华为诺亚方舟实验室机器翻译研究与应用

李良友

华为诺亚方舟实验室机器翻译团队

www.huawei.com

HUAWEI

目录

华为诺亚方舟实验室概况

基础研究

第三代人工智能

基础能力层

资源高效AI
...小数据、低能耗...

自动自治AI
...AutoML、元学习...

安全可信AI
...可解释、可推理、鲁棒性...

核心能力层

机器学习

语音语义

计算视觉

决策推理

能力沉淀至鲲鹏+昇腾生态

深入场景，差异化解决挑战问题

应用牵引，促进能力循环提升

解决商业场景挑战问题

创新应用研究

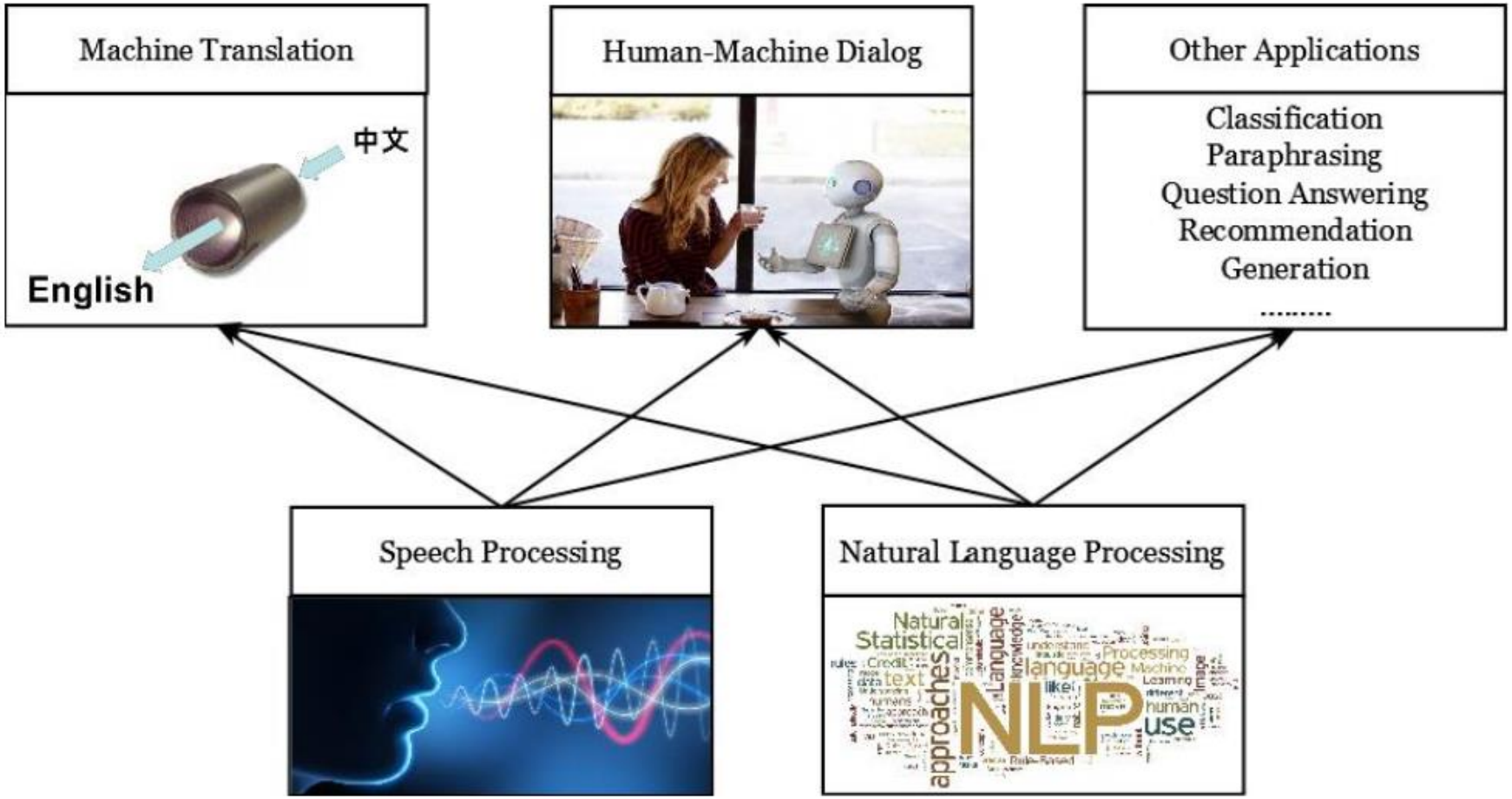
GTS大脑　供应大脑　终端计算视觉　终端语音语义

网络大脑　使能工程　终端搜索推荐　自动驾驶

网络智能　行业智能　终端智能　车载智能

MindSpore构建开发者生态

语音语义子实验室


Machine Translation
中文
English
Human-Machine Dialog
Other Applications
Classification
Paraphrasing
Question Answering
Recommendation
Generation
.....
Speech Processing
Natural Language Processing
Natural
Statistical
Language
Processing
language
text
approaches
NLP
human
use

语音语义子实验室全球布局

Beijing

Shenzhen

Hong Kong

London

Moscow

Montreal

目录

诺亚的机器翻译研究

- **神经机器翻译模型和算法**
- **多语言机器翻译**
- **低资源机器翻译**
- **实时语音翻译**
- **篇章级机器翻译**
- **机器翻译领域迁移**
- **鲁棒机器翻译**
- **……**

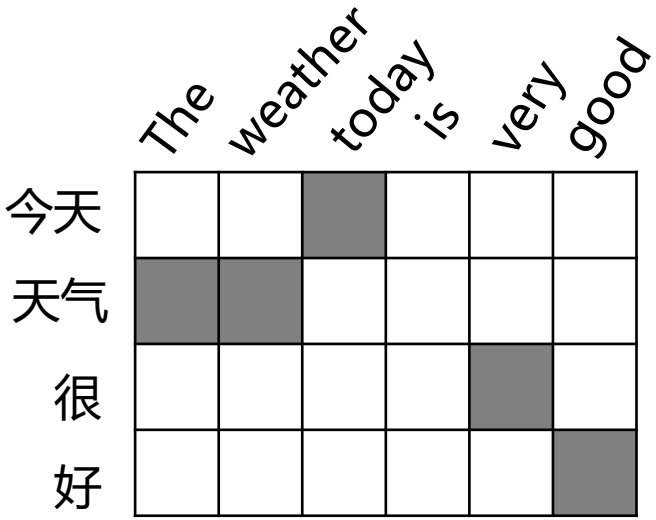
近期的一些工作简介

- 词语对齐
- 篇章机器翻译
- 低资源机器翻译
- 实时语音翻译

词语对齐

- **反映了词之间的互译关系，对于机器翻译非常重要**
  - 可解释性、可操控性、翻译知识挖掘等
  - 术语翻译和用户自定义词典翻译

- **神经机器翻译无法直接导出词语对齐信息**
  - 注意力机制的出发点是利用词语对齐信息并取得了很大的成功
  - 但是研究发现，从注意力矩阵导出的词语对齐是不准确的
  - 注意力不仅仅与词语对齐有关，而更多反映了词语的相关性
  - 一些研究引入额外的对齐模块代价高且需要监督信号

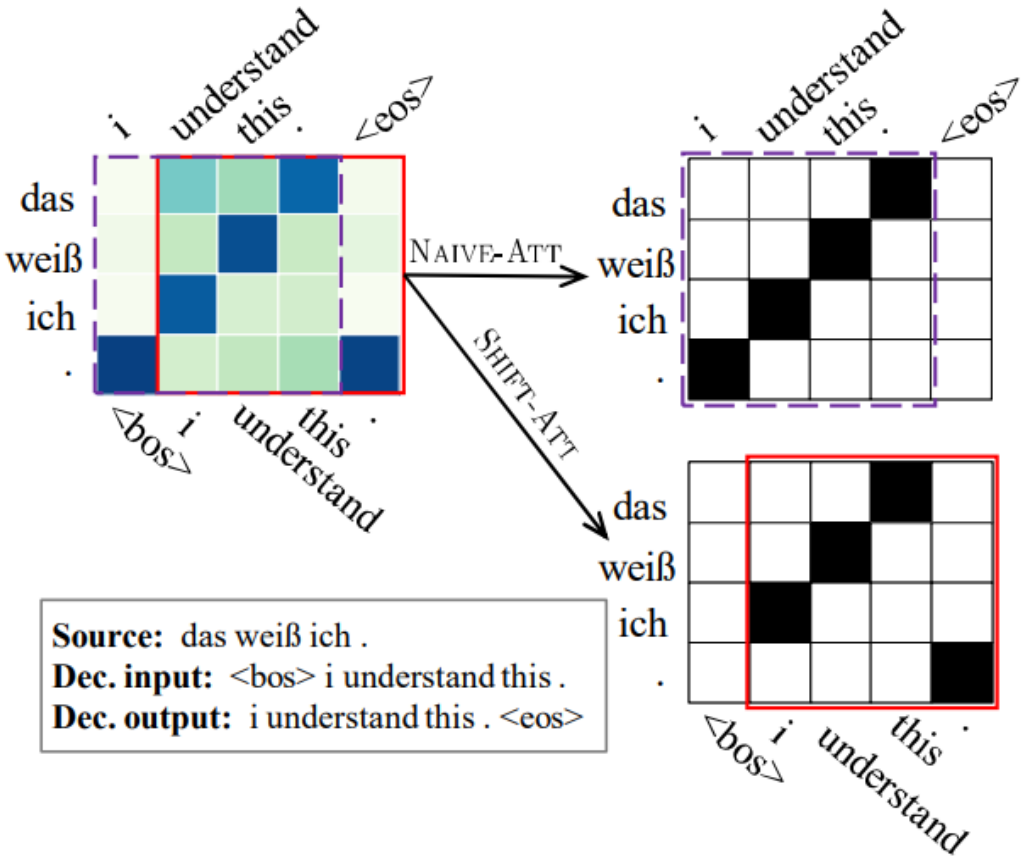


神经机器翻译的六个问题之一 (Koehn and Knowles, 2017)

