Unified), Hu, Fi – Zh		33.91

模型架构2



(a) Transformer



(b) Mixed Transfer Architecture

实验设置

- 数据集
  - Tanzil corpora

数据特性介绍

Languages	Train	Dev	Test	Source		Target	
				Voc.	Word	Voc.	Word
Arabic (Ar) – Zh	5.1M	2.0K	2.0K	1.0M	32.2M	0.5M	37.4M
Farsi (Fa) – Zh	1.4M	2.0K	2.0K	0.2M	10.4M	0.2M	10.0M
Azerbaijani (Az) – Zh	**20.1K**	0.5K	0.5K	25.1K	0.6M	0.2M	10.0M
Turkish (Tr) – Zh	4.4M	2.0K	1.0K	0.7M	30.6M	0.5M	35.9M
Uyghur (Ug) – Zh	10.9K	0.5K	0.5K	18.3K	0.4M	12.5K	0.4M
Uzbek (Uz) – Zh	**10.5K**	0.5K	0.5K	26.5K	0.5M	16.3K	0.3M

实验结果

共享词表的影响

Method	Parent	Child	BLEU
Transformer	N/A	Az – Zh	43.68
		Uz – Zh	40.99
MTL (non-shared)	Tr (2.4M) - Zh	Az – Zh	**45.97**
	Fa – Zh	Uz – Zh	42.15
	Tr (4.4M) - Zh	Az – Zh	46.81
	Ar – Zh	Uz – Zh	42.64
MTL (shared)	Tr (2.4M) - Zh	Az – Zh	**46.44**
	Fa – Zh	Uz – Zh	42.53
	Tr (4.4M) - Zh	Az – Zh	47.32
	Ar – Zh	Uz – Zh	42.89

共享和拉丁化的影响

Method	Parent	Child	BLEU
Transformer	N/A	Az – Zh	43.68
		Uz – Zh	40.99
Original (TL)	Tr (2.4M) - Zh	Az – Zh	45.82
	Fa – Zh	Uz – Zh	42.03
	Tr (4.4M) - Zh	Az – Zh	46.49
	Ar – Zh	Uz – Zh	42.51
MTL (shared)	Tr (2.4M) - Zh	Az – Zh	46.44
	Fa – Zh	Uz – Zh	42.53
	Tr (4.4M) - Zh	Az – Zh	47.32
	Ar – Zh	Uz – Zh	42.89
MTL (shared+latin)	Tr (2.4M) - Zh	Az – Zh	46.44
	Fa – Zh	Uz – Zh	42.82
	Tr (4.4M) - Zh	Az – Zh	47.32
	Ar – Zh	Uz – Zh	43.11

实验结果

混合模型+拉丁化+双轮的影响

Method	Parent	Child	BLEU
Vaswani et al. (2017)	N/A	Az – Zh	43.68
		Uz – Zh	40.99
Johnson et al. (2017)		Az – Zh	46.74
		Uz – Zh	43.67
Zoph et al. (2016)	Tr (2.4M) – Zh	Az – Zh	45.82
	Fa – Zh	Uz – Zh	42.03
	Tr (4.4M) – Zh	Az – Zh	46.49
	Ar – Zh	Uz – Zh	42.51
Passban et al. (2017)	Tr (4.4M), (2.4M) – Zh	Az – Zh	47.55
		Uz – Zh	44.21
MTL (shared)		Az – Zh	48.62
		Uz – Zh	45.41
MTL (shared+latin)		Az – Zh	48.62
		Uz – Zh	45.83

小结

- 我们提出了**两种基于迁移学习**的低资源语言NMT模型。
- 提出了一个父语言对和子语言对之间能够实现**统一化的转换**方法。
- 提出了更好的**选择高度接近的**父模型语言对的策略。
- 提出了父模型和子模型之间**词表共享**的策略。

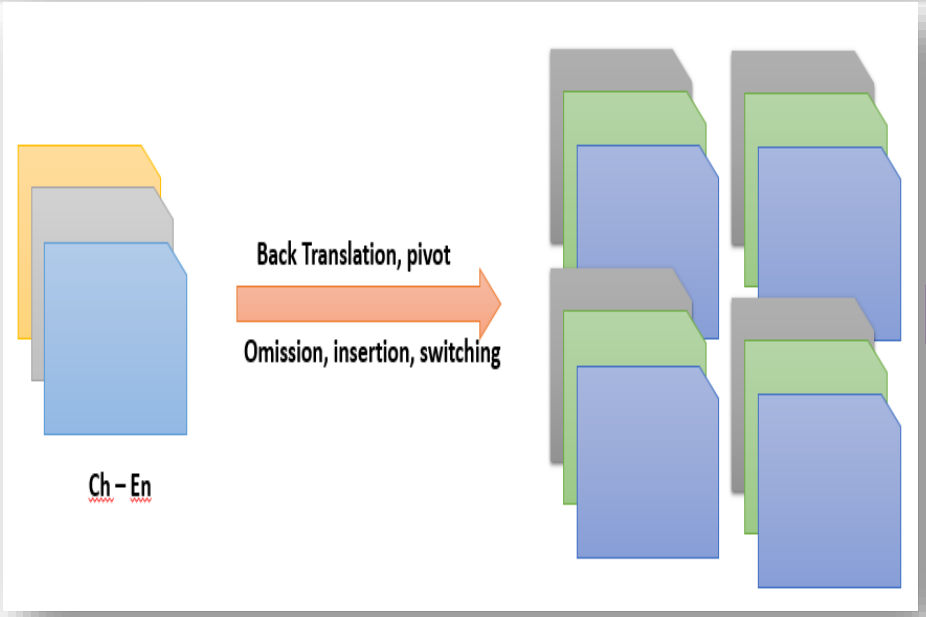
内容提要

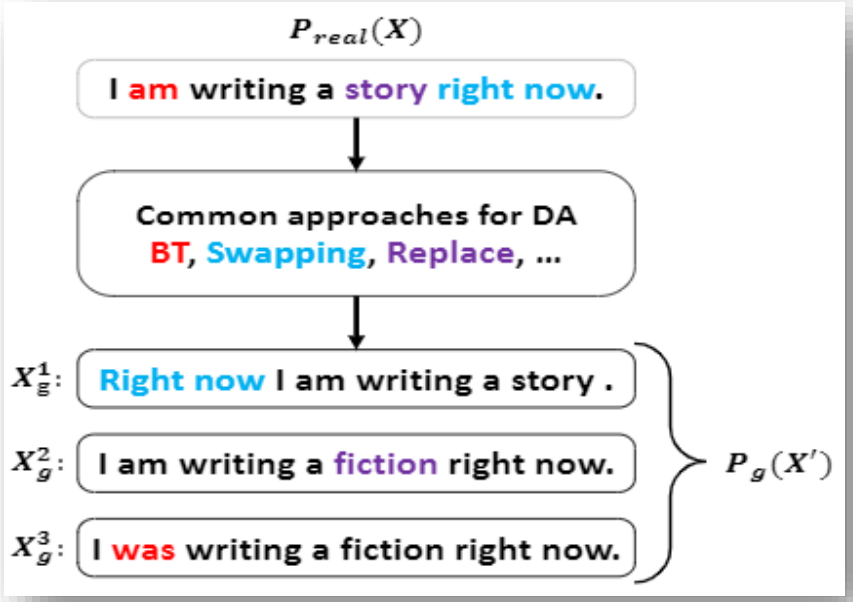
- ✓研究背景及意义
- ✓相关工作及研究现状
- ✓面临的问题与挑战
- ●研究工作
  - ✓基于自监督方法的预训练词切分模型
  - ✓基于迁移学习的低资源语言神经机器翻译
  - ●基于数据增强的低资源语言神经机器翻译
- ●总结与展望

