DataFunSummit

自然语言处理

在线峰会

机器翻译与同传论坛

2021.07.10（周六） 09：00~17：30

2021

Baidu 百度 | DataFunSummit

机器同传技术及应用

张睿卿　百度资深研发工程师

目录

CONTENTS

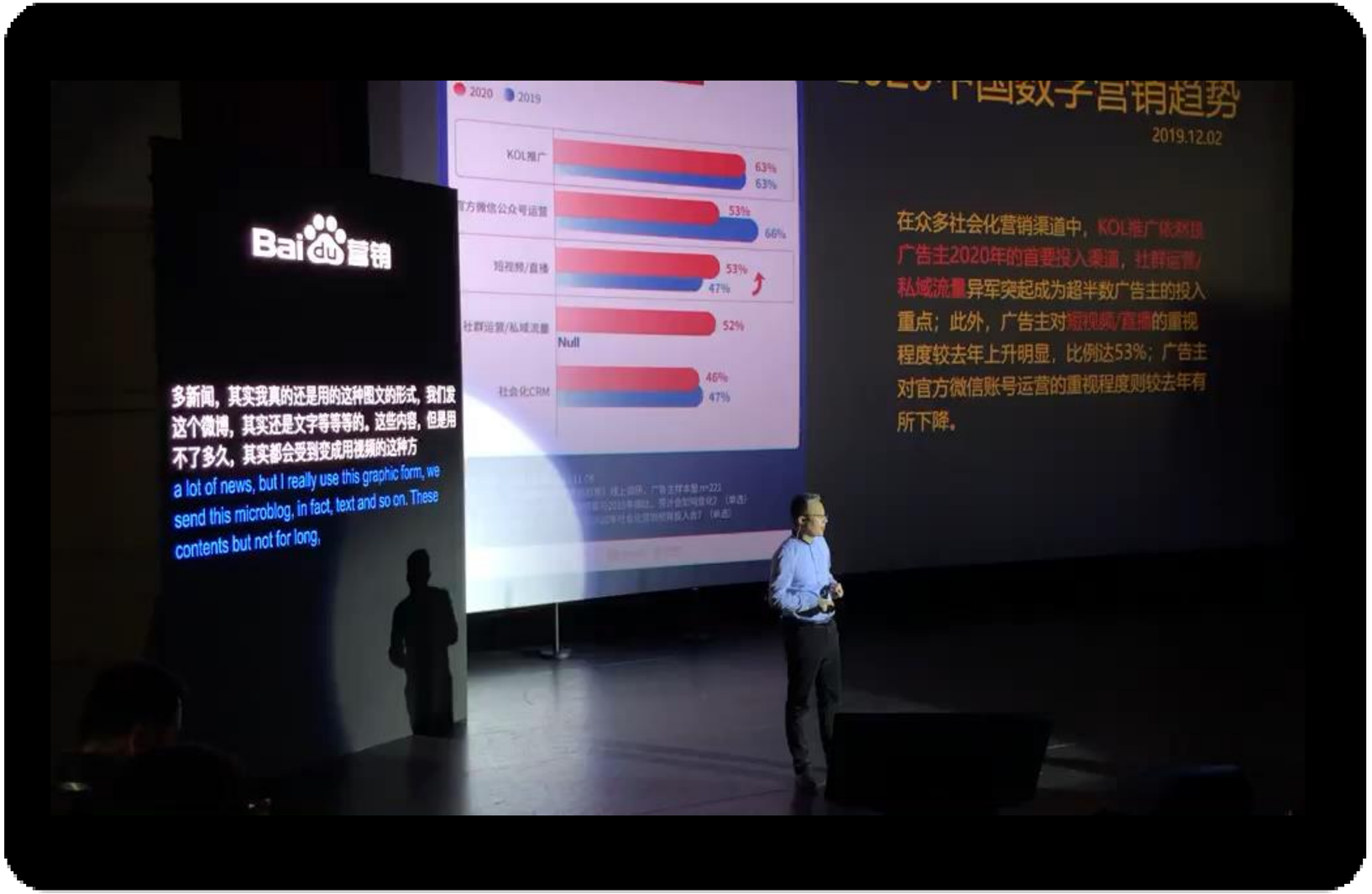
01

从机器翻译到机器同传

什么是机器同传

为什么要做机器同传

什么是机器同传


Bai du 营销
多新闻，其实我真的还是用的这种图文的形式，我们发这个微博，其实还是文字等等等的。这些内容，但是用不了多久，其实都会受到变成用视频的这种方
a lot of news, but I really use this graphic form, we send this microblog, in fact, text and so on. These contents but not for long,
2020 2019
KOL推广 63% 63%
官方微信公众号运营 53% 66%
短视频/直播 53% 47%
社群运营/私域流量 52%
Null
社会化CRM 46% 47%
2019.12.02
在众多社会化营销渠道中，KOL推广依然是广告主2020年的首要投入渠道，社群运营/私域流量异军突起成为超半数广告主的投入重点；此外，广告主对短视频/直播的重视程度较去年上升明显，比例达53%；广告主对官方微信账号运营的重视程度则较去年有所下降。