词语对齐

通过观察注意力矩阵发现

- 注意力矩阵可以很好地表现词语对齐，但是需要进行移位
- 做法：用解码器第i+1步而不是第i步的注意力代表目标语言词语y_i的对齐信息
- 解释：在解码器第i步，实际上是利用词语y_{i-1}作为输入预测y_i，此时解码器还未获得y_i的全部信息，而在第i+1步将y_i作为输入时才得到了y_i的全部信息



词语对齐

基于已训练的Transformer翻译模型

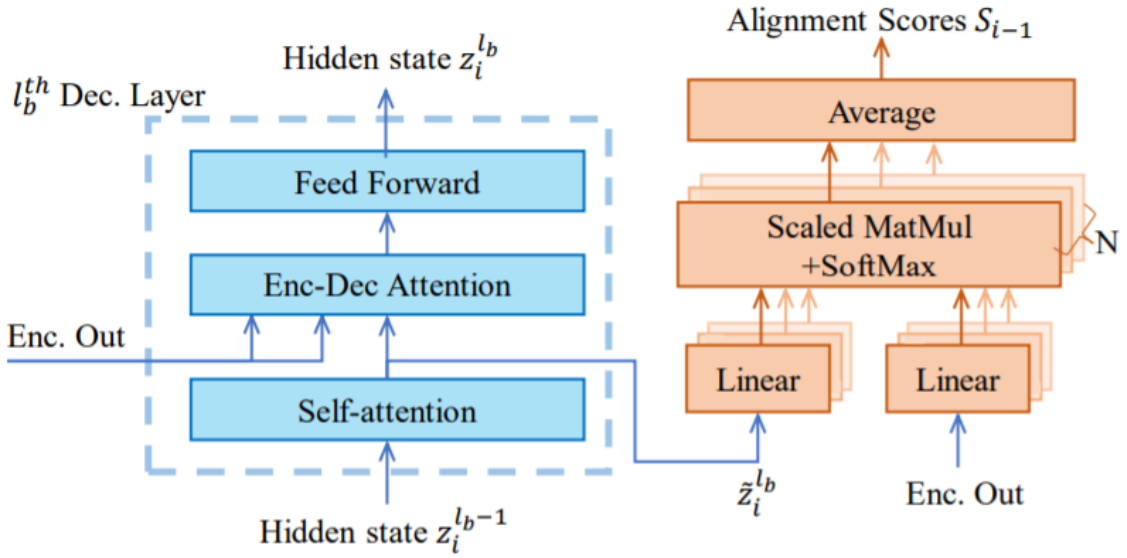
■ SHIFT-ATT方法

- 无需对模型做任何修改和重新训练
- 合并双向翻译模型得到的词语对齐
- 从解码器的多层注意力中选择最佳的一层

$$l_{\mathrm{b},\mathbf{x}\to\mathbf{y}}, l_{\mathrm{b},\mathbf{y}\to\mathbf{x}} = \underset{i,j}{\operatorname{argmin}} \operatorname{AER}(\boldsymbol{A}^{i}_{\mathbf{x}\to\mathbf{y}}, \boldsymbol{A}^{j}_{\mathbf{y}\to\mathbf{x}})$$

■ SHIFT-AET方法

- 系统中加入一个独立的对齐模块，可进一步提升词语对齐的质量
- 利用SHIFT-ATT得到的词语对齐训练
- 翻译模型保持不变



$$\mathcal{L}_a = -\frac{1}{|\mathbf{y}|}\sum_{i=1}^{|\mathbf{y}|}\sum_{j=1}^{|\mathbf{x}|}(\hat{A}^{p}_{i,j} \odot \log S_{i,j})$$

Yun Chen, Yang Liu, Guanhua Chen, Xin Jiang, Qun Liu. Accurate Word Alignment Induction from Neural Machine Translation. In Proceedings of EMNLP 2020.

词语对齐

- 实验结果

Method	Inter.	Fullc	de-en			fr-en			ro-en		
			de→en	en→de	bidir	fr→en	en→fr	bidir	ro→en	en→ro	bidir
Statistical Methods											
FAST-ALIGN (Dyer et al., 2013)	-	Y	28.5	30.4	25.7	16.3	17.1	12.1	33.6	36.8	31.8
GIZA++ (Brown et al., 1993)	-	Y	18.8	19.6	17.8	7.1	7.2	6.1	27.4	28.7	26.0
Neural Methods											
NAIVE-ATT (Garg et al., 2019)	Y	N	33.3	36.5	28.1	27.5	23.6	16.0	33.6	35.1	30.9
NAIVE-ATT-LA (Garg et al., 2019)	Y	N	40.9	50.8	39.8	32.4	29.8	21.2	37.5	35.5	32.7
SHIFT-ATT-LA	Y	N	54.7	46.2	45.5	60.5	46.9	55.1	66.1	60.4	65.3
SMOOTHGRAD (Li et al., 2016)	Y	N	36.4	45.8	30.3	25.5	27.0	15.6	41.3	39.9	33.7
SD-SMOOTHGRAD (Ding et al., 2019)	Y	N	36.4	43.0	29.0	25.9	29.7	15.3	41.2	41.4	32.7
PD (Li et al., 2019)	Y	N	38.1	44.8	34.4	32.4	31.1	23.1	40.2	40.8	35.6
ADDSGD (Zenkel et al., 2019)	N	N	26.6	30.4	21.2	20.5	23.8	10.0	32.3	34.8	27.6
MTL-FULLC (Garg et al., 2019)	N	Y	-	-	20.2	-	-	7.7	-	-	26.0
Statistical + Neural Methods											
MTL-FULLC-GZ (Garg et al., 2019)	N	Y	-	-	16.0	-	-	**4.6**	-	-	23.1
Our Neural Methods											
SHIFT-ATT	Y	N	20.9	25.7	17.9	17.1	16.1	6.6	27.4	26.0	23.9
SHIFT-AET	N	N	15.8	19.2	**15.4**	9.9	10.5	4.7	22.7	23.6	**21.2**

Table 2: AER on the test set with different alignment methods. *bidir* are symmetrized alignment results. The column Inter. represents whether the method is an interpretation method that can extract alignments from a pretrained vanilla Transformer model. The column Fullc denotes whether full target sentence is used to extract alignments at test time. The lower AER, the better. We mark best symmetrized interpretation results of vanilla Transformer with underlines, and best symmetrized results among all with boldface.

篇章机器翻译

- **当前的主流机器翻译系统是句子级的**
  - 输入的篇章段落首先被切分为句子
  - 系统并行地翻译这些独立句子，速度优势

- **篇章翻译需要考虑句子间的上下文**
  - 翻译一致性、连贯性等要求，文档、字幕、对话等翻译场景
  - 相比于句子级数据，篇章双语平行数据少
  - 篇章翻译模型通常考虑短距离上下文，保证效率的情况下引入最可能有用的上下文信息，长距离依赖建模容易影响效率和质量

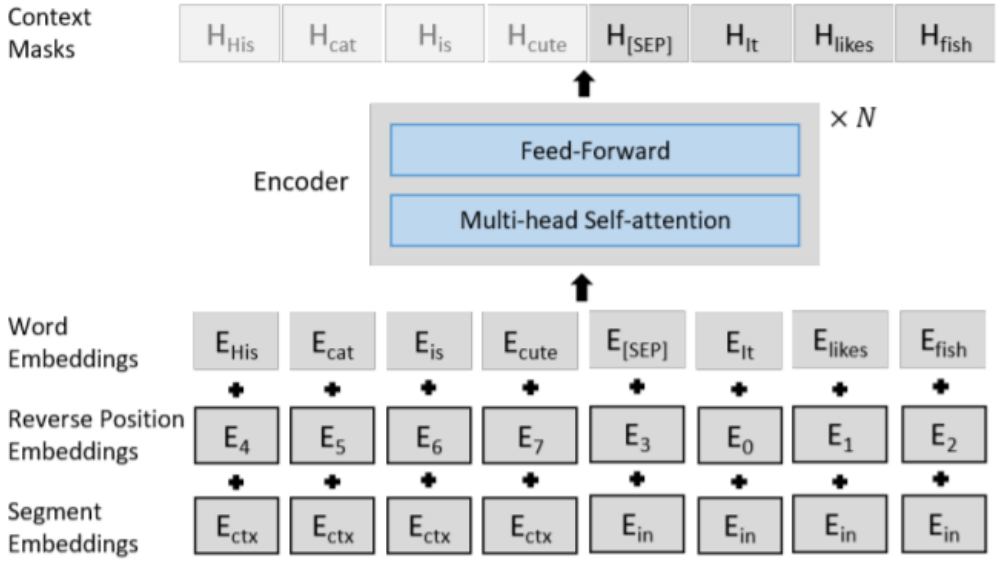
Source	与大多数欧洲人一样, 德国总理对美国总统的“美国优先”民族主义难以掩饰不屑。 ... 但她已进入第四个、也必定是最后一个总理任期。
Sent2Sent	Like most Europeans , the German chancellor has struggled to hide his disdain for the US president’s “America First” nationalism. ... But she has entered a fourth and surely last term as prime minister.