基于数据增强的低资源语言NMT

- 动机：
  - 低资源场景下数据增强是一种有效、常用的方法， 但是**很难保证**生成的**伪数据**的**流利度**和**忠实度**。
  - 前人常用的方法都是词级别的**随机**替换、交换以及删除的方法进行的，很难避免**语义和句法**错误。
  - 我们提出了一个**无监督的**、不需要任何外部信息（词典）的**约束采样方法**，并提出了一个**评价模型**。
  - 同时，我们提出了**基于复述表和词性的**数据增强模型，能够有效降低语义和句法错误。

数据增强背景





常用的几个方法

核心思想

Original pair

S: $s_1, \cdots, s_i, \cdots, s_n$

T: $t_1, \cdots, t_j, \cdots, t_n$

Augmented pair

S: $s_1, \cdots, s_i', \cdots, s_n$,

T: $t_1, \cdots, t_j', \cdots, t_n$,

语法错误（语义+句法）

Sentence [original / substituted]	Plausible
My sister drives a [car / motorbike]	yes
My uncle sold his [house / motorbike]	yes
Alice waters the [plant / motorbike]	no (semantics)
John bought two [shirts / motorbike]	no (syntax)

方法1

- 训练集 $\mathrm{x} = x_1, \cdots, x_i, \cdots, x_I$

  $y = y_1, \cdots, y_j, \cdots, y_J$

  $\{< x^{(m)}, y^{(m)} >\}_{m=1}^{M}$

- 增强目标 $P_r(x) \rightarrow P_r(\tilde{x})$
- 增强概率 $P_f(\tilde{x}|x)$
- 增强的分布

$$P_f(\tilde{x}) = \mathbb{E}_{x \sim P_r(x)}[P_f(\tilde{x}|x)]$$

x (I like the book) → $\tilde{x}$ (I like a movie)

- $d = 2, p_1 = 3, p_2 = 4$

  $w_1 = \text{"a"}$ $w_2 = \text{"movie"}$

- 按照Miao et al. (2018) 的方法实现约束采样过程

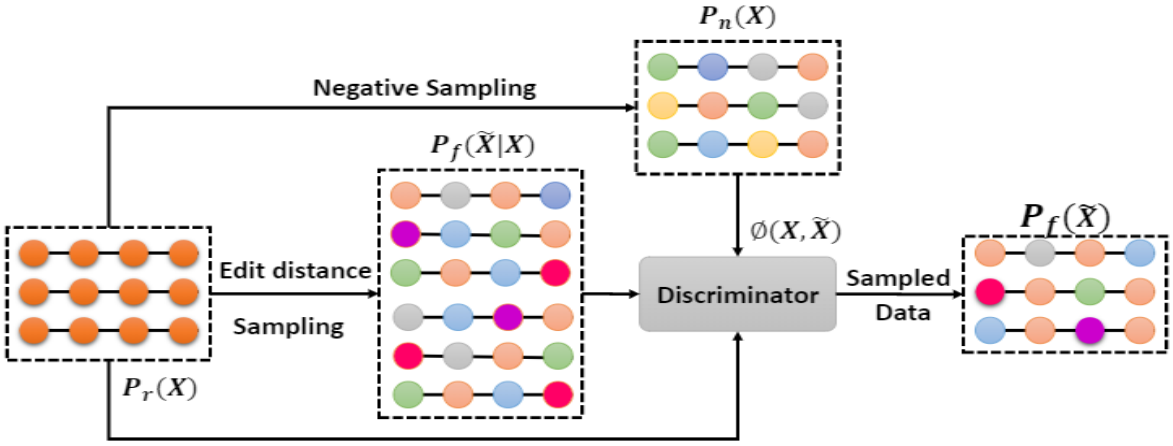
- 最终的增强过程类似于 Norouzi et al. (2016)

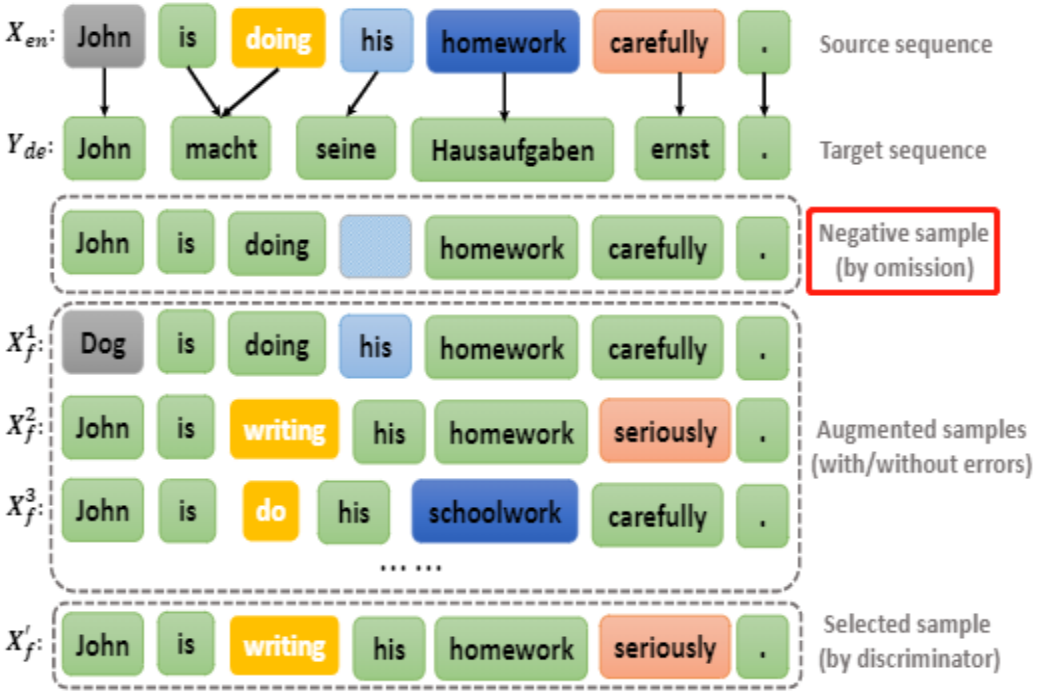
$$\begin{aligned} & P_f(\tilde{x}|x) \\ &= P(d, \boldsymbol{p}, w|x) \\ &= P(d|x)P(\boldsymbol{p}|x, d)P(w|x, d, \boldsymbol{p}) \\ &= P(d|x)\prod_{i=1}^{d} P(p_i|x, d)P(w_i|x, d, \boldsymbol{p}) \end{aligned}$$

- 使用Least-Square GAN (Mao et al., 2017) 实现评价子模型

$$\begin{aligned} \mathcal{L}_{\emptyset} = & \frac{1}{2}\mathbb{E}_{x \sim P_r(x)}[(\emptyset(x) - 1)^2] \\ & + \frac{1}{2}\mathbb{E}_{x \sim P_r(x)}\left[(\emptyset(x))^2\right] \end{aligned}$$

模型架构1


$P_n(X)$
Negative Sampling
$P_f(\tilde{X}|X)$
$\emptyset(X,\tilde{X})$
$P_f(\tilde{X})$
Edit distance
Sampling
Discriminator
Sampled
Data
$P_r(X)$


X_{en}: John is doing his homework carefully .
Source sequence
Y_{de}: John macht seine Hausaufgaben ernst .
Target sequence
John is doing homework carefully .
Negative sample (by omission)
X_f^1: Dog is doing his homework carefully .
X_f^2: John is writing his homework seriously .
X_f^3: John is do his schoolwork carefully .
Augmented samples (with/without errors)
X_f': John is writing his homework seriously .
Selected sample (by discriminator)