为什么要做机器同传

- 信息传递效率高
- 任务难度大： 15-20分钟需要休息
  - 源语言：监听、记忆、理解、
  - 目标语言：组织、修正、表达

https://aiic.net/directories/interpreters/lang/1

02

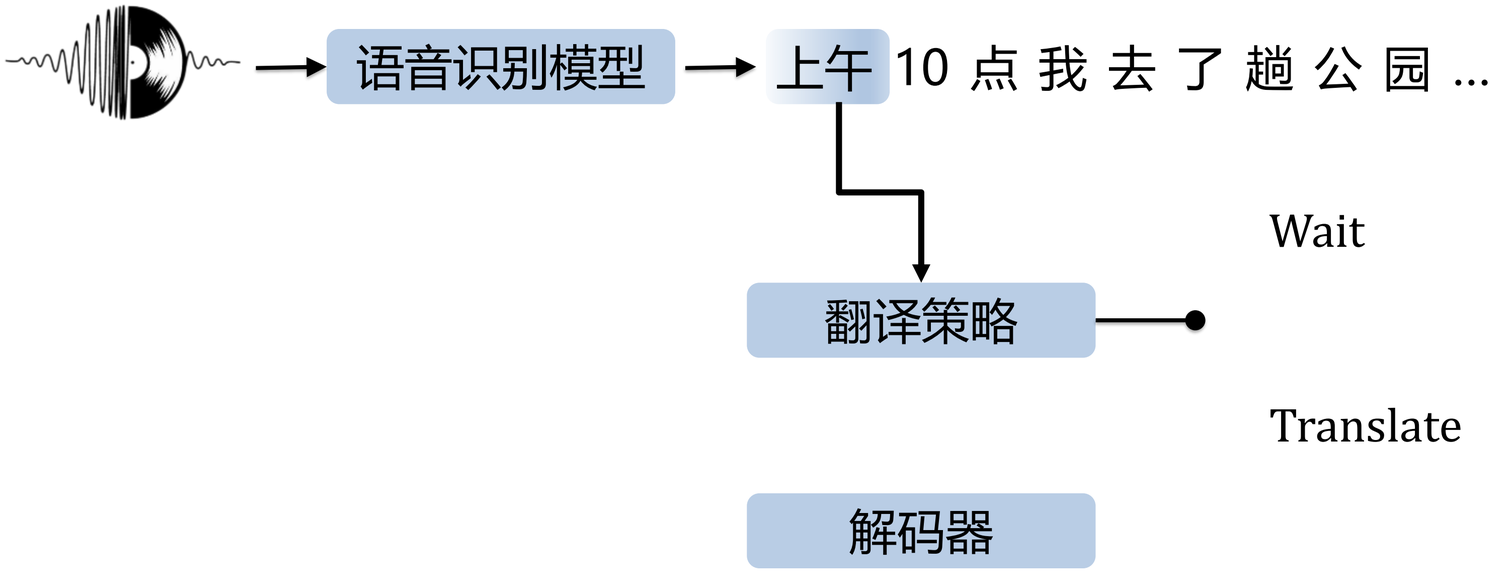
同传关键问题与解决方案

同传架构

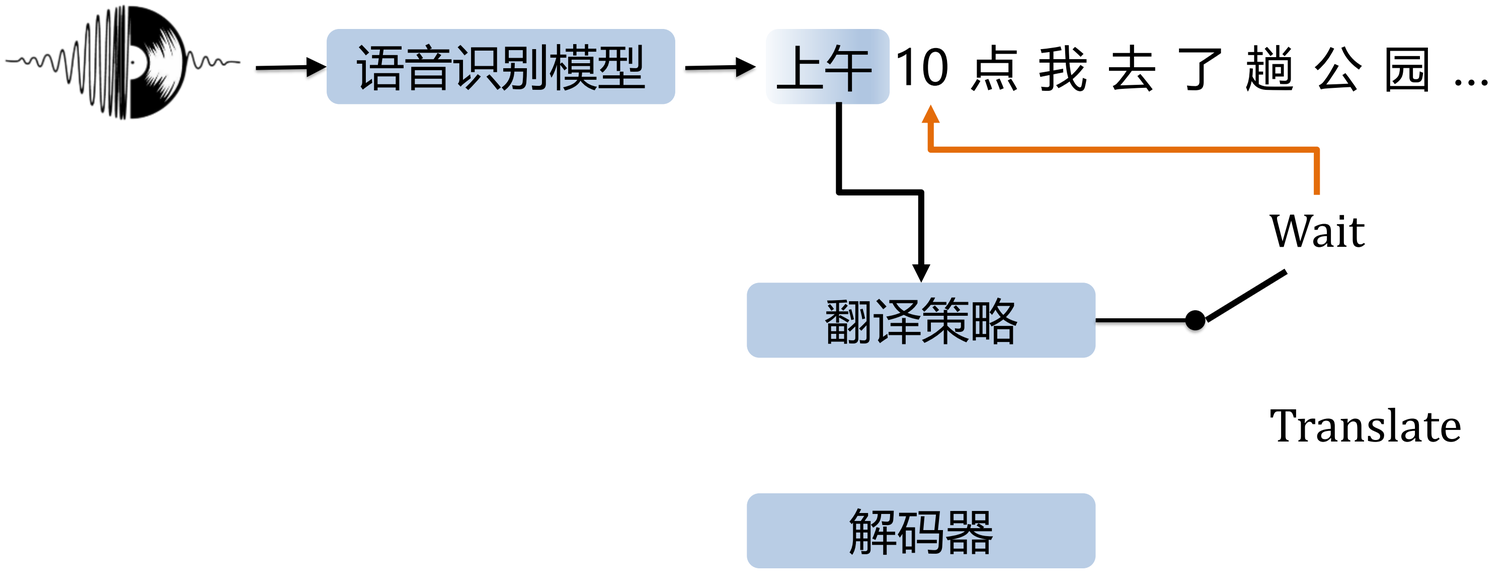
机器同传的几个挑战

解决方案

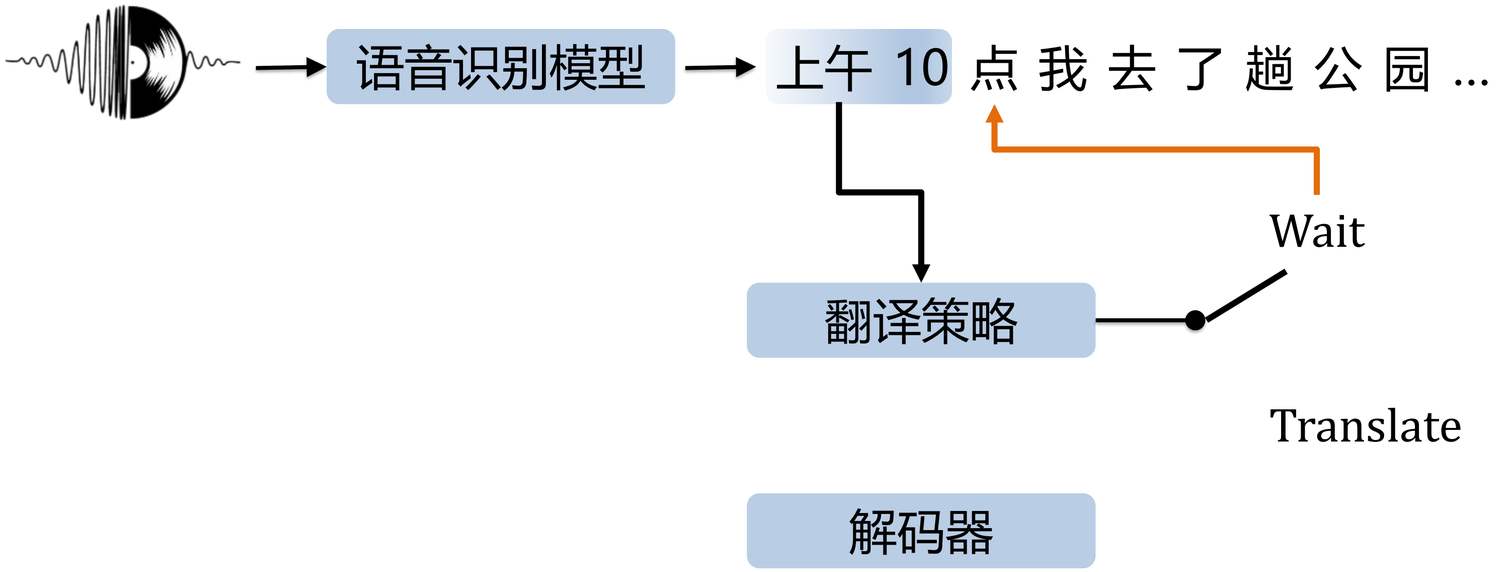
同传系统



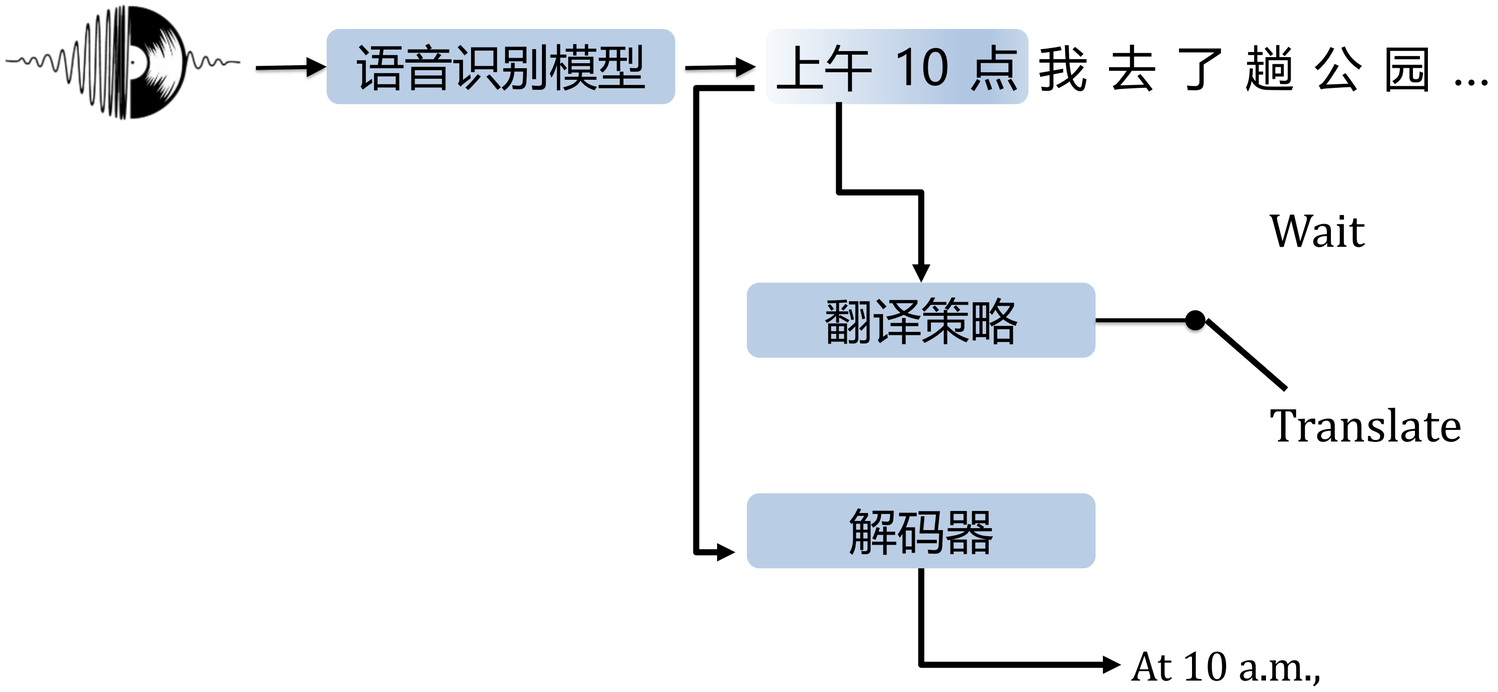