(Sun et al., 2020)

篇章机器翻译

长距离上下文的初步实验

- 预训练语言模型作为编码器初始化
- 当前语句和其前文拼接作为输入
- 句子类型编码用于区分当前语句和其前文
- 当前句的编码器输出表示输入解码器



- 预训练语言模型初始化有利于长下文的训练，基线模型训练不收敛
- 但是只使用预训练语言模型解决不了长上下文的模型质量问题，上述的方法缩短了长短上下文间的差距

Systems	Zh–En
Baseline	12.19
+Small Context	12.29 (+0.1)
+Large Context	Diverge
+BERT	13.23 (+1.04)
+Small Context	15.54 (+3.35)
+Large Context	14.54 (+2.35)
+BERT +Manipulation	-
+Small Context	15.50 (+3.31)
+Large Context	15.30 (+3.11)

Liangyou Li, Xin Jiang, Qun Liu. Pretrained Language Models for Document-Level Neural Machine Translation. arXiv:1911.03110v1

篇章机器翻译

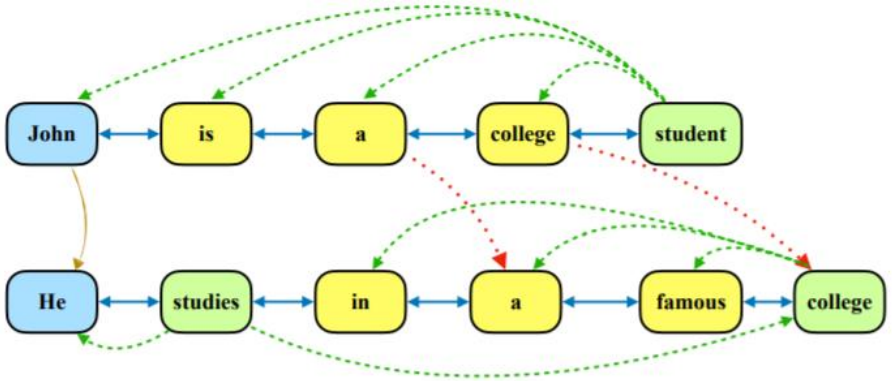
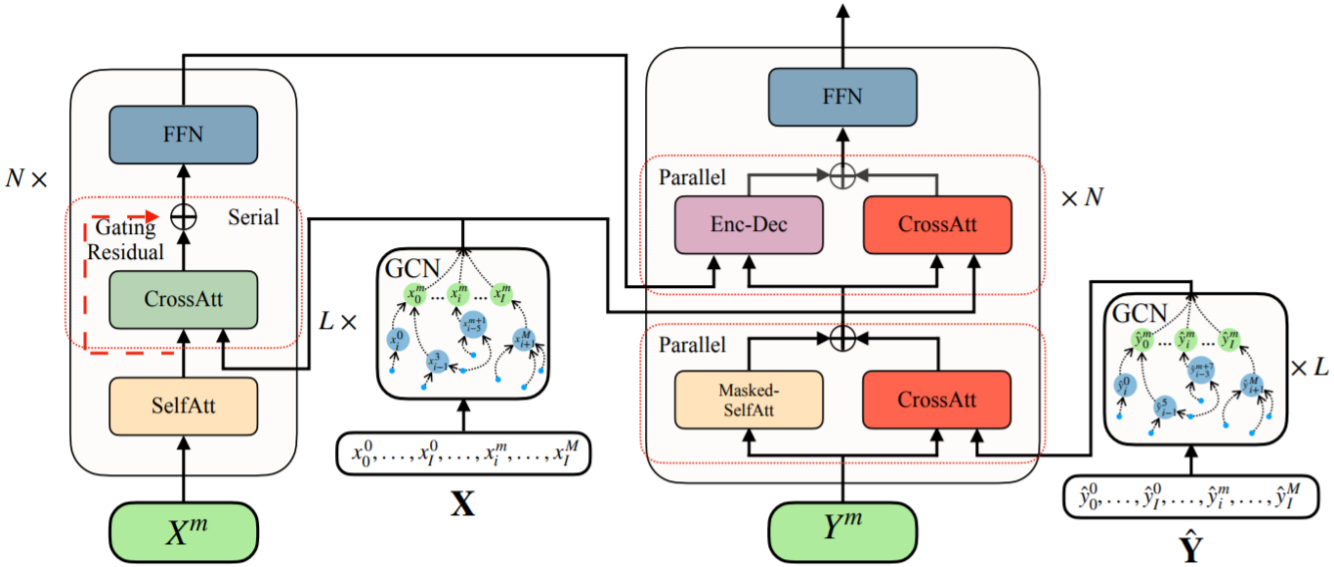


Figure 1: The diagram Illustrates the architecture of the proposed model. Solid lines in blue depict adjacency relations. Dash lines in green denote dependency relations. Lexical consistency is represented as dash-dotted lines in red. The brown line means a coreference relation. [1]

将文档表示成图结构，词作为图节点，四种关系构成边：邻居关系、依存关系、词汇衔接、指代。上下文相关词在图上的距离不等同于文档中的距离。

图卷积网络用于编码文档图，并利用关注网络将节点表示集成到翻译模型中。



篇章机器翻译

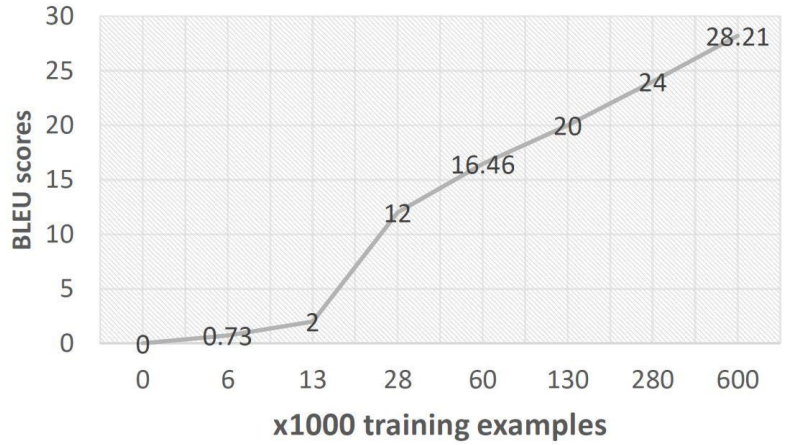
- **实验结果**

Model	En-Fr		Zh-En		En-DE		En-Ru		Para. △	Speed
	Dev	Test	Dev	Test	Dev	Test	Dev	Test		
BASE	29.56	35.77	10.92	16.7	35.91	34.89	28.44	29.05	-	24.9k
Constrained Context										
HAN	29.93	36.15	11.37	17.65	36.01	35.00	28.92	29.38	7.36 M	14.4k
CTX	30.23	36.67	11.37	17.57	36.05	35.23	28.79	29.31	22.06M	16.3k
CACHE	30.17	36.27	11.36	17.39	35.93	35.12	29.31	29.53	1.84 M	18.6k
Global Context										
MS	29.04	34.93	10.59	16.12	35.87	34.59	27.89	28.70	0.00 M	16.1k
HM-GDC	30.36	36.38	11.54	17.52	35.97	35.01	28.96	29.12	7.30 M	19.9k
SELECTIVE	30.53	36.87	11.57	17.86	36.14	35.47	29.67	29.64	8.39 M	7.7k
Our										
SRC–GRAPH	$30.84^{\Uparrow}$	$37.11^{\Uparrow}$	$11.75^{\uparrow}$	$18.31^{\Uparrow}_{\ddagger}$	36.21	$35.68^{\uparrow}$	$29.78^{\uparrow}$	29.72	22.59M	17.2k
+TGT	$31.62^{\Uparrow}_{\ddagger}$	$37.71^{\Uparrow}_{\ddagger}$	$12.01^{\Uparrow}_{\ddagger}$	$18.53^{\Uparrow}_{\ddagger}$	$36.34^{\uparrow}$	$35.94^{\Uparrow}_{\ddagger}$	$30.14^{\Uparrow}\ddagger$	$30.10^{\Uparrow}_{\ddagger}$	22.59M	15.9k

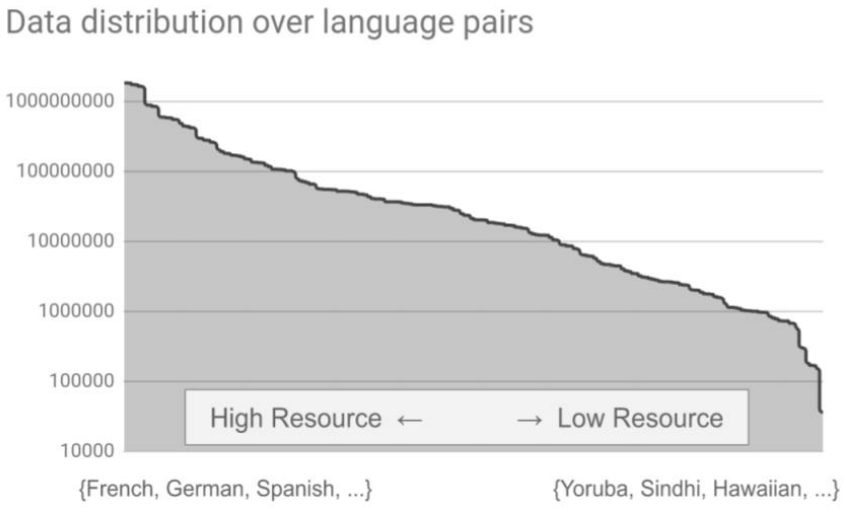
Mingzhou Xu, Liangyou Li, Derek. F. Wong, Qun Liu, Lidia S. Chao. Document Graph for Neural Machine Translation. arXiv:2012.03477

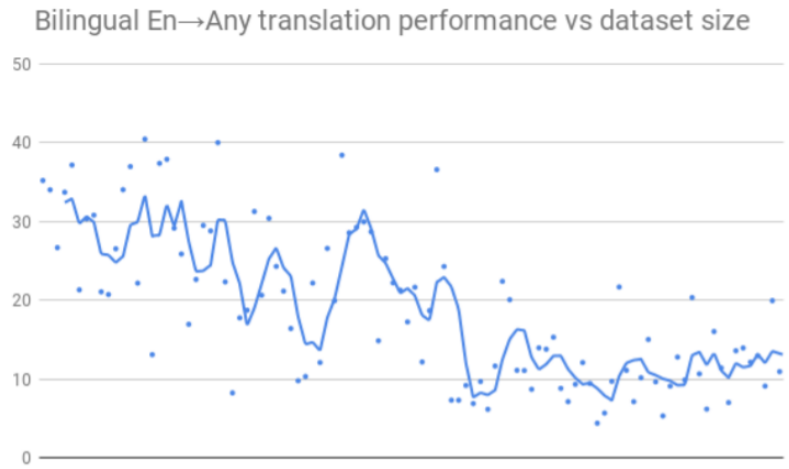
低资源机器翻译

- 神经机器翻译的质量与数据量相关
- 不同语言(对)数据量存在差异
- 在高资源语言对上取得了很多进步和应用，低资源翻译仍然存在挑战



(Gu et al., 2018)