实验设置

- 数据集
  - Tanzil 数据集 Az – En, Hi – En, Ug – En, Uz – En 和 Tr – En
  - WMT14 En – De; IWSLT14 De – En; IWSLT15 Vi – En
  - WMT17 baseline BT, Copy, LM

数据属性

Languages	Train	Dev	Test	Source			Target		
				Voc.	Word	Avglen.	Voc.	Word	Avglen.
Az – En	21.2K	1.0K	1.0K	24.6K	3.1M	14.50	18.9K	4.0M	18.77
Hi – En	182.0K			13.3K	7.3M	39.97	20.1K	5.0M	27.09
Ug – En	81.1K			18.3K	2.1M	25.61	19.5K	2.2M	26.45
Uz – En	134.6K			25.2K	2.2M	15.97	20.1K	2.5M	18.12
Tr – En	141.9K			86.0K	0.8M	5.85	54.4K	1.0M	6.86
En – De	4.5M	6.0K	2.7K	817.0K	116.1M	26.00	1.7M	109.7M	24.5
De – En	160K	7.3K	6.8K	113.5K	3.1M	19.35	53.3K	3.3M	20.44
Vi – En	140.5K	1.6K	1.3K	25.3K	3.5M	24.64	48.64K	2.9M	20.10

实验结果

低资源语言（Tanzil corpus）数据增强效果

Method	Az – En	Hi – En	Ug – En	Uz – En	Tr – En
Trans (Vaswani et al., 2017)	21.03	20.15	19.19	17.76	22.72
BT (Sennrich et al., 2016)	21.32	19.20	20.01	18.72	24.95
Copy (Currey et al., 2017)	20.34	19.80	20.35	17.48	23.82
Swap (Artetxe et al., 2017)	30.32	21.33	21.76	19.21	25.08
Drop (Iyyer et al., 2015)	21.19	20.78	21.72	19.30	25.77
Blank (Xie et al., 2017)	23.23	20.40	21.67	19.39	25.44
Smooth (Xie et al., 2017)	25.88	21.87	22.24	19.48	25.85
Switch (Wang et al., 2018)	**26.36**	**22.61**	**23.16**	19.65	25.54
SCA (Zhu et al., 2019)	25.32	22.16	22.90	**19.77**	**25.92**
$Augment_{source}$	27.37	23.53	22.94	**21.22**	26.17
$Augment_{target}$	26.76	22.74	23.43	19.76	26.04
$Augment_{source+target}$	**27.59**	**23.68**	**23.67**	20.14	**26.66**

实验结果

WMT14和IWSLT14, 15数据集上的数据增强效果

Method	WMT14	IWSLT14	IWSLT15
	En – De	De – En	Vi – En
Trans (Vaswani et al., 2017)	27.24	33.53	25.32
BT (Sennrich et al., 2016)	27.30	33.69	26.34
Copy (Currey et al., 2017)	27.27	34.62	26.45
Swap (Artetxe et al., 2017)	27.19	33.98	26.98
Drop (Iyyer et al., 2015)	27.22	34.68	27.35
Blank (Xie et al., 2017)	27.32	34.83	27.80
Smooth (Xie et al., 2017)	27.48	34.85	**29.31**
Switch (Wang et al., 2018)	27.39	34.75	28.58
SCA (Zhu et al., 2019)	**27.51**	**34.89**	29.23
$Augment_{source}$	27.57	34.93	**29.88**
$Augment_{target}$	27.63	34.98	29.61
$Augment_{source+target}$	**27.94**	**35.14**	29.79

实验结果

评价模型的影响

Discriminator	Tanzil Corpus		IWSLT Corpus	
	Ug – En	Tr – En	De – En	Vi – En
×	23.25	26.16	34.85	29.61
√	**23.67**	**26.66**	**35.14**	**29.79**

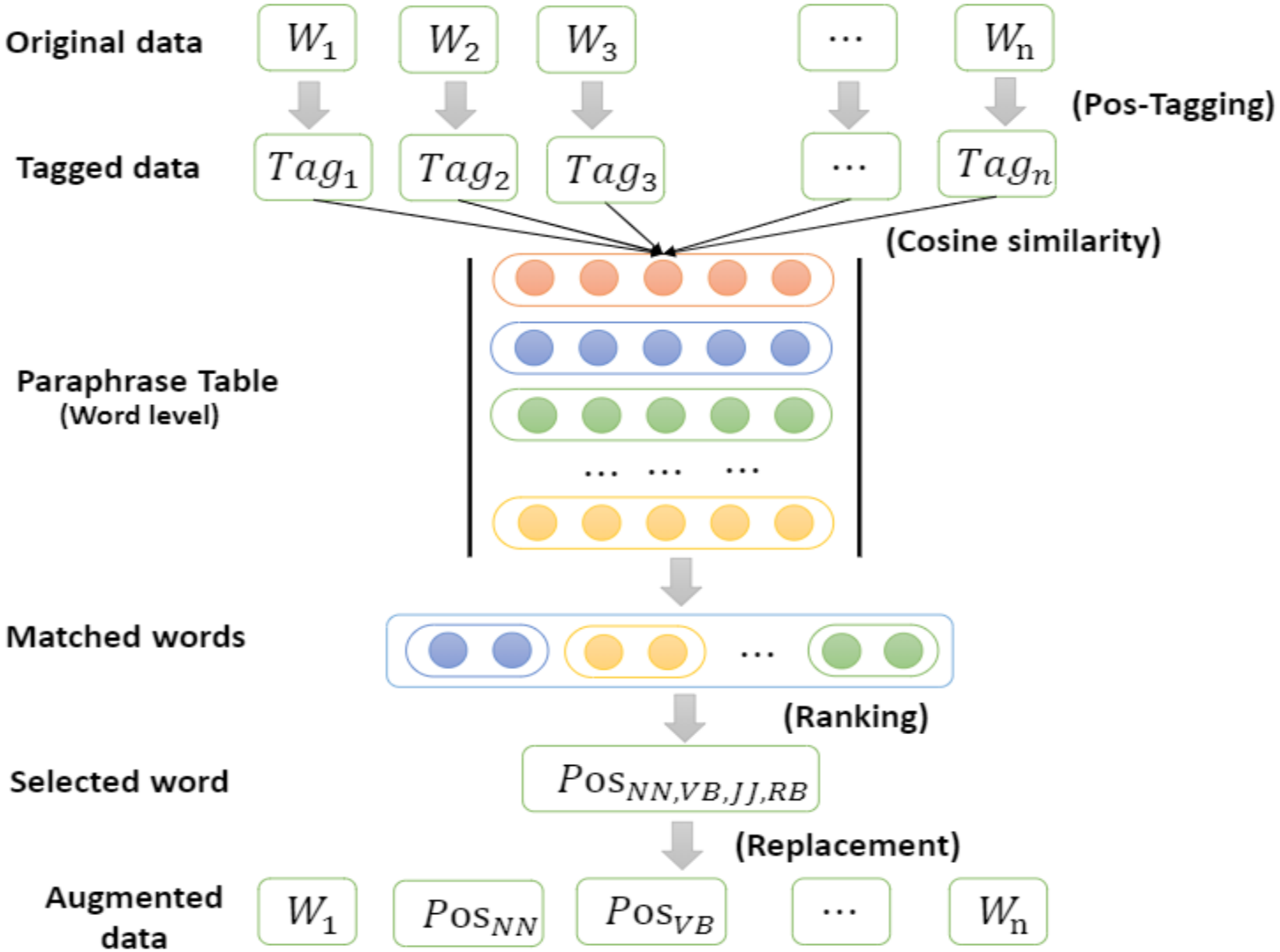
TER score (↓) 比较

Method	Az – En	Hi – En	Ug – En	Uz – En	Tr – En
Trans	66.72	66.73	69.51	76.96	68.74
BT	67.13	69.07	71.13	75.54	68.88
Copy	68.79	69.80	80.29	79.00	68.89
Swap	67.38	70.35	65.88	70.57	66.07
Drop	67.74	72.49	68.25	71.85	65.13
Blank	64.60	68.88	68.18	69.71	65.61
Smooth	60.64	67.51	66.85	70.95	64.39
Switch	**60.38**	63.59	**64.68**	**69.39**	65.60
SCA	60.90	**62.32**	65.98	71.47	**64.15**
Aug_{src}	**58.27**	**60.14**	**61.47**	**66.89**	**64.04**