同传系统



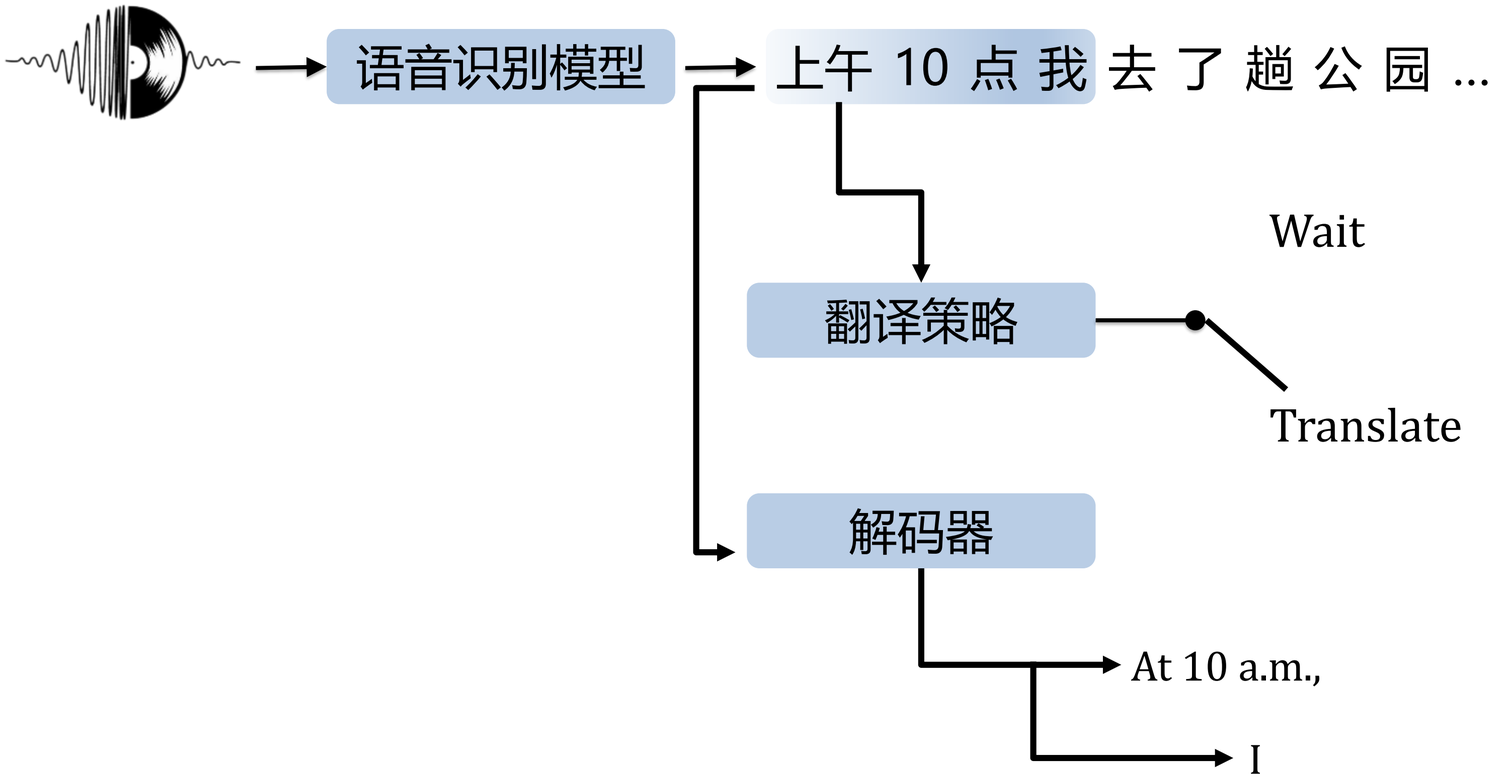
同传系统



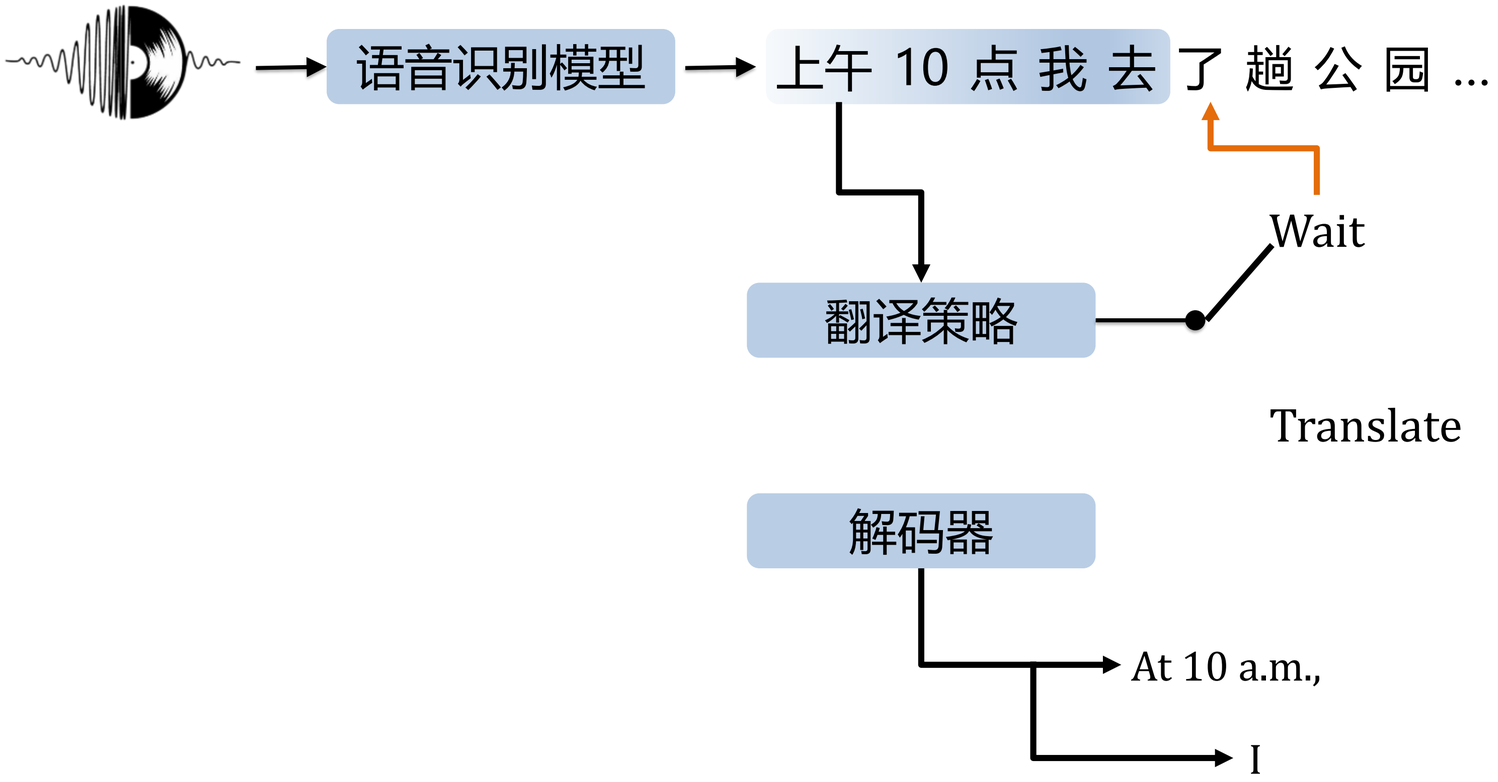
同传系统



同传系统



同传系统

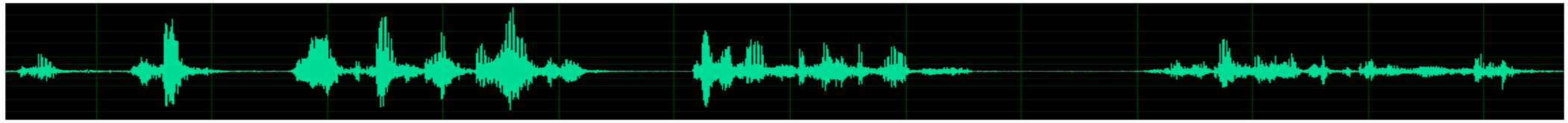


同传关键问题

挑战一、语音识别错误

演讲稿　很 高 兴　参 加 百 度 举 办 的　这 次 大 会 。　出 席 今 天 会 议 的 ， 还 有

语音信号



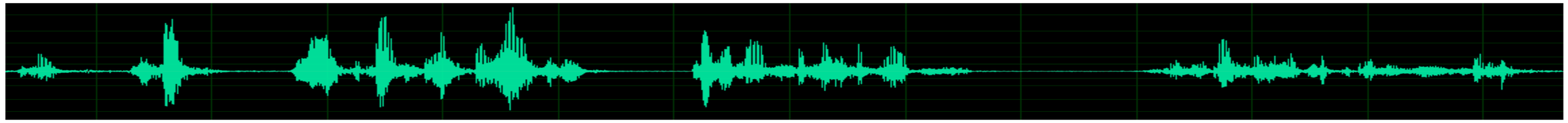
ASR　很 高 兴　参 加 百 度 举 办 的　这 次 大 会 。　出 席 经 典 会 的 ， 还 有