(Arivazhagan et al., 2019)

低资源机器翻译

利用额外数据

- 单语数据，如回译、预训练语言模型等
- 其他语言的单/双语数据：迁移、多语言等

	High-resource language		Low-resource language	
	monolingual	parallel	monolingual	parallel
no transfer				✔
(Zoph et al., 2016)		✔		✔
(Kim et al., 2019)		✔	✔	✔
BERT2RND			✔	✔
BERT2BERT			✔	✔
(Kocmi and Bojar, 2018)*		✔		✔
BBERT2BBERT*			✔	✔
BBERT transfer*	✔	✔	✔	✔
dual transfer (ours)	✔	✔	✔	✔

*共享词表

低资源机器翻译

- 从高资源A→B 迁移到 P→Q的通用流程

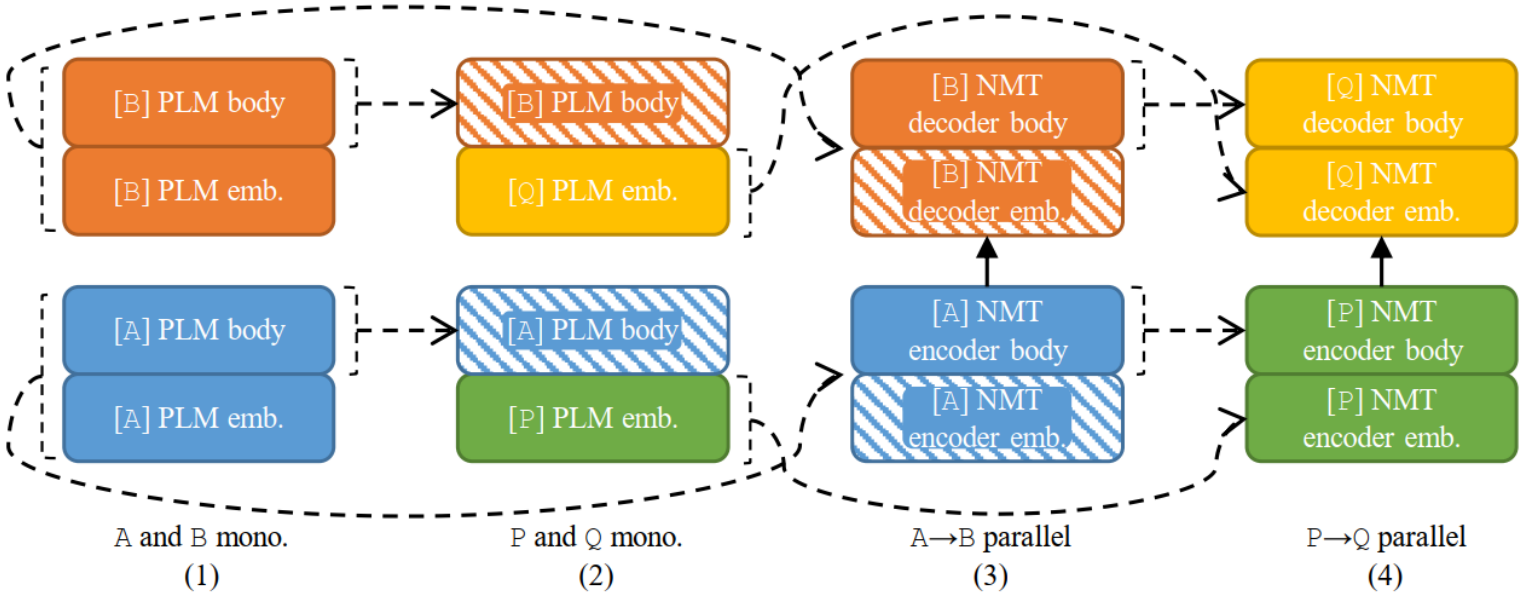


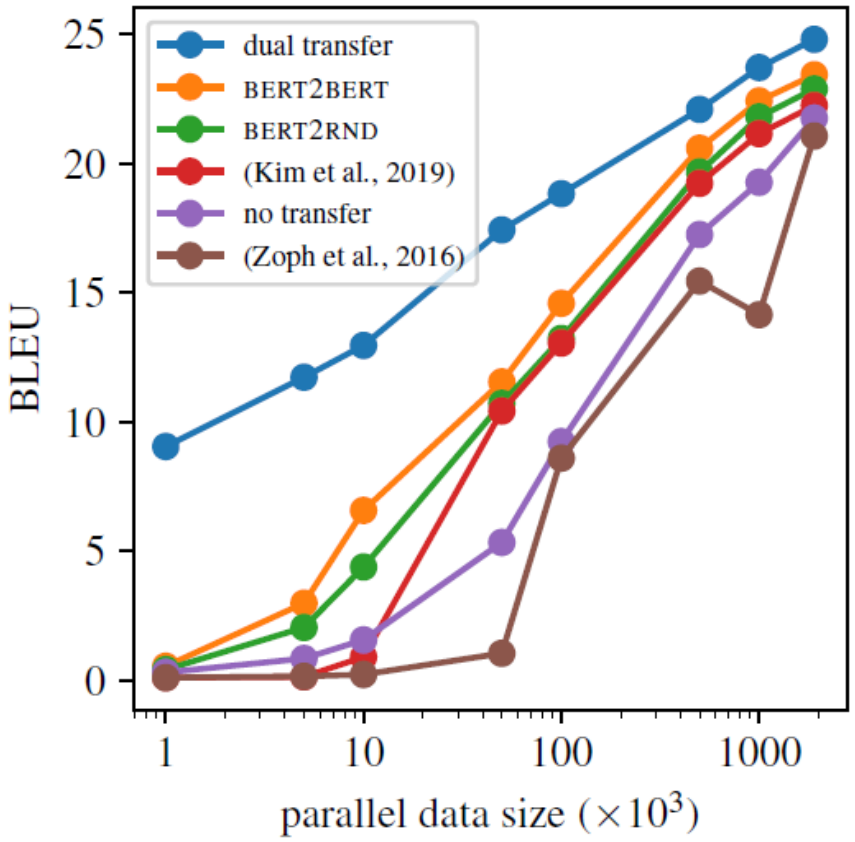
Figure 1: Dual transfer from PLM and high-resource A→B NMT to low-resource P→Q NMT. Dashed lines represent initialization. Parameters in striped blocks are frozen in the corresponding step, while other parameters are trainable. Different colors represent different languages. Data used in each step is also listed.

Meng Zhang, Liangyou Li, Qun Liu. Two Parents, One Child: Dual Transfer for Low-Resource Neural Machine Translation. In ACL 2021 Findings.

低资源机器翻译

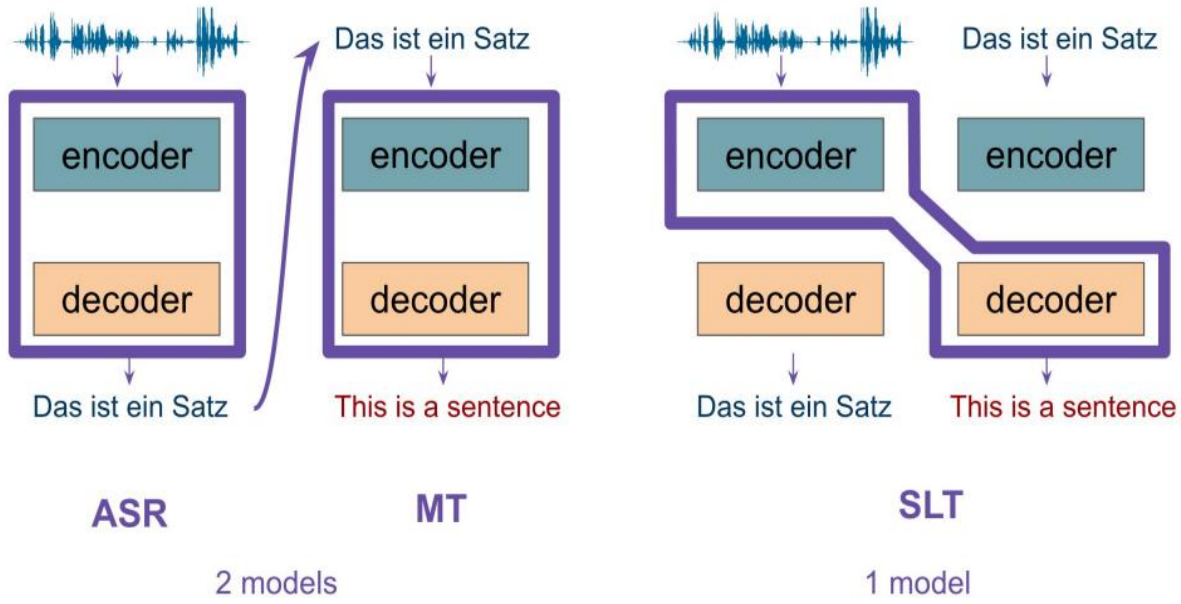
■ 实验结果

	et-en BLEU
no transfer	21.76
(Zoph et al., 2016)	21.07
(Kim et al., 2019)	22.25
BERT2RND	22.89
BERT2BERT	23.44
(Kocmi and Bojar, 2018)*	23.58
BBERT2BBERT*	23.90
BBERT transfer*	24.08
dual transfer	**24.81**