流利度 (PPL ↓) 比较

Method	Az – En	Hi – En	Ug – En	Uz – En	Tr – En
Trans	23.28	22.62	22.33	23.98	23.64
BT	22.78	21.49	20.23	22.19	23.47
Copy	22.88	21.51	21.30	23.70	23.60
Swap	21.16	18.19	17.34	19.34	23.26
Drop	**20.73**	21.30	17.81	**18.58**	23.40
Blank	23.14	21.94	17.35	19.48	23.48
Smooth	22.24	17.25	18.52	20.31	23.45
Switch	21.57	**17.17**	**16.81**	18.94	23.19
SCA	21.20	16.42	17.34	19.42	**22.86**
Aug_{src}	18.47	16.31	16.66	18.27	21.98

模型架构2


Original data
W_1 W_2 W_3 … W_n
(Pos-Tagging)
Tagged data
Tag_1 Tag_2 Tag_3 … Tag_n
(Cosine similarity)
Paraphrase Table
(Word level)
Matched words
(Ranking)
Selected word
$Pos_{NN,VB,JJ,RB}$
(Replacement)
Augmented data
W_1 Pos_{NN} Pos_{VB} … W_n

实验设置

- 数据集
  - Tanzil 数据集 Az, Hi, Ug, Uz 和Tr
  - WMT14 De – En
  - LDC Zh – En
  - IWSLT15 Vi – En
  - PPDB 复述表 (只用词级别)
  - WMT17 BT, word2vec

词性统计

POS_{Label}	Az – En	De – En	Hi – En	Vi – En	Tr – En	Uz – En	Zh – En
POS_{NN}	779.2K	798.7K	961.0K	785.5K	4.1M	498.2K	8.0M
POS_{VB}	865.3K	62.2K	1.1M	513.5K	4.8M	537.3K	4.0K
POS_{JJ}	192.6K	235.3K	246.1K	207.2K	1.0M	122.9K	2.8M
POS_{RB}	192.4K	188.5K	240.1K	150.1K	1.0M	118.4K	912.4K
POS_{ALL}	2.0M	1.3M	2.5M	1.7M	11M	1.3M	15.6M

数据属性

Languages	Train	Dev	Test	Source			Target		
				Voc.	Word	Avglen.	Voc.	Word	Avglen
Az – En	21.2K	1.0K	1.0K	24.6K	3.1M	14.50	18.9K	4.0M	18.77
Hi – En	182.0K			13.3K	7.3M	39.97	20.1K	5.0M	27.09
Uz – En	134.6K			25.2K	2.2M	15.97	20.1K	2.5M	18.12
Tr – En	141.9K			86.0K	0.8M	5.85	54.4K	1.0M	6.86
De – En	160K	7.3K	6.8K	113.5K	3.1M	19.35	53.3K	3.3M	20.44
Vi – En	140.5K	1.6K	1.3K	25.3K	3.5M	24.64	48.64K	2.9M	20.10
Zh – En	1.0M	0.9K	0.9K	191.3K	22.7K	22.68	97.8K	28.4M	28.42

实验结果 --- Zh – En

BLEU score (↑) 比较

Method	10K	50K	100K	500K	1M
Trans	4.73	15.73	23.01	29.54	30.61
BT	4.75	25.88	22.95	29.63	30.56
Swap	4.80	16.01	22.94	29.43	30.59
Drop	4.84	16.41	23.10	29.75	31.22
Switch	4.89	16.49	23.18	30.23	31.24
SCA	4.94	16.53	23.28	30.40	31.34
POS_{NN}	5.12	16.76	23.36	30.51	31.37
POS_{VB}	5.39	16.85	23.43	30.60	31.42
POS_{JJ}	**5.46**	16.88	**23.74**	**30.66**	31.46
POS_{RB}	5.07	**16.99**	23.58	30.63	**31.51**

METEOR score (↑) 比较

Method	10K	50K	100K	500K	1M
Trans	10.91	21.93	28.83	32.97	34.04
BT	10.95	22.09	28.71	33.00	33.58
Swap	11.24	22.13	28.88	32.87	33.66
Drop	11.28	23.41	28.90	33.22	34.05
Switch	11.35	23.65	29.03	33.49	34.24
SCA	11.37	23.73	29.04	33.57	34.32
POS_{NN}	11.41	23.77	29.05	33.58	34.33
POS_{VB}	11.81	23.78	29.06	33.61	34.34
POS_{JJ}	11.85	23.88	29.08	**33.92**	34.36
POS_{RB}	**11.40**	**23.91**	**29.07**	33.62	**34.38**

复述表规模的影响

Method	Az – En	Hi – En	Uz – En
$Paraphrase_{T-S}$	25.89	25.56	20.42
$Paraphrase_{T-X}$	26.38	26.23	21.10
$Paraphrase_{T-XL}$	27.09	29.99	21.79

实验结果

低资源语言 BLEU score (↑) 比较

Method	Az – En	Hi – En	Tr – En	Uz – En	De – En	Vi – En
Trans (Vaswani et al., 2017)	21.03	20.15	22.72	16.76	33.53	25.32
BT (Sennrich et al., 2016)	21.32	19.20	24.95	18.72	33.69	26.34
Swap (Artetxe et al., 2017)	20.32	21.33	25.08	19.21	33.98	26.98
Drop (Iyyer et al., 2015)	21.19	20.78	25.77	19.30	34.68	27.35
Switch (Wang et al., 2018)	26.36	22.61	25.54	19.65	34.75	28.58
SCA (Zhu et al., 2019)	25.32	22.16	25.92	19.77	34.89	29.23
POS_{NN}	26.45	22.87	25.98	19.82	34.96	29.65
POS_{VB}	26.58	23.01	26.07	20.32	34.91	29.80
POS_{JJ}	26.71	24.29	26.00	20.41	34.92	**30.05**
POS_{RB}	**26.73**	**24.82**	**26.16**	**20.62**	35.02	29.93

随机替换单词和同词性替换的影响

Method	Az – En	Hi – En	Uz – En	De – En	Vi – En	Zh – En_{100K}
Baseline	26.36	22.61	19.77	34.89	29.23	23.28
Paraphrase + random	26.01	25.16	20.49	34.90	29.24	23.33
Paraphrase + POS-Taggings	**27.09**	**26.99**	**21.79**	**35.39**	**30.15**	**24.19**