语音识别错误

同传关键问题

难点二、平衡翻译效果和时延

演讲稿　很 高 兴　参 加 百 度 举 办 的　这 次 大 会 。　出 席 今 天 会 议 的 ， 还 有



语音信号

ASR　很 高 兴　参 加 百 度 举 办 的　这 次 大 会 。　出 席 经 典 会 的 ， 还 有

翻译结果

I’ m glad to　participate in this conference held by Baidu

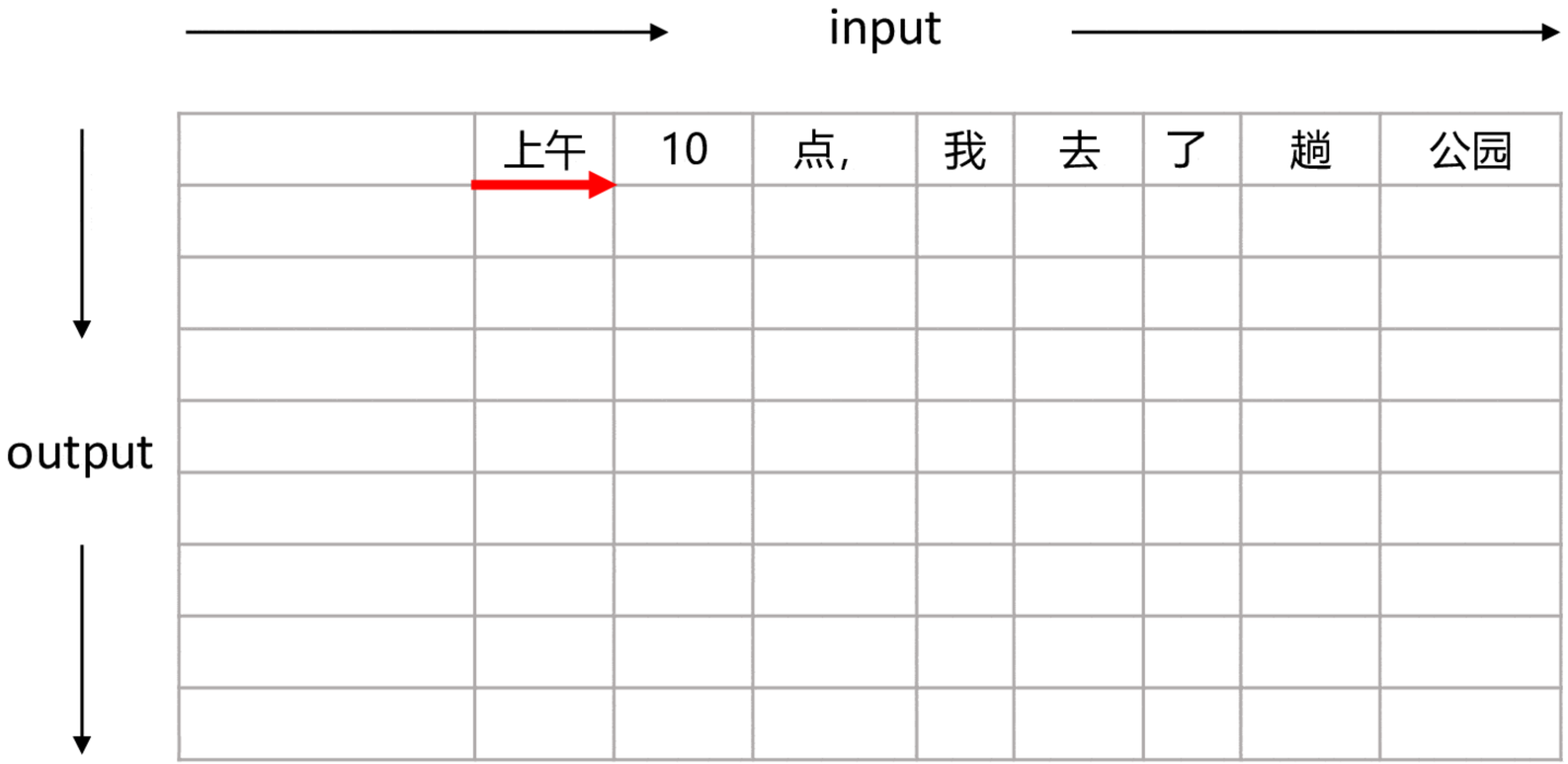
很 高 兴　参 加 百 度　举 办 的　这 次 大 会

I’ m glad to　join Baidu　to hold this conference.