实时语音翻译

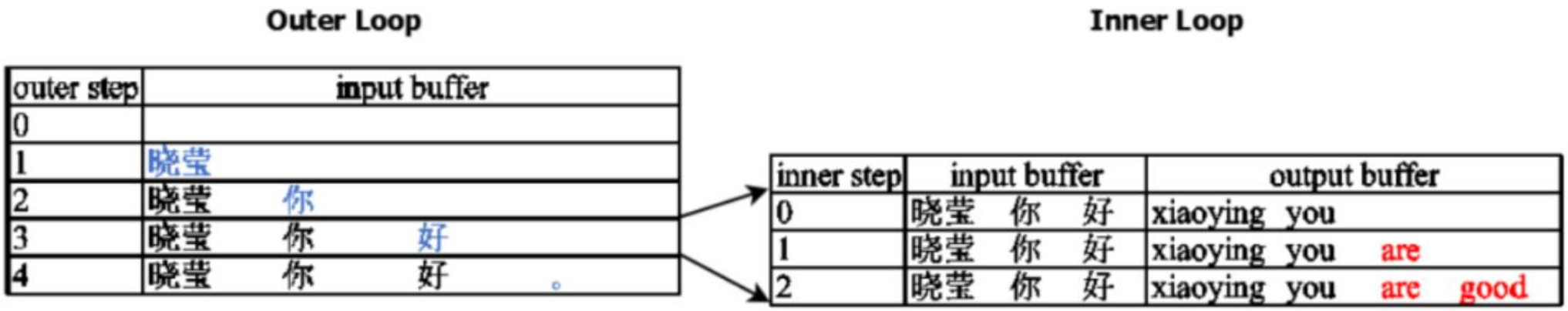
- **级联模型**
  - 优点：每个模块都有大量的数据，更加容易训练，复用现有服务，不需要单独部署语音翻译模型
  - 缺点：错误传递、延迟高，语音信息丢失、不同模块训练数据领域不匹配
- **端到端模型**
  - 优点：延迟低、避免错误传递、利用全部语音信息
  - 缺点：语料稀缺、要直接处理语音和翻译之间的模态差异，任务复杂建模困难、不易干预



级联模型和端到端模型 Link

实时语音翻译

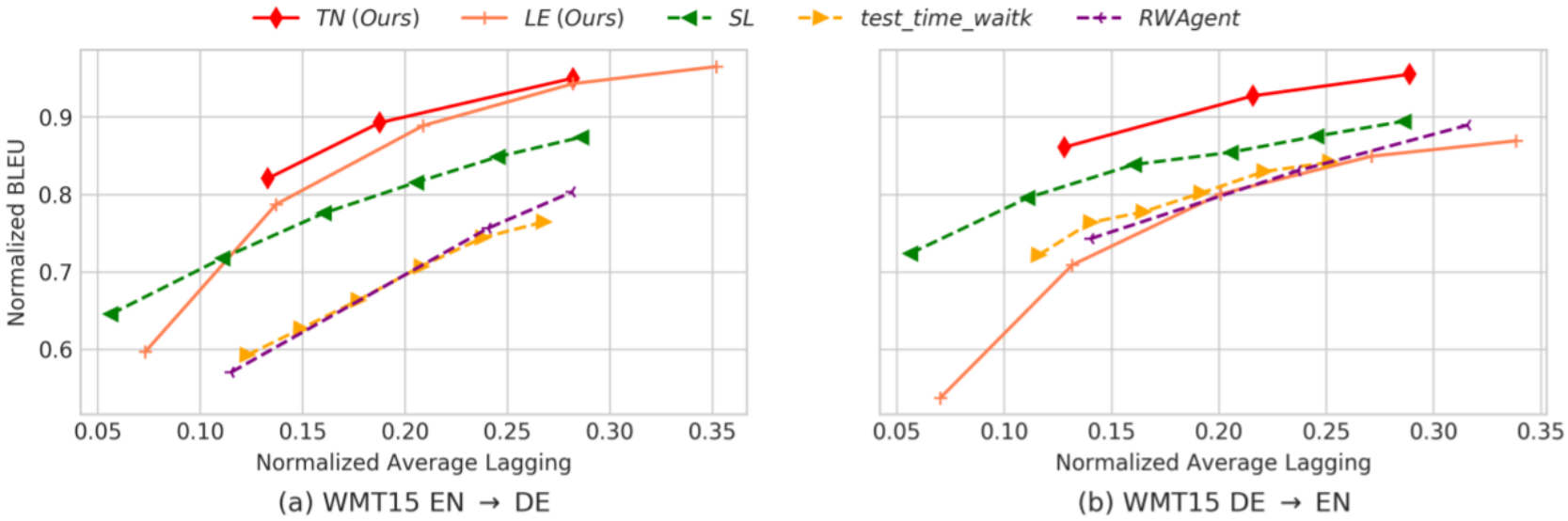
- **基于级联模型的实时语音翻译**
  - ASR模型、翻译模型、输入缓存、输出缓存
  - 外循环：读取ASR的文本输出流到输入缓存
  - 内循环：翻译输入缓存中的内容并输出到输出缓存中



Yun Chen, Liangyou Li, Xin Jiang, Xiao Chen, Qun Liu, A General Framework for Adaptation of Neural Machine Translation to Simultaneous Translation. In Proceedings of AACL-IJCNLP 2020.

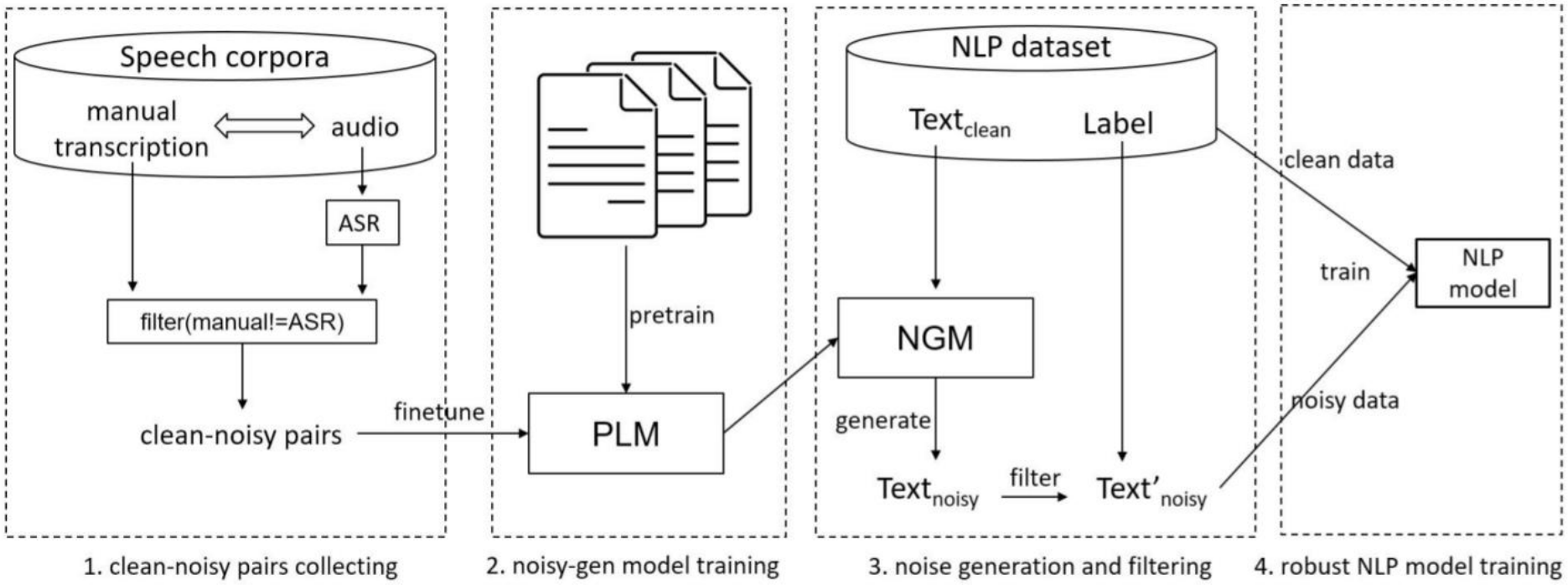
实时语音翻译

- **流式输入的连续翻译：**重新编码+强制解码
- **内循环的翻译长度，即翻译停止条件**
  - 翻译输出句尾结束符，或者达到预定义的长度
  - 一个额外可训练的控制器输出停止符



实时语音翻译

- 对语音识别错误鲁棒的机器翻译



Tong Cui, Jinghui Xiao, Liangyou Li, Xin Jiang, Qun Liu. An Approach to Improve Robustness of NLP Systems against ASR Errors. arXiv:2103.13610v1

实时语音翻译

■ 实验结果

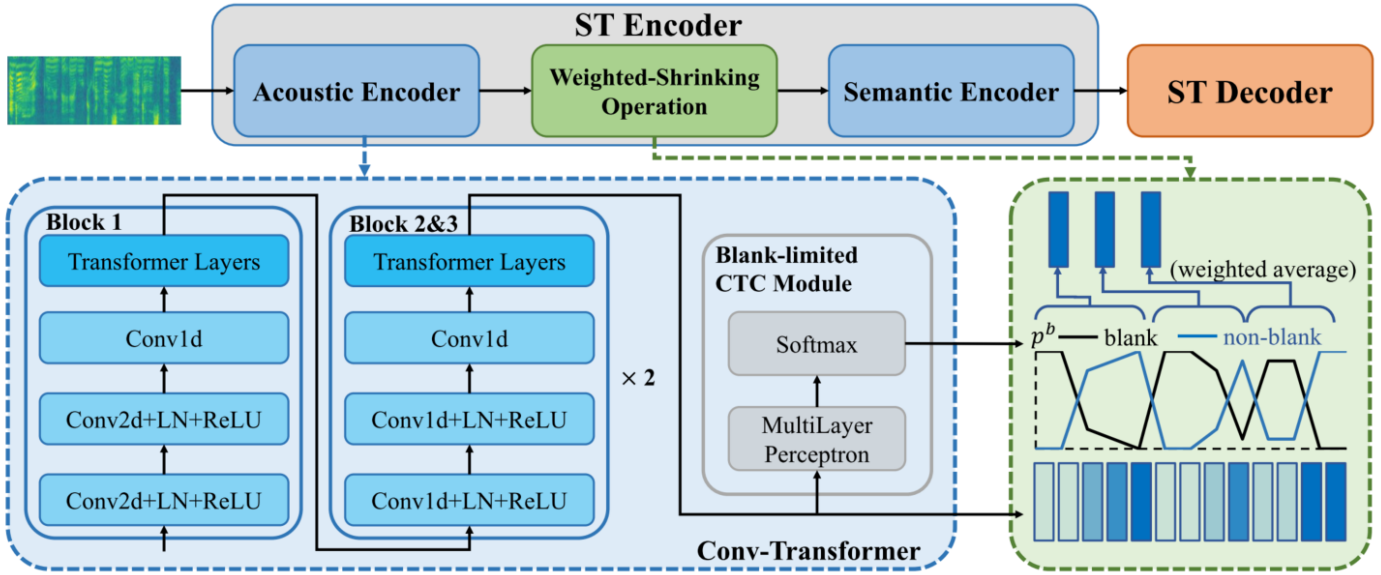
	En-Zh					En-De			
	CoVoST-v2		MSLT			CoVoST-v2		MSLT	
	Manual	ASR	Manual	ASR	ASR-other	Manual	ASR	Manual	ASR
Clean	**50.5**	28.2	**45.3**	28.2	34.1	34.0	17.2	26.2	13.9
+RS	50.5	28.2	45.3	26.7	31.6	**34.1**	17.0	25.9	13.7
+SS	50.3	29.7	44.4	28.9	35.8	34.0	18.0	25.6	14.3
+Ours	50.2	**30.8**	44.4	**32.0**	**38.6**	34.0	**19.0**	**26.6**	**20.2**