加快训练速度

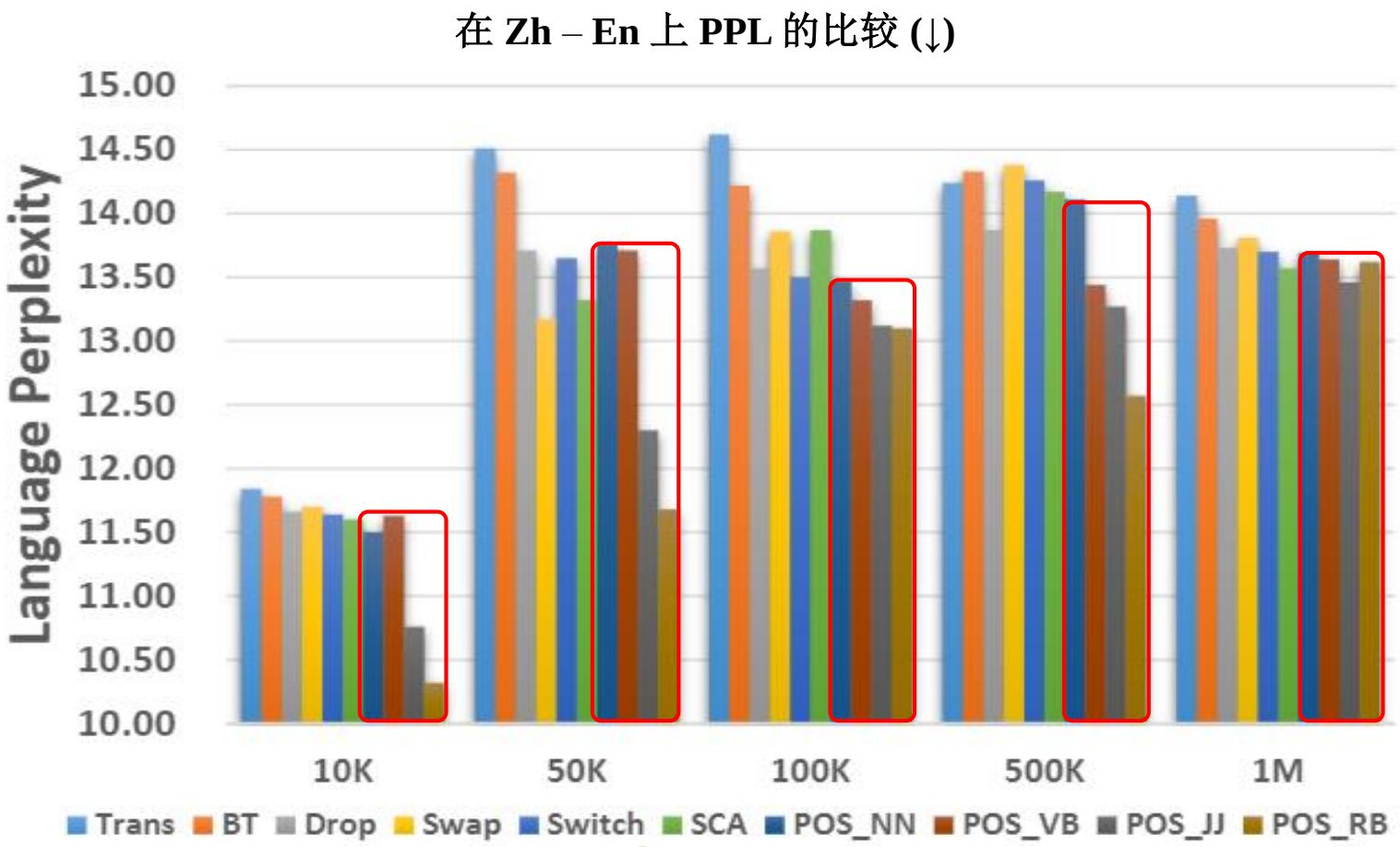
实验结果

替换多种不同词性的影响比较

Method	Az – En	Hi – En	Uz – En	De – En	Vi – En	Zh – En_{100K}
POS_{NN}	26.45	22.87	25.98	19.82	34.96	23.36
POS_{VB}	26.58	23.01	26.07	20.32	34.91	23.43
POS_{JJ}	26.71	24.29	26.00	20.41	34.92	23.74
POS_{RB}	26.73	24.82	26.16	20.62	35.02	23.58
POS_{NN-VB}	26.02	25.50	21.23	34.55	29.96	24.06
POS_{NN-JJ}	26.58	25.08	20.73	34.84	30.03	23.82
POS_{NN-RB}	26.68	24.49	20.71	35.04	29.84	24.11
POS_{VB-JJ}	25.56	24.76	20.32	35.16	**30.15**	24.01
POS_{VB-RB}	26.76	25.23	21.47	34.98	30.10	23.89
POS_{JJ-RB}	26.88	25.56	20.88	**35.39**	29.92	23.93
$POS_{NN-VB-JJ}$	26.90	24.43	21.55	35.35	29.77	23.85
$POS_{NN-VB-RB}$	26.91	**26.99**	20.98	34.75	30.13	**24.19**
$POS_{NN-JJ-RB}$	26.93	24.82	20.91	34.68	30.09	23.98
$POS_{VB-JJ-RB}$	**27.09**	25.11	**21.79**	35.32	28.90	23.90

针对于低资源语言**同时替换多个词性**时的影响比较明显，对于高资源语言**未必**是越多越好

实验结果



本方法的语言复杂度相比于前人工作低很多，说明本方法生成的数据更流利

混合实验结果 --- TL+DA

TL+DA的混合策略对LRLs NMT的影响

Method	Az – Zh	Uz – Zh	Ug – Zh
Transformer	43.68	43.11	28.28
Transfer_Learning_Multi_Round	–	–	**33.91**
Transfer_Learning_Mixture	48.62	45.83	–
Data_Augmentation_Constrained_Sampling	**49.04**	46.39	33.59
Data_Augmentation_POS_Paraphrase	48.83	**47.11**	–
Data Augmentation + Transfer Learning	**49.78**	**48.05**	**35.63**

- 总结
  - 虽然基于**迁移学习**和**数据增强**的混合策略对低资源神经机器翻译有帮助，但是在不同的语言之间模型性能还是不太一致；
  - 最后还是靠实验结果来**启发式地**选择具体的方法；
  - 简单、易用、**有效**