同传翻译策略错误

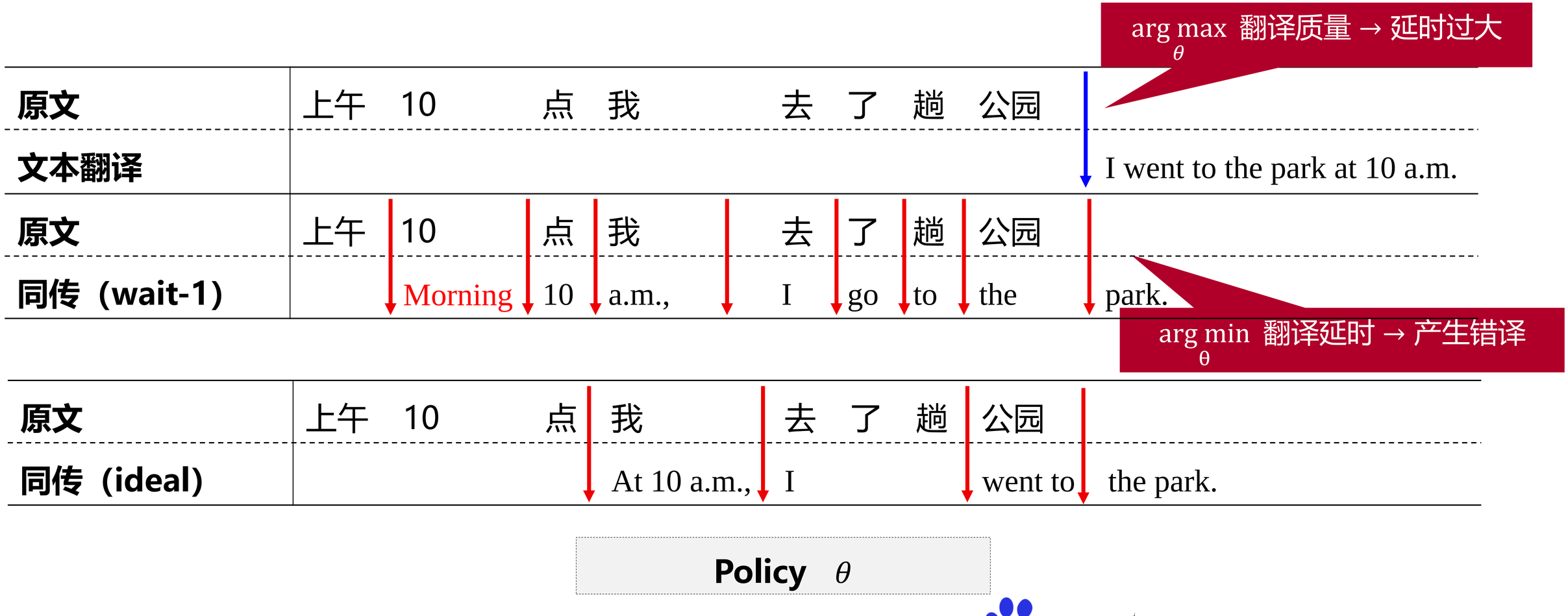
同传关键问题

难点二、平衡翻译效果和时延



同传关键问题

难点二、平衡翻译效果和时延

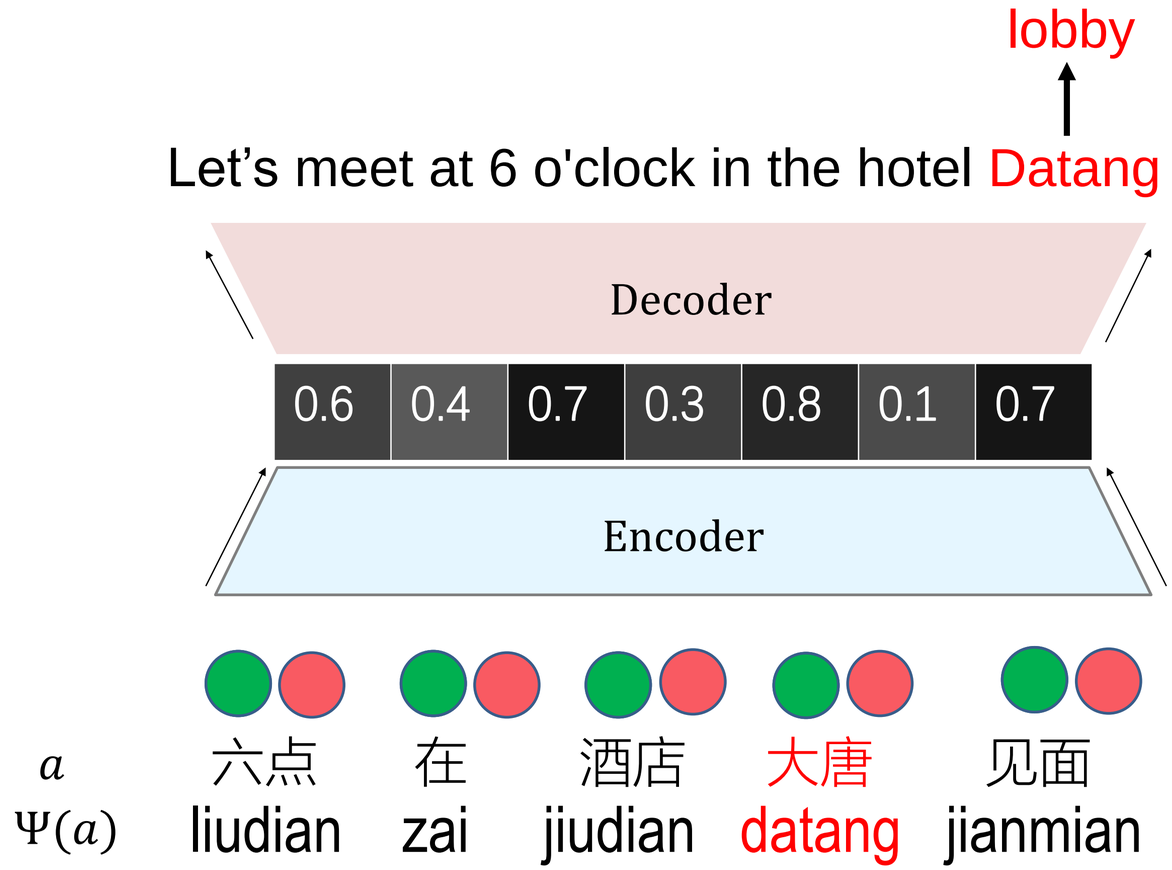


同传关键问题

挑战	解决方案
语音识别错误	更鲁棒的ASR / 纠错
高翻译准确和低延时之间矛盾	Policy

同传ASR错误解决方案

NMT容错模型——鲁棒的NMT模型



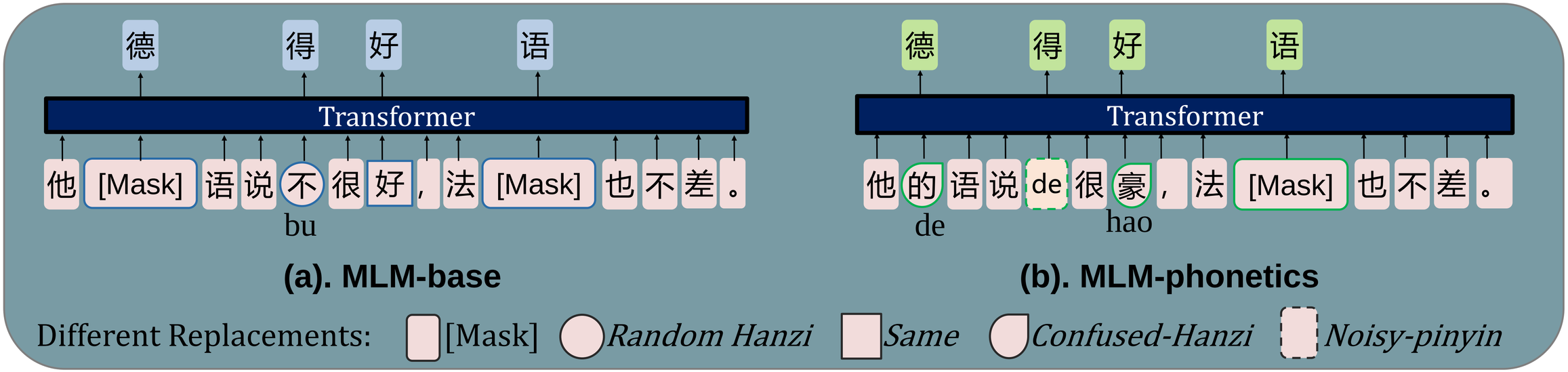
文本向量　　音节向量

$$\pi([a, \psi(a)]) = (1-\beta) * \pi(a) + \beta * \pi(\psi(a))$$

Liu et al., Robust Neural Machine Translation with Joint Textual and Phonetic Embedding, ACL 2019

同传ASR错误解决方案——文本纠错模型

ASR纠错模型——将语音信息融入预训练



Zhang et al., Correcting Chinese Spelling Errors with Phonetic Pre-training, Findings of ACL 2021