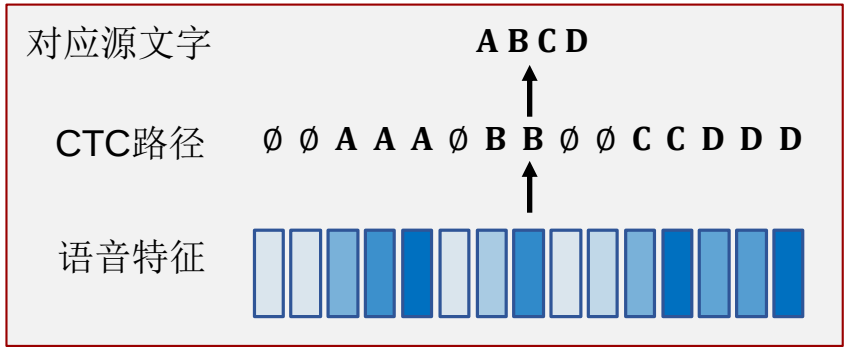
Table 6: BLEU scores for the results of MT models on manual transcriptions and ASR results of test sets. "Clean" refers to models trained on clean training data. "+SS", "+RS", "+Ours" refer to models trained on augmented data. "ASR-other" refers to the ASR results provided by MSLT dataset.

实时语音翻译

- **端到端模型**
  - 使用单向编码器用于实时输入
  - 交替使用卷积层和自关注层实现逐渐下采样
  - 基于CTC预测结果进行加权收缩

$$\boldsymbol{h}_{t'} = \sum_{t \in seg\, t'} \boldsymbol{h}_t \frac{\exp(\mu(1-p_t^b))}{\sum_{s \in seg_{t'}} \exp(\mu(1-p_s^b))}$$





Xingshan Zeng, Liangyou Li, Qun Liu. RealTranS: End-to-End Simultaneous Speech Translation with Convolutional Weighted-Shrinking Transformer. In ACL 2021 Findings

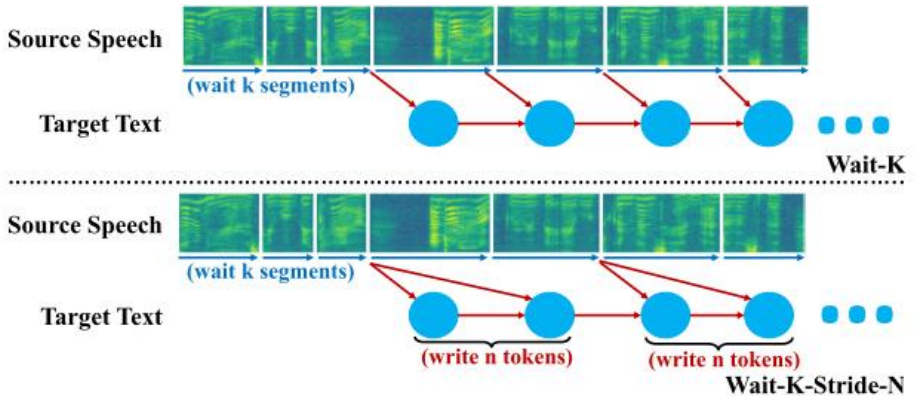
实时语音翻译

■ 端到端模型

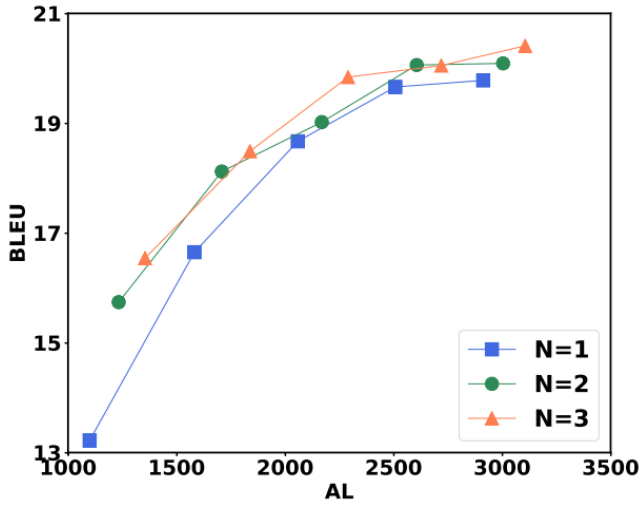
› 观察发现在单向编码时CTC倾向于预测Blank，引入Blank惩罚因子提高准确率

$$\mathcal{L}'_{CTC} = \mathcal{L}_{CTC} + \lambda \sum_{\boldsymbol{x}\in\mathcal{D}} \sum_{\pi_t \in \boldsymbol{\pi}(\boldsymbol{x})} p(\pi_t = \phi|\boldsymbol{x})$$

› Wait-K-Stride-N实时控制策略在输出N个词时可以进行局部的柱搜索

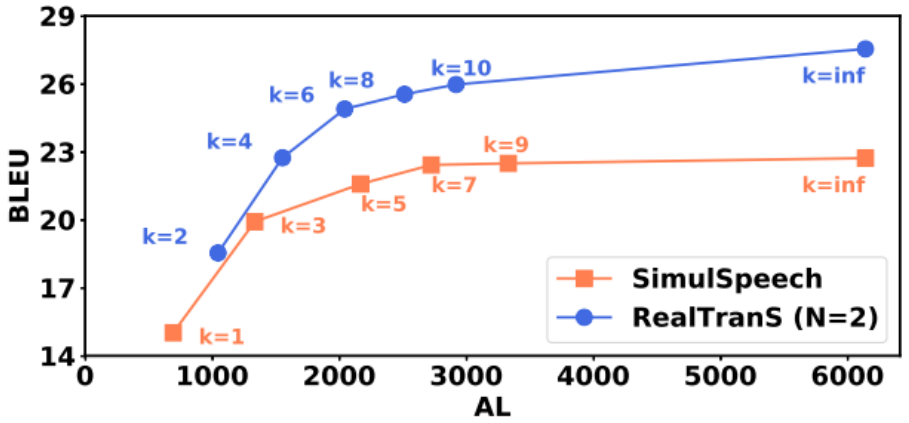


Model	Diff≤2	Diff≤4	Diff≤6	BLEU
Full Model	**52%**	**74%**	**85%**	**27.40**
- CTC PT	48%	70%	82%	26.63
- BP	35%	53%	67%	25.76
+ Bi-Enc	42%	59%	72%	26.58



实时语音翻译

■ 实验结果



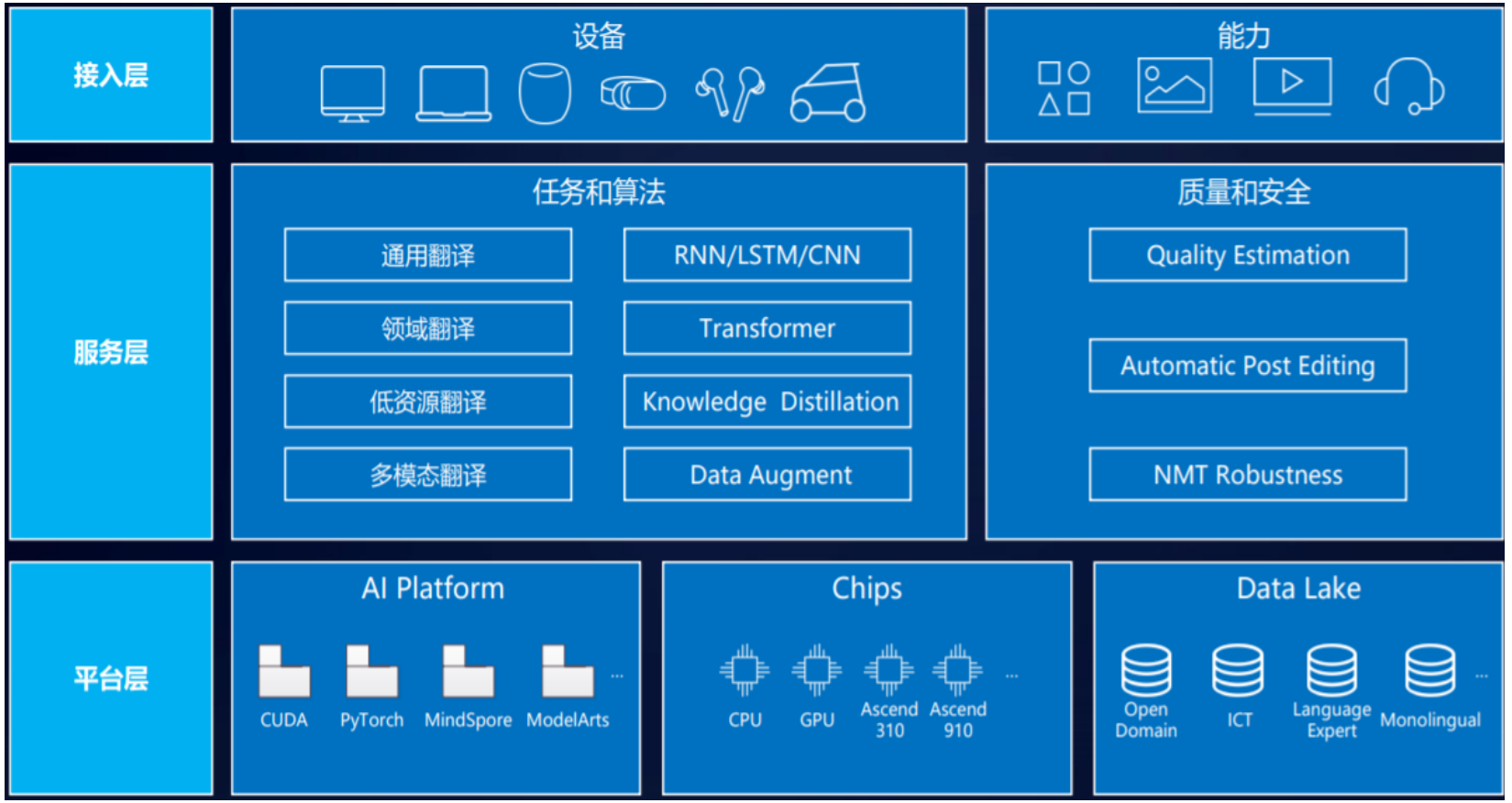
Model	K=2	K=4	K=6	K=8	K=10	K=inf
Cascaded	14.92	19.22	22.11	23.33	24.47	26.79
RealTranS	**18.45**	**22.65**	**24.79**	**25.41**	**25.82**	**27.40**