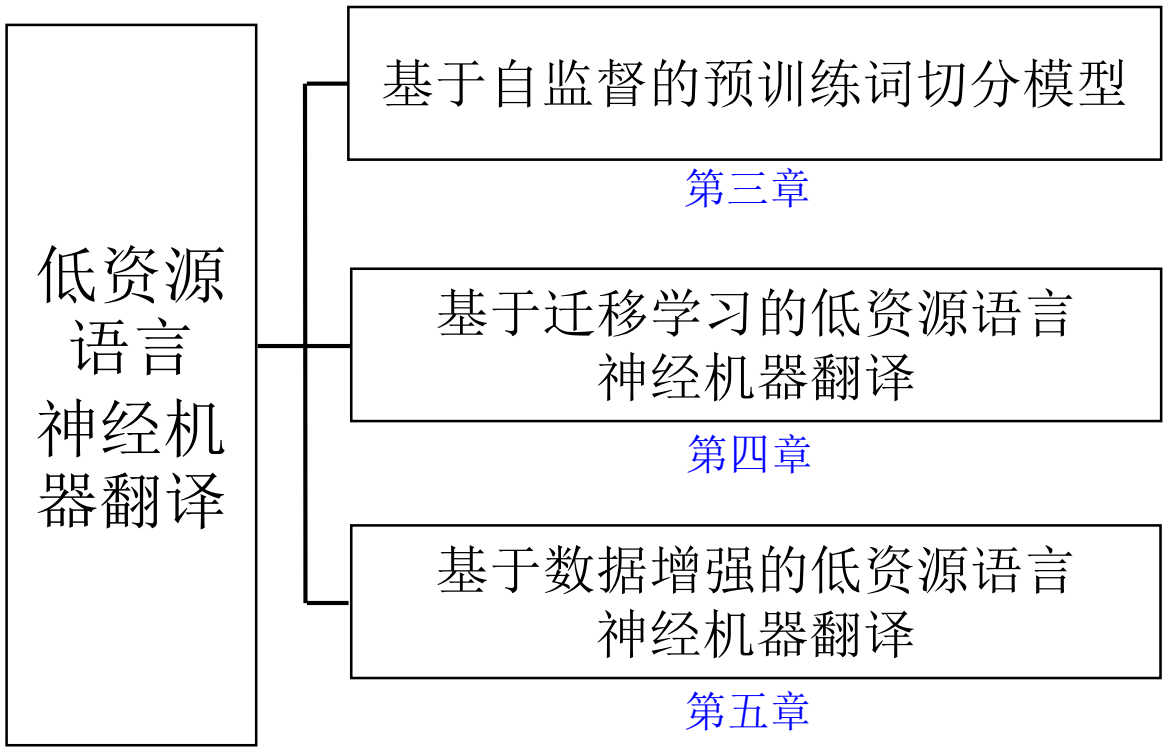
小结

- 我们提出了**两种基于数据增强**的低资源语言NMT模型。
- 提出了一个基于约束采样的数据增强方法，在8种语言对上和前人的9个基线系统比较，实验结果说明我们提出的模型有一定的效果。
- 为了尽可能的**减少生成的数据包含的语义或者句法错误**，我们提出了质量评价模型。
- 提出了一个基于**复述表**和**同词性单词替换**的数据增强方法。
- 两种不同的数据增强模型的流利度均有提高，说明我们的数据生成方法**一定程度上避免了**伪数据的语法错误。

内容提要

- ✓研究背景及意义
- ✓相关工作及研究现状
- ✓面临的问题与挑战
- ✓研究工作
  - ✓基于迁移学习的低资源语言神经机器翻译
  - ✓基于数据增强的低资源语言神经机器翻译
  - ✓基于混合模型的无监督领域自适应模型
- ●总结与展望

研究工作总结



针对于低资源场景下已有的词切分系统性能鲁棒性差、模型架构复杂

针对于低资源场景下无法高效并充分利用高资源语言帮助低资源语言

针对于低资源场景下生成的伪数据质量不好，难以避免语义和句法错误

研究工作展望

- 在研究词切分时，可以进一步探索最简单、最稳定的方法；此外，可以进一步研究**完全无监督**的词切分方式。
- 在利用高资源语言方面，可以探索高资源语言和**完全零资源语言**之间的连接，甚至如何利用**完全不接近**的高资源语言帮助低资源语言。
- 在数据生成方面，可以从**短语级别**和**句子级别**出发进一步研究，同时可以尝试不同的**采样方式**来训练评价模型。

论文

- **Mieradilijiang Maimaiti**, Xiaohui Zou. Discussion on Bilingual Cognition in International Exchange Activities. International Conference on Intelligence Science (ICIS2018), 2018.11. (EI检索)
- **Mieradilijiang Maimaiti**, Shunpeng Zou, Xiaoqun Wang, and Xiaohui Zou. How to Understand: Three Types of Bilingual Information Processing? International Conference on Computer System and Information Processing (ICCSIP2018), 2018.11. (EI检索)
- **Mieradilijiang Maimaiti**, Yang Liu, Huanbo Luan, and Maosong Sun. Multi-Round Transfer Learning for Low-Resource NMT Using Multiple High-Resource Languages. In ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2019.1. (CCF C类，SCI检索[3区]期刊)
- **Mieradilijiang Maimaiti**, Yang Liu∗, Huanbo Luan, and Maosong Sun. Enriching the Transfer Learning with Pre-trained Lexicon Embedding for Low-Resource Neural Machine Translation. TSINGHUA SCIENCE AND TECHNOLOGY (TST), 2020.8. (SCI检索[3区]期刊)
- **Mieradilijiang Maimaiti**, Zegao Pan, Yang Liu∗, Huanbo Luan, and Maosong Sun. Improving the Data Augmentation for Low-Resource NMT Using Pos-Tagging and Paraphrase Embedding. In ACM Transactions on Asian and Low-Resource Lan-guage Information Processing (TALLIP), 2021.12. (CCF C类，SCI检索[3区]期刊)

论文

- **Mieradilijiang Maimaiti**, Yang Liu∗, Zhixing Tan, Jiacheng Zhang, Huanbo Luan, and Maosong Sun. Data Augmentation for Low-Resource NMT Guided by Constrained Sampling. In INTERNATIONAL JOURNAL OF INTELLIGENT SYSTEMS (IJIS), 2020.12. (CCF C类, SCI检索[1区]期刊) [minor revision]
- **Mieradilijiang Maimaiti**, Yang Liu∗, Yuanhang Zheng, Gang Chen, Kaiyu huang, Ji Zhang, Huanbo Luan, and Maosong Sun. Segment, Mask, and Predict: Self-supervised Word Segmentation. In Proceedings of the 2021 conference on Empirical Methods in Natural Language Processing, 2021.05. (CCF B类，会议) [under review]
- Yuanhang Zheng, Zhixing Tan, Meng Zhang, **Mieradilijiang Maimaiti**, Huanbo Luan, Maosong Sun, Qun Liu, and Yang Liu∗. Self-Supervised Quality Estimation for Machine Translation. In Proceedings of the 2021 conference on Empirical Methods in Natural Language Processing, 2021.05. (CCF B类，会议) [under review]
- Zhe Li∗, **Mieradilijiang Maimaiti∗**, Jiabao Sheng, Zunwang Ke, Wushour Slamu∗,Qinyong Wang, Xiuhong Li. An Empirical Study on Deep Neural Network Models for Chinese Dialogue Generation. In Symmetry-Basel, 2020.10. (SCI检索[3区]期刊)

专利

- 已授权
  - **米尔阿迪力江·麦麦提**，刘洋，栾焕博，孙茂松，“一种基于无监督领域自适应的神经网络机器翻译方法”，计算机发明专利，授权号：CN 107038159 A;申请号：201710139214.0，**已授权**；(对应论文第3章)
  - 孙茂松,**米尔阿迪力江·麦麦提**，刘洋，栾焕博，“神经网络机器翻译模型的训练方法和装置”，计算机发明专利，授权号：CN 109117483 B;申请号：201810845896.1，**已授权**；(对应论文第4章)

软著

- 已登记
  - **米尔阿迪力江 麦麦提**，“维吾尔语人工词性标注及语料库构建系统”，登记号：2016SR031180，已发布，**已登记**，2016年2月；
  - **米尔阿迪力江 麦麦提**，“维吾尔语自动词性标注系统”，登记号：2016SR052763，已发布，**已登记**，2016年3月；
  - **米尔阿迪力江 麦麦提**，“维吾尔语自动词干提取与词性标注系统”，登记号：2016SR379408，已发布，**已登记**，2016年12月；
  - 孙茂松,**米尔阿迪力江 麦麦提**，刘洋，栾焕博，“基于Python的多语种多文种自动编码转换工具软件(1.0)”，登记号：2019SR0110291，未发布，**已登记**，2019年1月；
  - 孙茂松,**米尔阿迪力江 麦麦提**，刘洋，栾焕博，“面向低资源语言的多语种在线机器翻译系统(1.0)”，登记号：2019SR0108620，已发布，**已登记**，2019年1月；