同传ASR错误解决方案——文本纠错模型

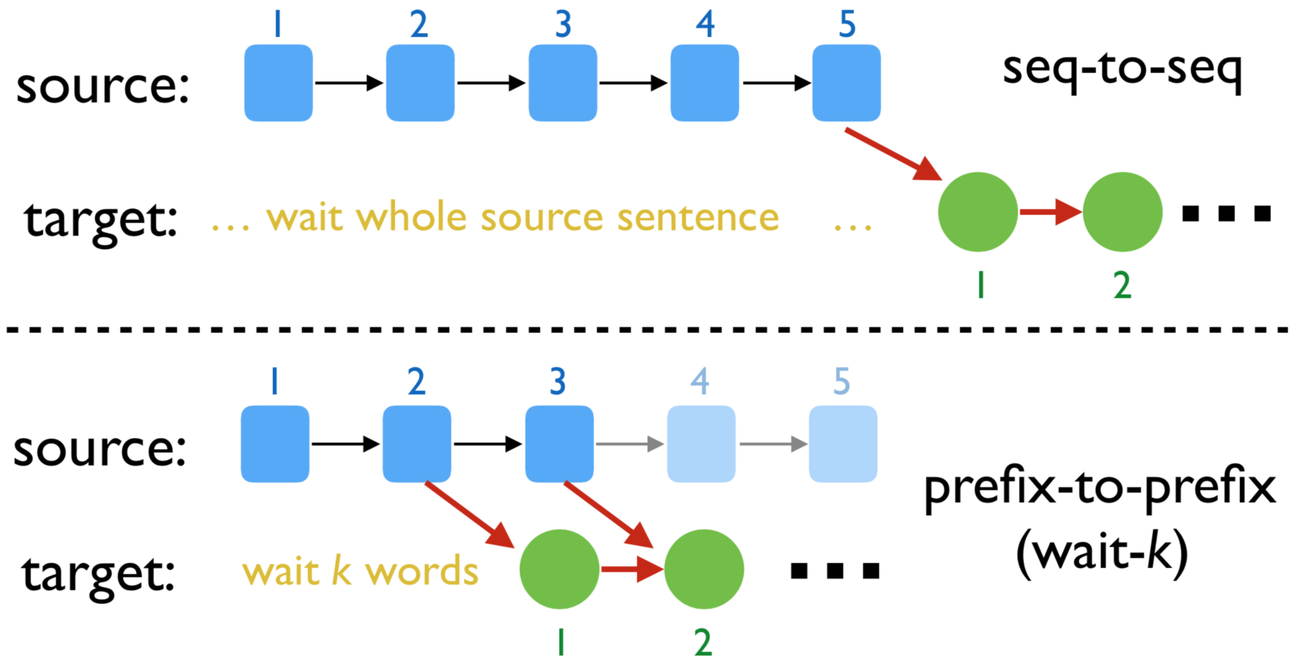
ASR纠错模型——将语音信息融入预训练

	Detection			Correction		
	Prec.	Rec.	F1	Prec.	Rec.	F1
SIGHAN13						
FASPell (2019)	76.2	63.2	69.1	73.1	60.5	66.2
Pointer Networks (2019) (character-level)	56.8	91.4	70.1	79.7	59.4	68.1
*Soft-Masked BERT**	81.1	75.7	78.3	75.1	70.1	72.5
SpellGCN (2020)	80.1	74.4	77.2	78.3	72.7	75.4
ERNIE	76.6	71.9	74.2	73.0	68.5	70.6
MLM-phonetics(Ours)	**82.0**	**78.3**	**80.1**	**79.5**	**77.0**	**78.2**
SIGHAN14						
FASPell (2019)	61.0	53.5	57.0	59.4	52.0	55.4
Pointer Networks (2019) (character-level)	63.2	82.5	71.6	79.3	68.9	73.7
*Soft-Masked BERT**	65.2	70.4	67.7	63.7	68.7	66.1
SpellGCN (2020)	65.1	69.5	67.2	63.1	67.2	65.3
ERNIE	63.5	69.3	66.3	60.1	65.6	62.8
MLM-phonetics(Ours)	**66.2**	**73.8**	**69.8**	**64.2**	**73.8**	**68.7**
SIGHAN15						
FASPell (2019)	67.6	60.0	63.5	66.6	59.1	62.6
Pointer Networks (2019) (character-level)	66.8	73.1	69.8	71.5	59.5	64.9
Soft-Masked BERT (2020)	73.7	73.2	73.5	66.7	66.2	66.4
*Soft-Masked BERT**	67.6	78.7	72.7	63.4	73.9	68.3
SpellGCN (2020)	74.8	80.7	77.7	72.1	77.7	75.9
ERNIE	73.6	79.8	76.6	68.6	74.4	71.4
MLM-phonetics(Ours)	**77.5**	**83.1**	**80.2**	**74.9**	**80.2**	**77.5**

Zhang et al., Correcting Chinese Spelling Errors with Phonetic Pre-training, Findings of ACL 2021

同传平衡翻译质量、延时解决方案

可控延时的翻译模型——WaitK



Ma et al., STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency, ACL 2019

同传平衡翻译质量、延时解决方案

可控延时的翻译模型——WaitK

Ma et al., STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency, ACL 2019

同传平衡翻译质量、延时解决方案

可控延时的翻译模型——WaitK

Training samples for Wait-1

Source	*Target*
他	
他还说	He
他还说现在	He said
他还说现在正在	He said he
他还说现在正在为	He said he is
他还说现在正在为这一	He said he is making
他还说现在正在为这一会议	He said he is making preparation
他还说现在正在为这一会议作出	He said he is making preparation for
他还说现在正在为这一会议作出安排	He said he is making preparation for this
他还说现在正在为这一会议作出安排。	He said he is making preparation for this meeting.

Ma et al., STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency, ACL 2019

同传平衡翻译质量、延时解决方案

可控延时的翻译模型——WaitK

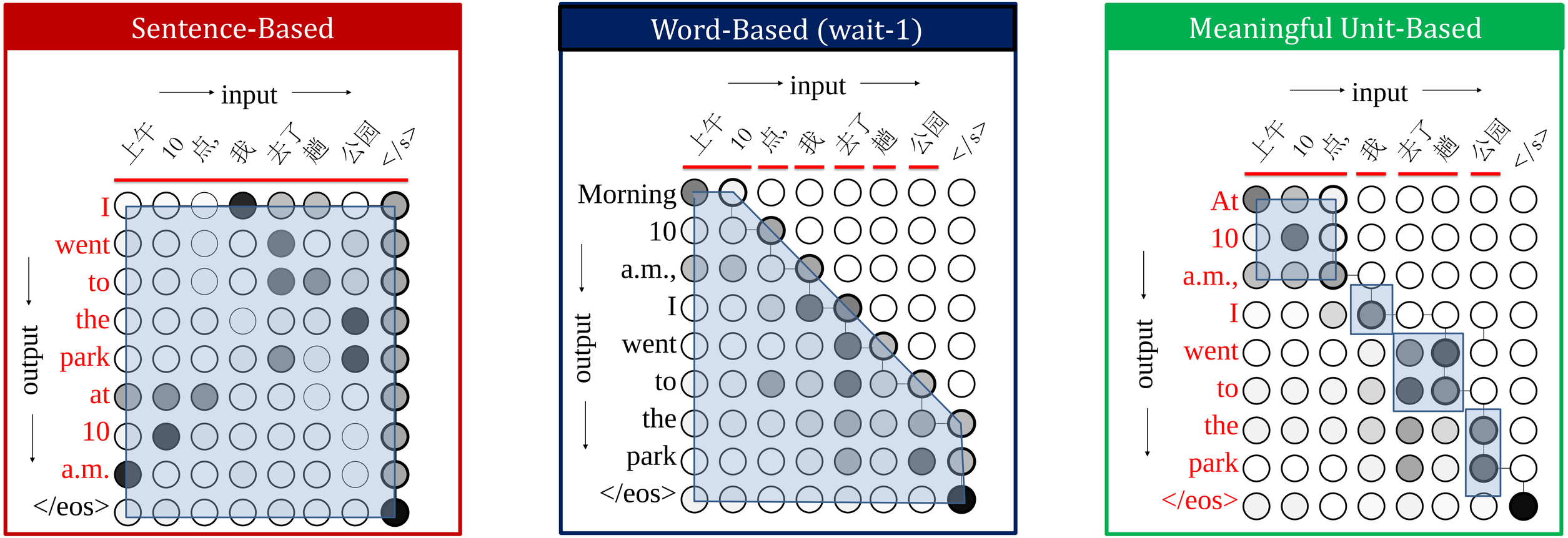
Training samples for Wait-1

Source	*Target*
他	
他还说	He
他还说现在	He said
他还说现在正在	He said he
他还说现在正在为	He said he is
他还说现在正在为这一	He said he is making
他还说现在正在为这一会议	He said he is making preparation
他还说现在正在为这一会议作出	He said he is making preparation for
他还说现在正在为这一会议作出安排	He said he is making preparation for this
他还说现在正在为这一会议作出安排。	He said he is making preparation for this meeting.

Ma et al., STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency, ACL 2019

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit



Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

信息单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

信息单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

信息单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

语义单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

语义单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						
$M'_{nmt}(x_{\leq 4})$	At 10 a.m., I						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

语义单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						
$M'_{nmt}(x_{\leq 4})$	At 10 a.m., I						
$M'_{nmt}(x_{\leq 5})$	At 10 a.m., I went there						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

语义单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						
$M'_{nmt}(x_{\leq 4})$	At 10 a.m., I						
$M'_{nmt}(x_{\leq 5})$	At 10 a.m., I went there						
$M'_{nmt}(x_{\leq 6})$	At 10 a.m., I went to						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

语义单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						
$M'_{nmt}(x_{\leq 4})$	At 10 a.m., I						
$M'_{nmt}(x_{\leq 5})$	At 10 a.m., I went there						
$M'_{nmt}(x_{\leq 6})$	At 10 a.m., I went to						
$M'_{nmt}(x_{\leq 7})$	At 10 a.m., I went to the park.						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit

语义单元：翻译内容不会随下文而变化的 最小片段。

如何定义信息单元？ 目标：翻译准确、延时较小

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						
$M'_{nmt}(x_{\leq 4})$	At 10 a.m., I						
$M'_{nmt}(x_{\leq 5})$	At 10 a.m., I went there						
$M'_{nmt}(x_{\leq 6})$	At 10 a.m., I went to						
$M'_{nmt}(x_{\leq 7})$	At 10 a.m., I went to the park.						
Extracted meaningful units	上午 10 点 | 我 | 去了 趟 | 公园 | 。						

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit++

Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	At 10 a.m., I went to the park.						
Extracted meaningful units	上午 10 点 | 我 | 去了 趟 | 公园 | 。						