Dataset	Method	BLEU
En-Fr	Transformer+KD (Liu et al., 2019)	17.02
	TCEN-LSTM (Wang et al., 2020b)	17.05
	Curriculum PT (Wang et al., 2020c)	17.66
	LUT (Dong et al., 2020b)	17.75
	STAST (Liu et al., 2020)	17.81
	COSTT (Dong et al., 2020a)	17.83
	Transformer+AFS (Zhang et al., 2020)	18.56*
	RealTranS (ours)	18.97
		18.30*
En-De	Transformer+MAM (Chen et al., 2020)	21.87*
	Transformer+ML (Indurthi et al., 2019)	22.11*
	Transformer+AFS (Zhang et al., 2020)	22.38*
	Fairseq S2T (Wang et al., 2020a)	22.70*
	Espnet ST (Inaguma et al., 2020)	22.91*
	STAST (Liu et al., 2020)	23.06
	RealTranS (ours)	23.53
		22.99*

目录

华为翻译

- 三大应用场景
  - 手机HMS+端侧AI
  - 华为云
  - 内部IT
- 已上线30+语种
- 文本、图片、语音等
- 跨部门合作，软硬件协同优化
  - 数据增强
  - 模型设计与训练
  - 模型压缩与加速
  - 推理框架优化

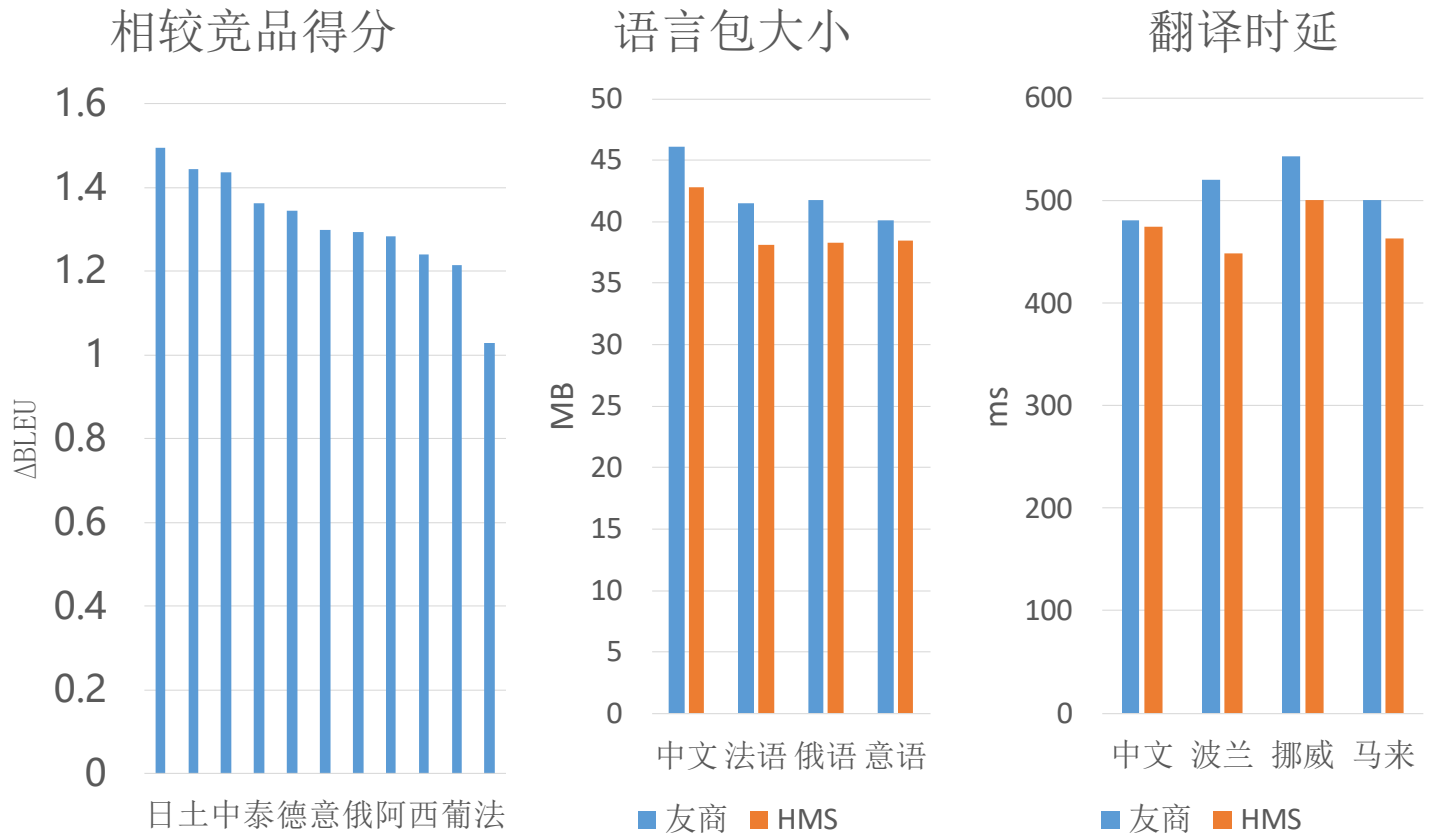


翻译竞争力水平

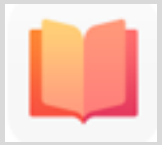
云测翻译人工评估

语种	HMS翻译相较竞品得分
英-中	127.34%
英-俄	111.93%
英-葡	109.43%
英-西	106.01%
英-瑞	104.88%
英-阿	102.99%
英-法	102.28%