学术报告

- **Mieradilijiang Maimaiti**, “Introduction of Low-Resource Neural Machine Translation”, Xinjiang University, Urumqi, Xinjiang, P.R. China, Jun.15.2017–Jun.18.2017;
- **米尔阿迪力江·麦麦提**，“MLWS2017维吾尔语词干提取评测报告”，第十六届全国少数民族语言文字信息处理学术研讨会，桂林，中国，2017年9月21-22日；
- **Mieradilijiang Maimaiti**, “Unsupervised Domain Adaptation for Low-Resource Neural Machine Translation”, Peking University, Beijing, P.R. China, Sept.14.2018
- **Mieradilijiang Maimaiti**, “Discussion on Bilingual Cognition in International Ex-change Activities”, International conference on intelligence science (ICIS2018), Beijing, P.R. China, Nov.1.2018–Nov.4.2018
- **Mieradilijiang Maimaiti**, “Improving the Low-Resource Neural Machine Translation between Morphologically Rich Languages”, Xinyi Tech INC, Shen Zhen, P.R. China, July.15.2019
- **Mieradilijiang Maimaiti**, “Data Augmentation for Low-Resource Neural Machine Translation Using Different Sampling”, Minzu University of China, Beijing, P.R. China, Nov. 12. 2019

参与的项目

- 国家科技支撑计划项目
  - 项目名称：“少数民族网络舆情综合分析与云服务关键技术研究及应用示范”
  - 批准号：2014BAK10B03
  - 状态：已结题
  - 主要贡献：开发多语种网站数据采集器，并开发了维-汉、藏-汉、蒙-汉双向在线机器翻译系统

1.5.245:9662/index.html

清华大学多语种翻译系统

چىڭخۇا ئۇنىۋېرسىتى كۆپ تىللىق تەرجىمە سىستېمىسى

维文 >> 汉语 通用领域 翻译

维文
汉语
藏文
蒙文

ئىنقلابىي قۇربانلارنى خاتىرىلەش كۈنىدە خەلق قەھرىمانلىرىغا گۈل سېۋېتى تەقدىم قىلىش مۇراسىمى بېيجىڭدا داغدۇغلىق ئۆتكۈزۈلدى.

革命先烈纪念日人民英雄敬献花篮的仪式在北京隆重举行

最多可以输入500个字符

Email: miradel51@126.com Tel: 13051308938 Wechat: 821777278

参与的项目

- 国家自然科学基金重点项目
  - 项目名称：“跨语言社会舆情分析基础理论与关键技术研究”
  - 批准号：61331013
  - 状态：已结题
  - 主要贡献：开发了跨语言信息检索系统，分别爬取了维吾尔语、藏语、蒙语网站，调用了自主研发的小语种机器翻译接口实现跨语言搜索

清华大学跨语言信息检索系统

查询语言

- 中文

文档语言

- 维吾尔语
- 藏语
- 蒙古语

请输入关键词...

编号	标题	日期	点击率
1	ཨཨ།།བོད་སྐྱོང་ས་ཁུལ་ཨོག་གཞི་བྱོན་ཅེ་ཤོས་ཀྱི་ལྷག་ཆག་འགྲན་ཚོགས་གཏེར་བ།	2018-07-25	ས་གཞུས: 213
2	ཞི་ཕིན་ཡིང་ཁྱབ་ཐོ་ཤེ་ཁི་ཀའི་ནུང་ཐུང་ལ་ཐ་ཆུ་ག་དང་མཐའ་ཚོལ་གནང་བ།	2018-07-25	ས་གཞུས: 191
3	ཡིང་ཁ་ཙོན་བུ་བཙོལ་ཁང་ཁི་དགེ་ཆུར་ཀྱི་སྐྱོང་བཙན་སྐྱབ་ཚོན་མཐོང་སྐོར་ཞུགས་གནང་བ།	2018-07-25	ས་གཞུས: 171
4	བ།ཉ་ཆེན་སྐུ་ཕྲེང་བསུ་གཅིག་པ་མཆོག་གནོད་གཡག་ཁང་བུ་ཞལ་ས་ཚོགས་ཤོར་འགོད་ནས་ཆུལ་ཀྱིད་དུ་ཞིང་བཤགས་རྟེན་བའི་བསྟན་འཛིན་ཞུགས་པ།	2018-07-24	ས་གཞུས: 163
5	ཁི་ཏྲོན་ཇ་བའི་མང་ཚོགས་ཁྲི་ཙོན་ཀྱིས་སྐུ་ཁྲ་གནས་ཁ་བསྒོར་བ།	2018-07-24	ས་གཞུས: 151

Email: miradel51@126.com Tel: 13051308938 Wechat: 821777278

清华大学跨语言信息检索系统

查询语言

- (•) 中文

文档语言

- (•) 维吾尔语
- ( ) 藏语
- ( ) 蒙古语

中国梦

中国梦

点击率	日期	标题	编号
ئاۋاتلىقى: 261	2018-07-26	شى جىنپىڭ كېسەك ئالتۇن دۆلەتلىرى سودا-سانائەت مۇنبىرىدىكە قاتناشتى ھەم مۇھىم سۆز قىلدى	1
ئاۋاتلىقى: 229	2018-07-26	شى جىنپىڭ رامافۇسا زۇكۈك ئۇنىكۇزگەن قارشى ئېلىش ۋە جۇكگو - جەنۇبىي ئافرىقا دىيلوماتىك...	2
ئاۋاتلىقى: 173	2018-07-26	سىرىدىۋىنىڭ كىچىك قىزى خۇشەمۇ بوللىۋۇدقا كىرىشى مۇمكىن	3
ئاۋاتلىقى: 149	2018-07-26	داۋالىنىش سۇغۇرتىسى ماليە يارىدەم پۇلى ئولچىمى كىشى بېشىغا بىرگىدىن 40 يۈەن قوشۇلدى	4
ئاۋاتلىقى: 123	2018-07-26	گوۋۇيۈەن تەكشۈرۈش گۇرۇپپىسى جىلىن ئولكىسىدىكە بىرىپ دىلونى تەكشۈرۈپ بىر تەرەپ قىلىش خىزمىتىنى ...	5

Email: miradel51@126.com Tel: 13051308938 Wechat: 821777278

清华大学跨语言信息检索系统

搜索

实现中华民族的伟大复兴

جەمئىي ئىزدەپ تاپقان ئۇچۇر: 7872 ، ئىشلەتكەن ۋاقىت:267 مكرو سكۇنت

باش شۇجى شى جىنپىڭ تەيۋەندىكى ھەر ساھە ئەربابلىرى بىلەن كۆرۈشتى- تەڭرىتاغ تورى

ئالدۇرۇپ، تىنچلىقنى قەدىرلەپ، تەرەققىياتنى بىرلىكتە پىلانلاپ، بىر نىيەتتە ئىتتىپاقلىشىپ جۇڭخۇا مىللەتلىرىنىڭ ئۇلۇغ گۈللىنىشىنى ئەمەلگە ئاشۇرۇش ئۈچۈن ...

http://uy.ts.cn/topic/kangri/2015_09/01/content_449260.htm 2017_02_28 22:52:34

شانلىق مۇسابە، پولاتتەك پاكىت- تەڭرىتاغ تورى

ئىز قالدۇردى. ھازىر شىنجاڭنىڭ تەرەققىياتى يېڭى تارىخىي باشلىنىش نۇقتىسىدا تۇرۇۋاتىدۇ، مەملىكەت خەلقى بىلەن بىللە، جۇڭخۇا مىللەتلىرىنىڭ ئۇلۇغ گۈللىنىشىنى ئەمەلگە

شانلىق يىپەك يولىنىڭ يېڭى سەھىپىسىنى بىرلىكتە ئاچايلى- تەڭرىتاغ تورى

قاتارلىق ئەۋزەللىكلەرنى ئەملي ھەمكارلىشىش، ئىقتىسادنى سىجىل ئاشۇرۇش، خىرىسلارغا ئورتاق تاقابىل تۇرۇش ئەۋزەللىكىگە ئىلاندۇردى. يىپەك يولى ئىقتىساد بەلبېغىنى ئورتاق قۇرۇش جۇڭخۇا مىللەتلىرىنىڭ ئۇلۇغ ...