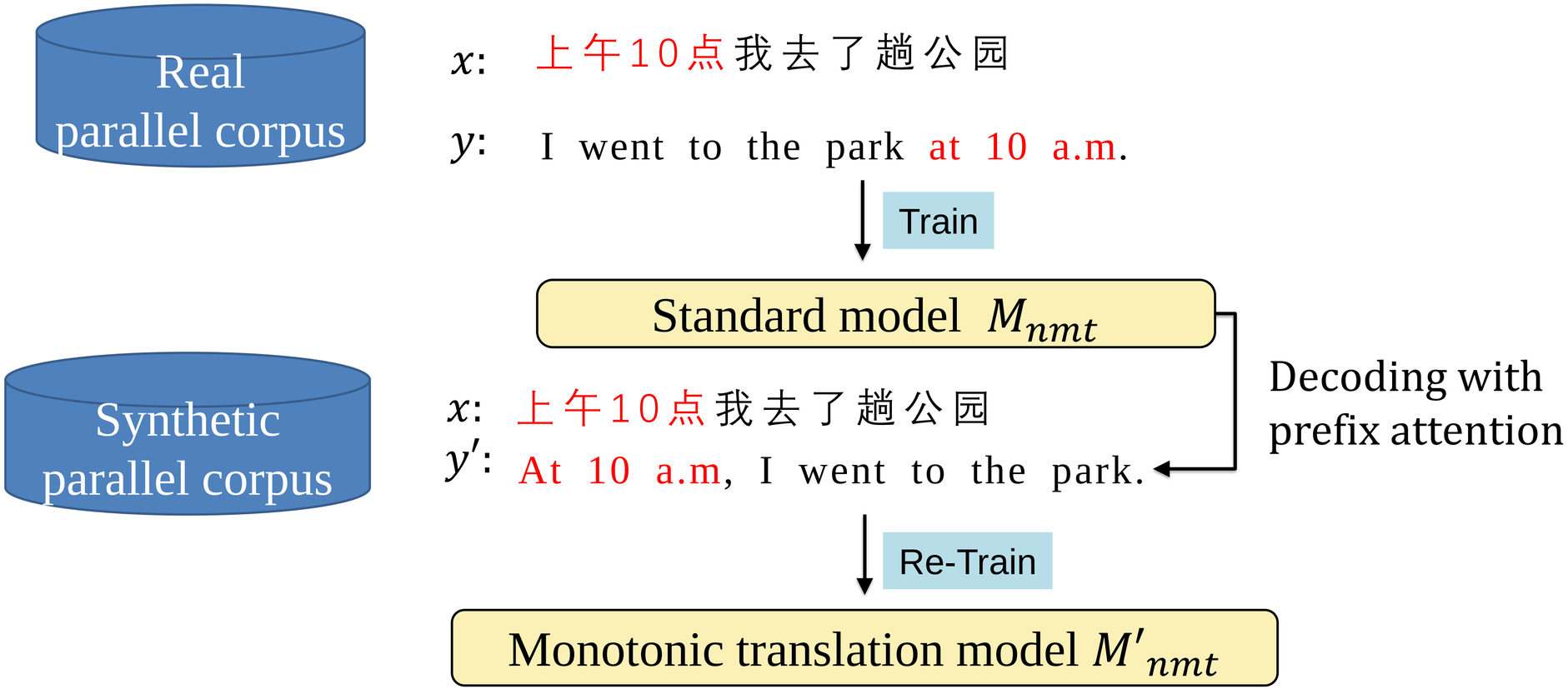
Source	shàngwǔ	10	diǎn	wǒ	qùle	tàng	gōngyuán
	上午	10	点	我	去了	趟	公园
full translation	**I went to the park at 10 a.m.**						
$M'_{nmt}(x_{\leq 1})$	Morning						
$M'_{nmt}(x_{\leq 2})$	Morning 10						
$M'_{nmt}(x_{\leq 3})$	At 10 a.m.						
$M'_{nmt}(x_{\leq 4})$	At 10 a.m., I						
$M'_{nmt}(x_{\leq 5})$	At 10 a.m., I went there						
$M'_{nmt}(x_{\leq 6})$	At 10 a.m., I went to						
$M'_{nmt}(x_{\leq 7})$	At 10 a.m., I went to the park.						
Extracted meaningful units	上午 10 点 我 去了 趟 公园 。						

翻译长距离调序导致无法生成短MU

Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

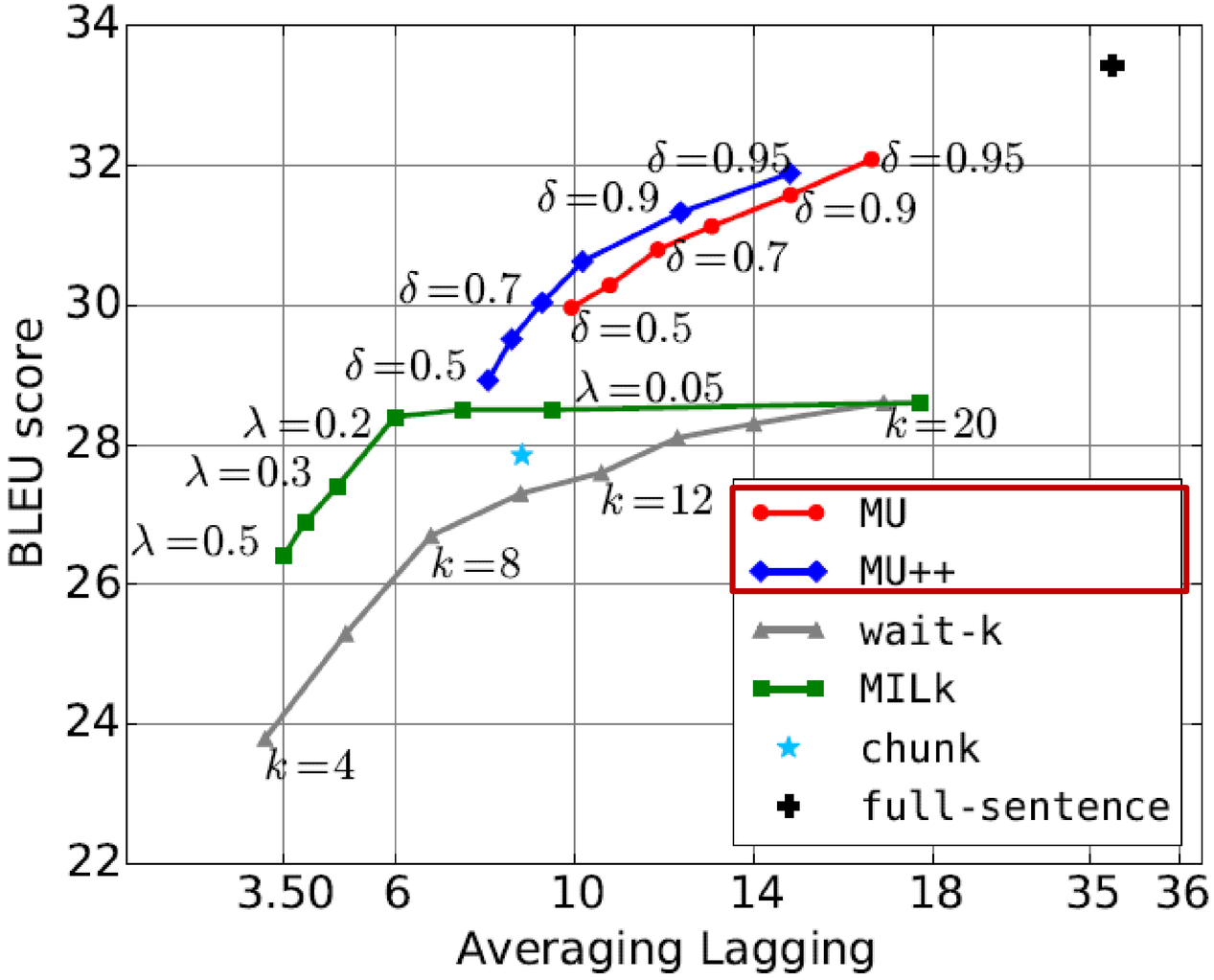
自适应同传切分模型——Meaningful Unit++



Zhang et al. Learning Adaptive Segmentation Policy for Simultaneous Translation. In EMNLP 2020

同传平衡翻译质量、延时解决方案

自适应同传切分模型——Meaningful Unit++



Experiment on WMT15 En-De

- **Wait-k**: First waiting for k words, then emiting one token after reader each word
- **chunk**: Generate MU training corpus according to GIZA++
- **MILK**: train the policy together with the NMT model in an end-to-end framework.