端侧翻译：速度相当但质量和模型大小优于竞品



案例：阅读 & 翻译

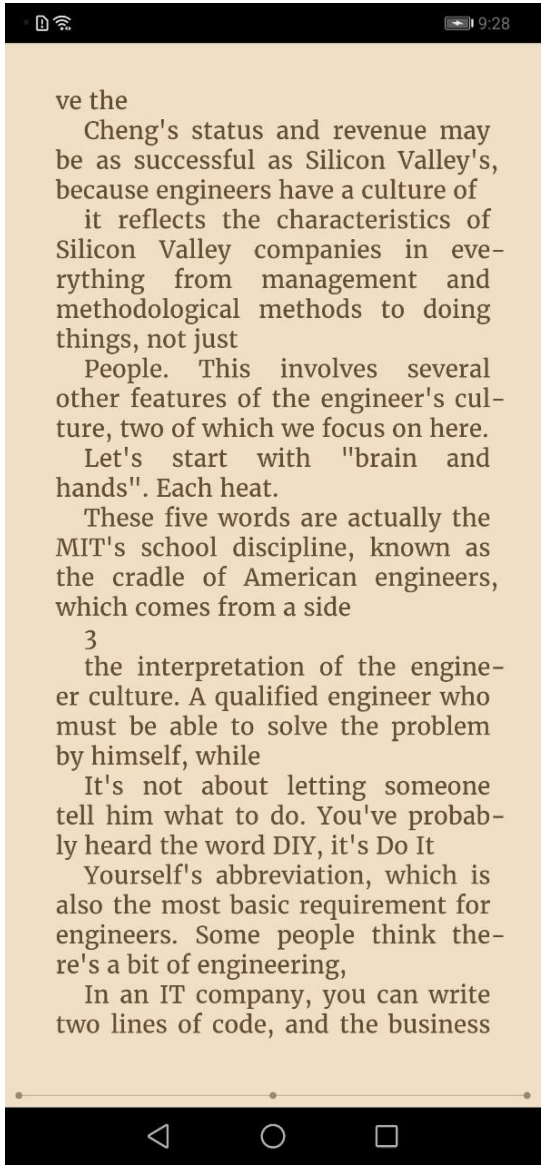


阅读APP

HMS ML kit

机器翻译

新闻阅读类、浏览器类、阅读器类APP集成翻译功能，为用户提供多语言翻译，助力APP开发者全球化。

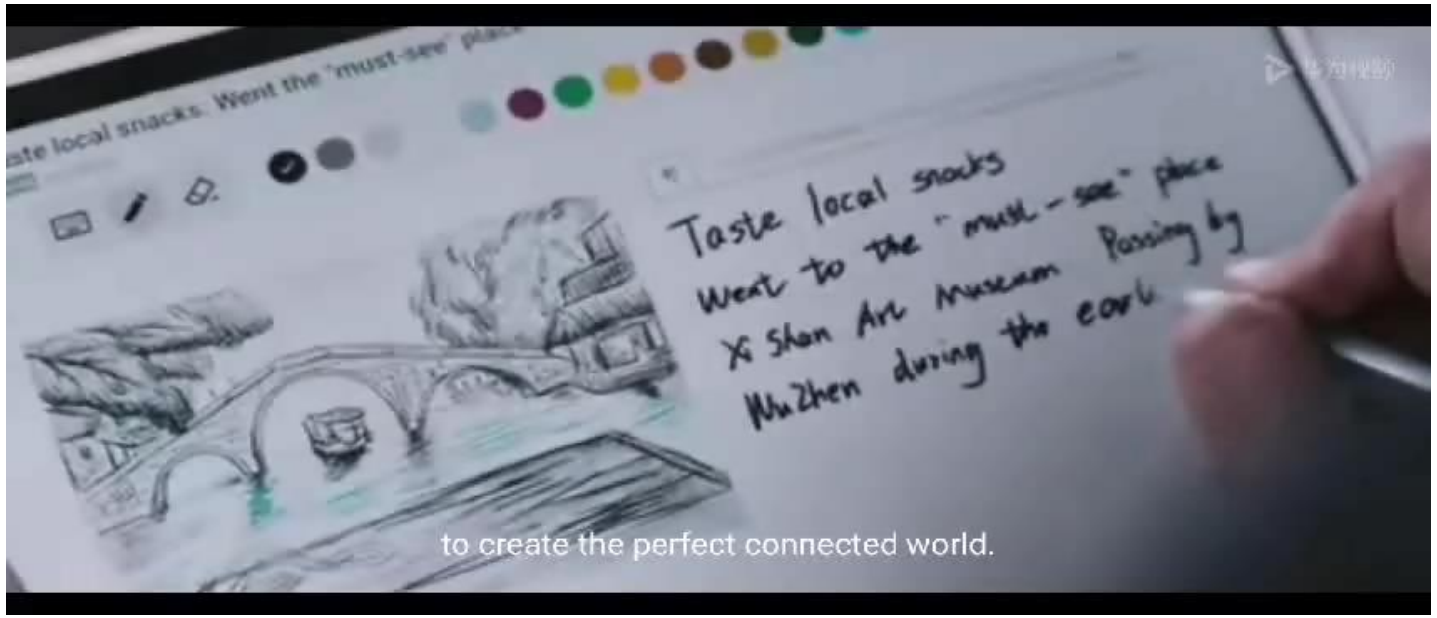


案例：字幕 & 翻译

华为视频

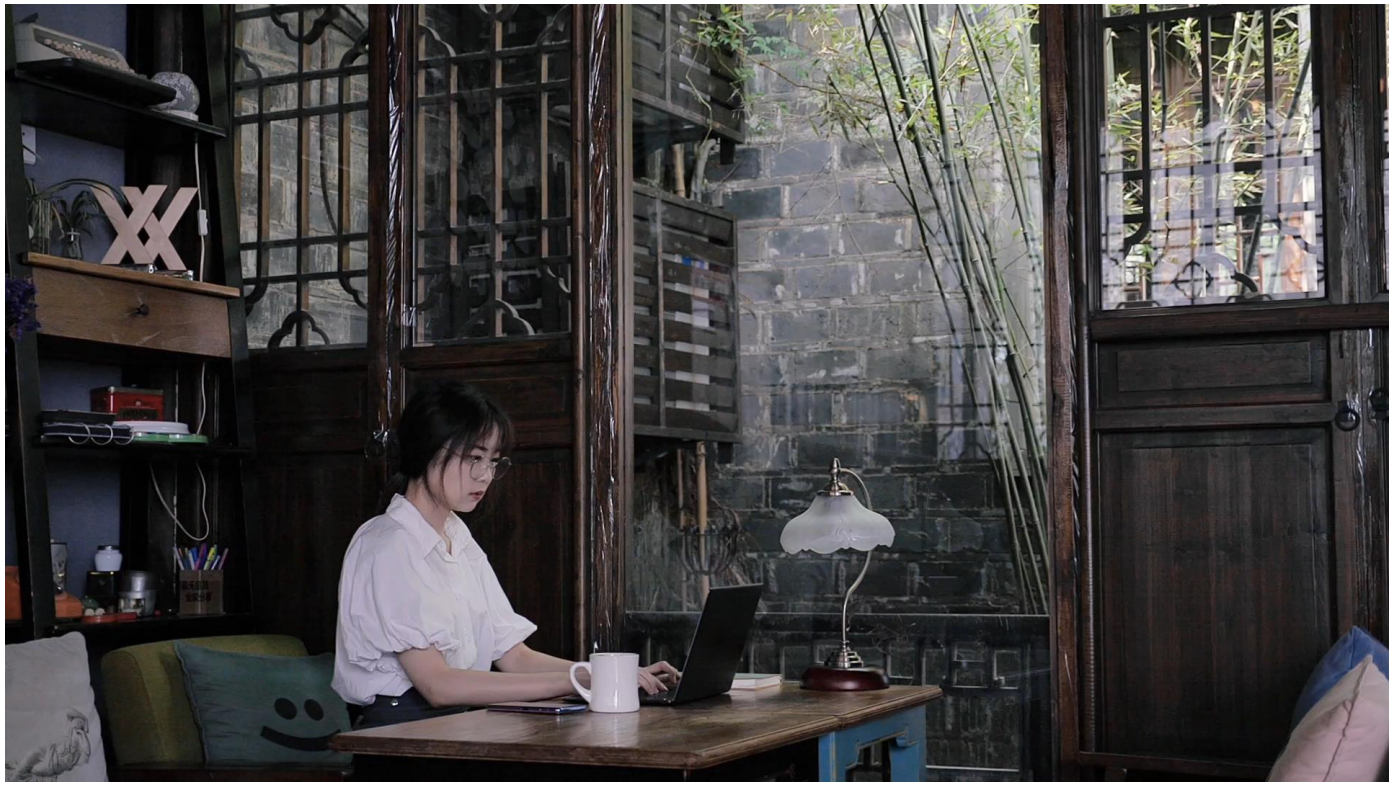
HMS ML kit

机器翻译

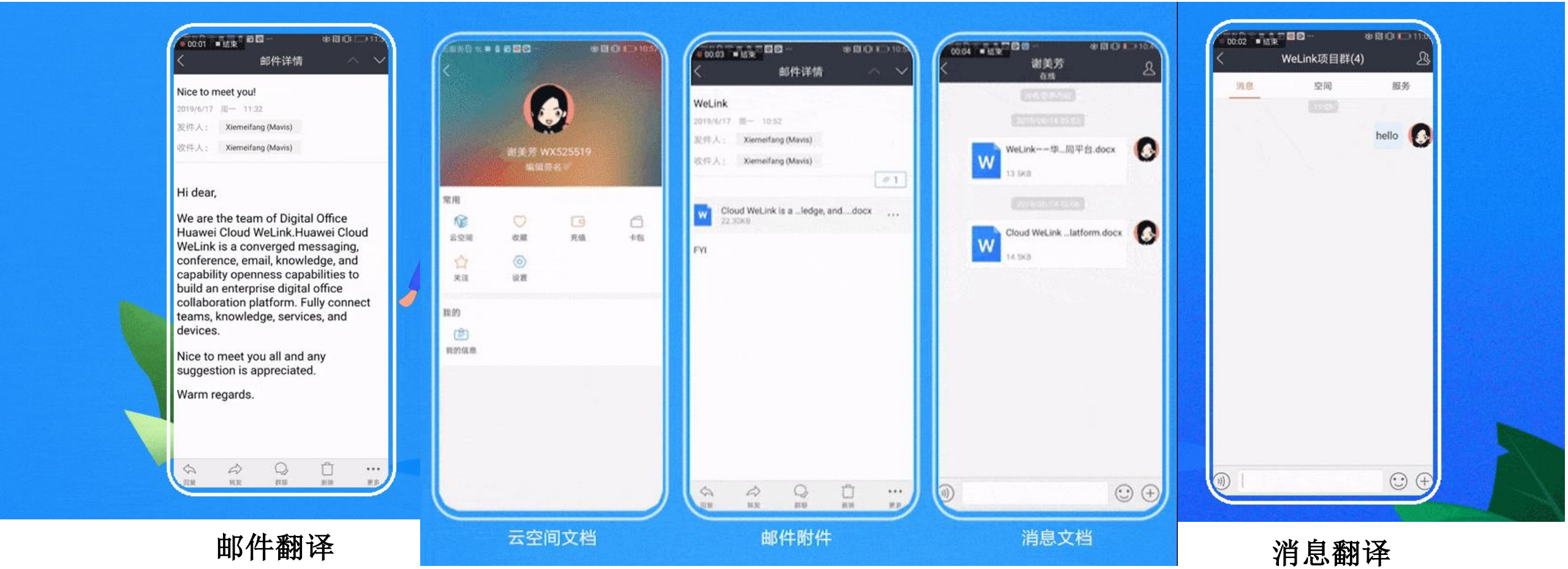


案例：通话 & 翻译

- 离线语音识别+离线翻译+离线语音合成打造端到端的语音翻译体验



案例：办公 & 翻译



邮件翻译

云空间文档

邮件附件

消息文档

消息翻译

文档翻译

目录

总结与展望

- **介绍了诺亚方舟实验室和机器翻译的近期研究成果**
  - 诺亚方舟实验室致力于人工智能的算法创新和应用
  - 机器翻译团队负责机器翻译方向的技术研发
  - 介绍了近期的一些研究：词对齐、篇章翻译、低资源翻译、实时语音翻译等
  - 与华为其他的业务和产品部门合作将机器翻译应用在不同场景
- **机器翻译应用仍然面临挑战**
  - 扩展到更多语言面临数据稀缺问题
  - 不同设备在内存和算力方面的差异
  - 不同场景的要求，领域、模态等
  - ……

欢迎大家加入诺亚方舟实验室！

Thank you

www.huawei.com