http://uy.ts.cn/zhuanti/2014_09/04/content_372758.htm 2018_07_05 22:29:24

ئالدىنقىلارغا ۋارىسلىق قىلىپ، كېيىنكىلەرگە يول ئېچىش- تەڭرىتاغ تورى

ۋە جۇڭخۇا مىللەتلىرىنىڭ ئۇلۇغ گۈللىنىشىنى ئەمەلگە ئاشۇرۇشقا مۇناسىۋەتلىك» ، دەپ كۆرسەتتى. مەركەز ھەر مىللەت خەلقىگە چوڭقۇر...

http://uy.ts.cn/topic/60zhounian/2015_09/10/content_451221.htm 2018_07_05 22:33:45

شانلىق مۇسابە، پولاتتەك پاكىت- تەڭرىتاغ تورى

. ھازىر شىنجاڭنىڭ تەرەققىياتى يېڭى تارىخىي باشلىنىش نۇقتىسىدا تۇرۇۋاتىدۇ، مەملىكەت خەلقى بىلەن بىللە، جۇڭخۇا مىللەتلىرىنىڭ ئۇلۇغ گۈللىنىشىنى ئەمەلگە ئاشۇرۇشت...

http://uy.ts.cn/topic/60zhounian/2015_09/26/content_458155.htm 2018_07_08 09:07:16

جۇڭگو خەلق سىياسىي مەسلىھەت كېڭىشى مەملىكەتلىك 12-نۆۋەتلىك كومىتېتى 4-يىغىنىنىڭ ...

مىللەتلىرىنىڭ ئۇلۇغ گۈللىنىشىنى ئەمەلگە ئاشۇرۇش كېرەك. دۆلىتىمىزنىڭ سىرتقا قارىتىلغان خىزمەتلىرىدىكى ئومۇمىي ئورۇنلاشتۇرمىسىغا ئاساسەن، سىرت بىلەن بولغان دوستانە ئالاقىنى چوڭقۇر...

http://uy.ts.cn/2016lianghui/2016_03/15/content_509696.htm 2017_05_22 23:02:04

جۇڭگو خەلق سىياسىي مەسلىھەت كېڭىشى مەملىكەتلىك 12 - نۆۋەتلىك كومىتېتى 4 - يىغىنى...

参与的项目

- 973项目
  - 项目名称：“面向三元空间的互联网中文信息处理理论与方法”
  - 批准号：2014CB340500
  - 状态：已结题
  - 主要贡献：开发了神经网络维-汉双向机器翻译系统，目前正在扩充到“一带一路”沿线国家的几种语言之间的翻译

清华大学多语... | 清华大学跨语... | 多语种翻译系统 | 多语种翻译系统 | 天山网_百度搜... | 天山网 - 新疆... | تەڭرىتاغ تورى | ئىش ئورنىدا چىك تۇ...

.196:9666/index.html

晒照为大s庆

多语种翻译系统

كۆپ تىللىق تەرجىمە سىستېمىسى

维吾尔语 >> 汉语 通用领域 翻译

ئاممانا تۇزگەن چاسا ئەترەت ئالدىغا جۇڭگو كوممۇنىستىك پارتىيەسى مەركىزي كومىتېتى، مەملىكەتلىك خەلق قۇرۇلتىيى دائىمي كومىتېتى، گوۋۇيۈەن، مەملىكەتلىك سىياسي كېڭەش، مەركىزي ھەربي كومىتېت، ھەرقايسى دېموكراتىك پارتىيە–گۇرۇھلار، مەملىكەتلىك سودا–سانائەتچىلەر بىرلەشمىسى ۋە پارتىيە–گۇرۇھسىز ۋەتەنپەرۋەر زاتلار، ھەرقايسى خەلق تەشكىلاتلىرى ۋە ھەر ساھە ئاممىسى، پىشقەدەم جەڭچىلەر، پىشقەدەم يولداشلار ۋە ئىنقلابي قۇربانلارنىڭ ئائىلە تاۋابىئاتلىرى، جۇڭگو پىيونېرلار ئەترىتىنىڭ نامىدا تەقدىم قىلىنغان چوڭ تېپتىكى توققۇز گۈل سېۋىتى قاتار تىزىلغاندى.

最多可以输入500个字符

群众组成方队，面对的是中共中央。,全国人民代表大会常务委员会,国务院；,全国政协：,中央军事委员会；,各民主党派,全国工商联及无党派爱国人士,各人民团体和各界群众,老战士,老同志和革命先烈的家属。,以中国少先队命名的九个大型花束排列。

参与的项目

- 企业合作项目1
  - 企业名称：北京捷通华声科技股份有限公司
  - 项目名称：“基于深度学习的维汉机器翻译”
  - 批准号：N/A
  - 状态：已结题
  - 主要贡献：担任维-汉双向机器翻译引擎研发技术顾问，也提供技术指导以及相关的代码。

灵云
AIcloud.com

中文 ⇄ 维文

10月6日晚，“我爱你中国”主题灯光秀把夜色中的乌鲁木齐装点得分外绚丽。各族市民和众多来疆游客观灯赏景，欢声笑语，为中华人民共和国69周年华诞送上衷心的祝福。

10 - ئاينىڭ 6 - كۈنى كەچتە ، « سىزنى سۆيىمەن » دېگەن تېمىدىكى چىراق نۇرى كېچىدىكى ئۇرۇمچى شەھىردە نومۇر ئېلىپ ، گۈزەل تۇسكە كىردى .ھەر مىللەت شەھەر ئاھالىلىرى ۋە شىنجاڭغا كەلگەن نۇرغۇن سایاھەتچىلەرنىڭ بانۇس سەيلىسى ، كۆلكە سادالىرى جۇڭخۇا خەلق جۇمھۇرىيىتى قۇرۇلغانلىقىنىڭ 69 يىللىقىنى قىزغىن تەبرىكلىدى .

شكرا لك

شكريا

köszönöm

谢 谢！

ཐཱངཀཱ ཝ

བཀའ་དྲིན་ཆེ།

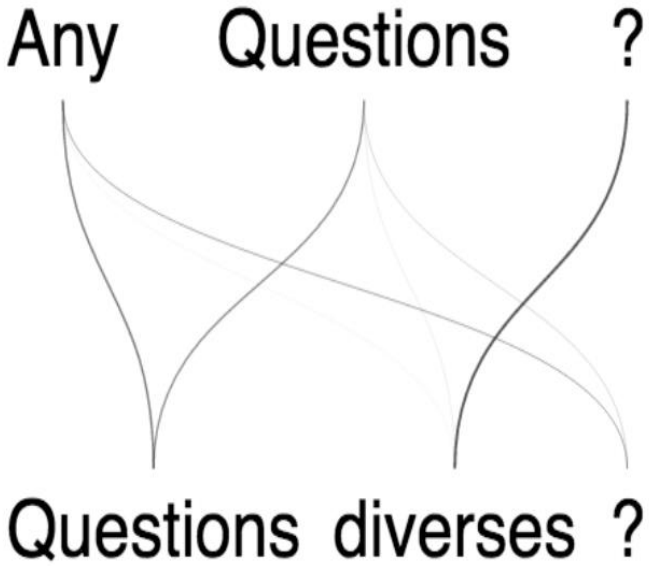
ره خمه ت!

הדות

Teşekkür

Kiitos

感谢各位老师参加我的博士答辩！
希望得到您的宝贵意见和建议！



This inspiration comes from Dzmitry Bahdanau @ *ICLR2014* .