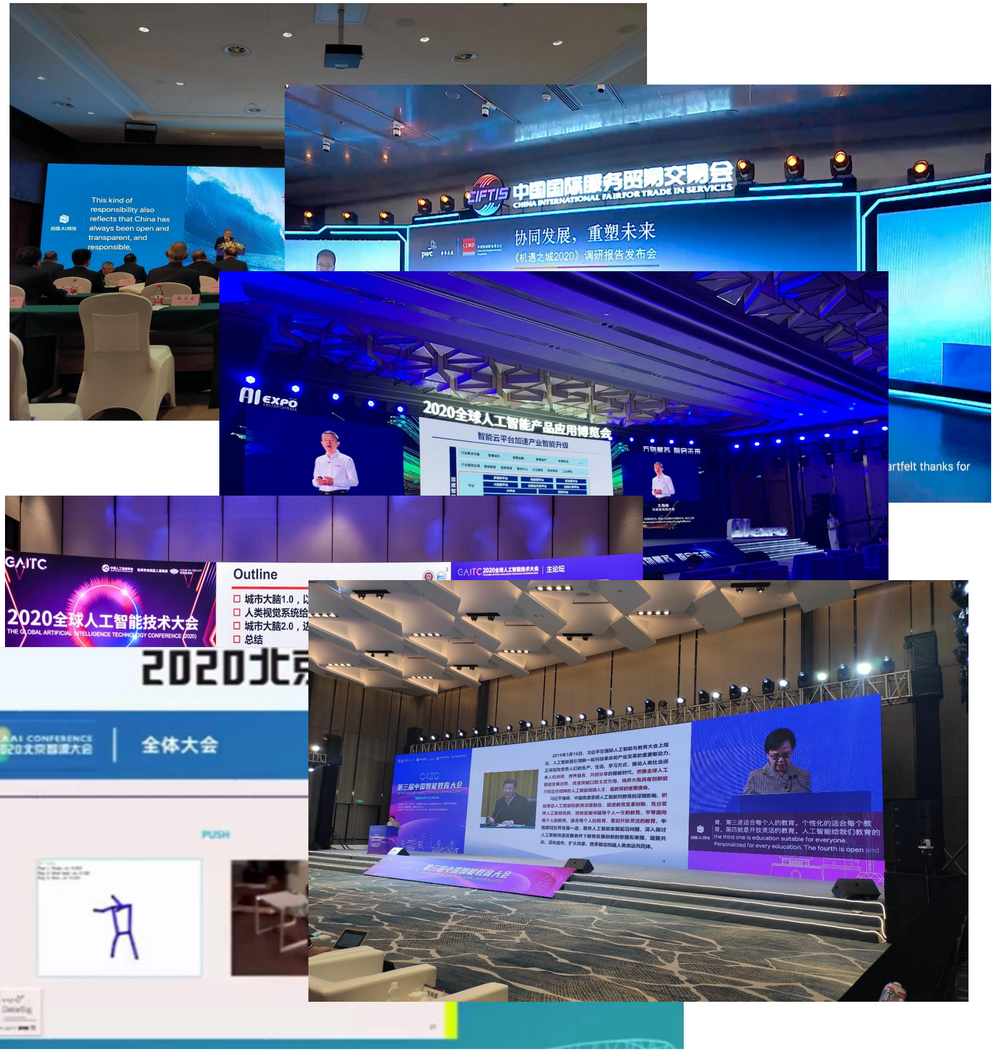
03

中英同传产业应用

业界同传应用场景

中英同传关键问题——数据

中英同传应用——百度AI同传

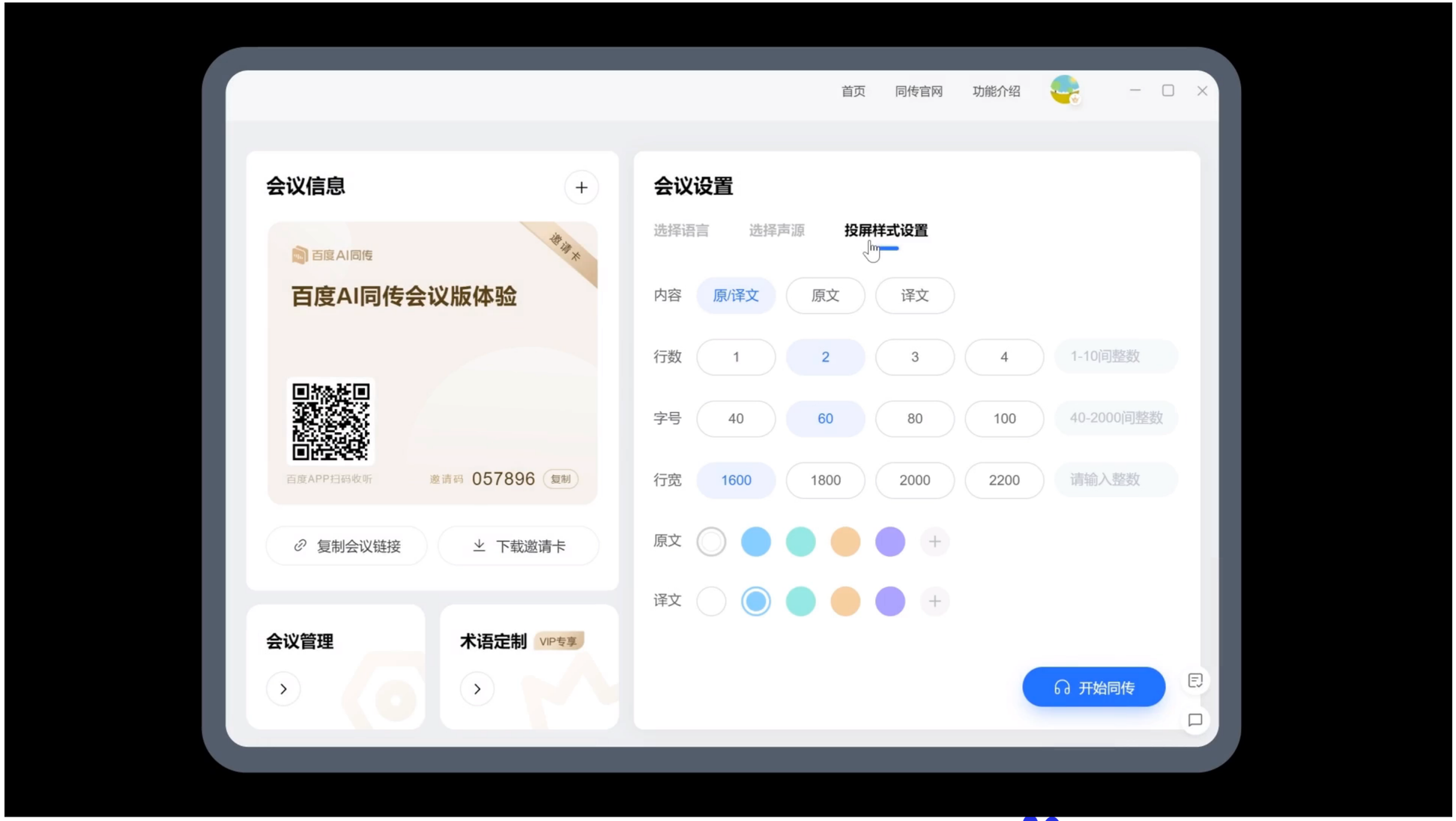


支持百余场线上线下会议

小度无线耳机同传功能

https://api.fanyi.baidu.com/simultaneous?fr=pinzhuan

中英同传应用——百度AI同传（企业版）



https://api.fanyi.baidu.com/simultaneous?fr=pinzhuan

中英同传应用——百度AI同传助手

https://api.fanyi.baidu.com/simultaneous?fr=pinzhuan

中英同传应用——中英语音翻译数据集BSTC

Speech Translation	Languages	Hours
F-C (2013)	Es→En	38
KIT-Disfluency (2014)	De→En	13
BTEC (2016)	En→Fr	17
MSLT V1.0 (2016)	En↔Fr/De	23
	En→Zh/Jp	6
MSLT V1.1 (2017)	Zh→En	5
	Jp →En	9
Travel (2017)	Am→En	8
Aug-LibriSpeech (2018)	En→Fr	236
MuST-C (2019)	En→8 Euro langs	3617
Europarl-ST (2020)	9 Euro langs	1642
Covost (2020a; 2020b)	En↔21 langs	2880
Simultaneous Translation	**Languages**	**Hours**
CIAIR (2004)	En↔Jp	182
EPPS (2009)	En↔Es	217
Simul-Trans (2014)	En↔Jp	22
BSTC (ours)	Zh→En	68

{
"offset": "105.975",
"duration": "3.287",
"wav": "2.wav",
"transcript": "但是你们的每个人都有多个设备，啊有手持设备，有手机。",
"Streaming ASR":

Type: partial 但是
Type: partial 但是你们
Type: partial 但是你们的没
Type: partial 但是你们的没个人都
Type: partial 但是你们的没个人都有多个
Type: final 但是你们的没个人都有多个设备
Type: partial 啊有
Type: partial 啊有首
Type: partial 啊有手持摄
Type: final 啊有手持设备
Type: partial 首
Type: partial 手机

"translation": "In fact, every one of you has multiple digital devices, handheld devices and mobile phones.",
"interpreter A": "But actually you own several devices, mobile devices, mobile phones.",
"interpreter B": "But every of you have multiple equipments with you hand held equipment like phone, smartphone.",
"interpreter C": "But every one of you have multi devices, we have mobile phones."
}

Zhang et al. BSTC: A Large-Scale Chinese-English Speech Translation Dataset. In 2nd workshop on AutoSimulTrans (at NAACL21)

04

未来展望

数据

鲁棒性

可控性

评估

未来挑战

1.数据缺乏

2.评估方法

3.同传应用场景

Zhang et al. BSTC: A Large-Scale Chinese-English Speech Translation Dataset. In 2nd workshop on AutoSimulTrans

Baidu 百度 | DataFunSummit

THANKS!

今天的分享就到这里...

Ending