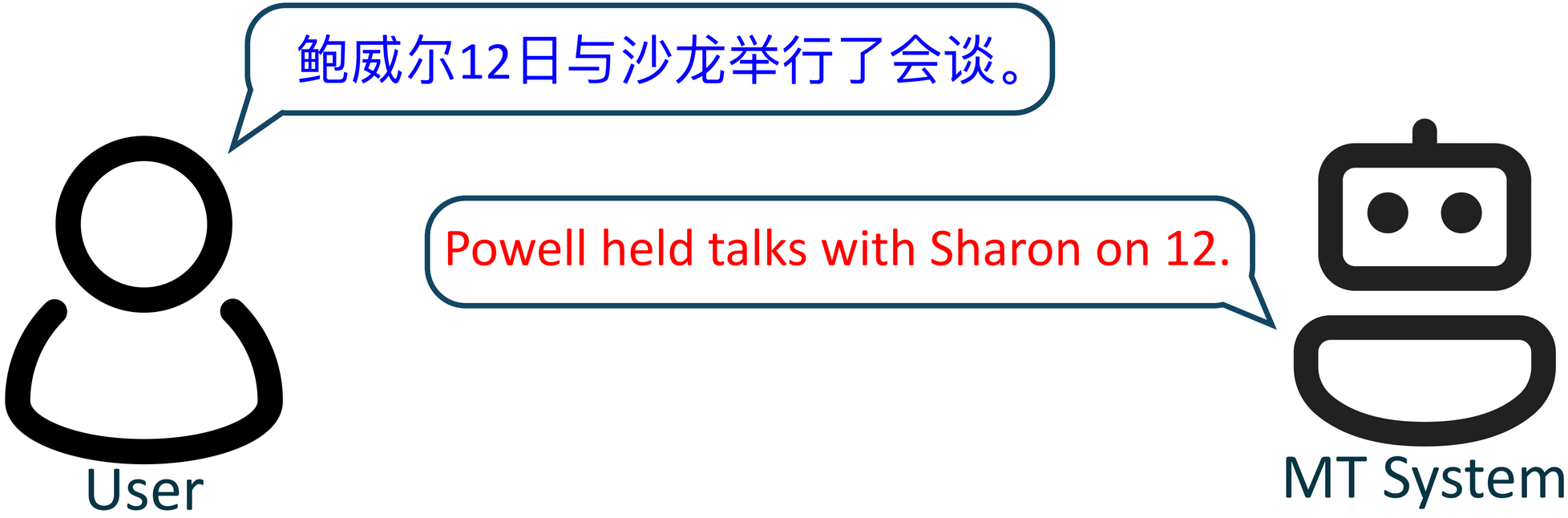
从机器翻译到同声传译：挑战与进展

张绍磊

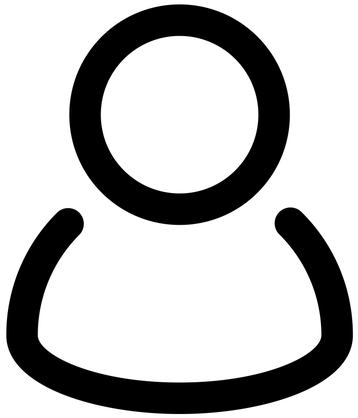
中国科学院计算技术研究所

2022.3.27　MLNLP学术研讨会

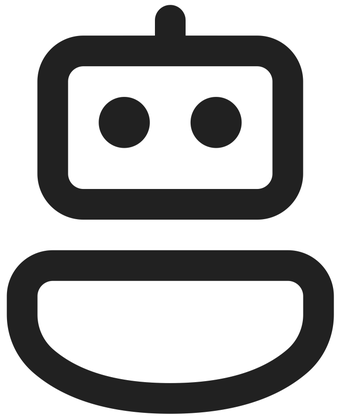
机器翻译（Machine Translation）



同声传译（Simultaneous Translation）

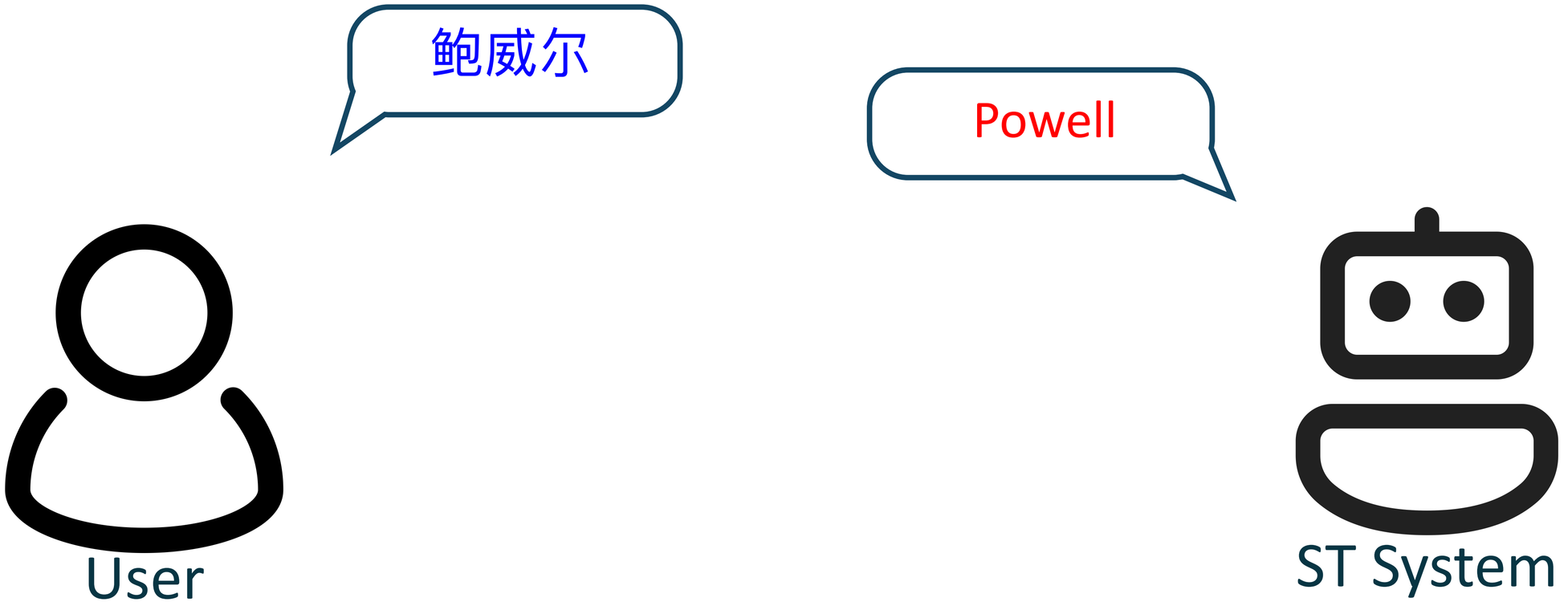


User



ST System

同声传译（Simultaneous Translation）



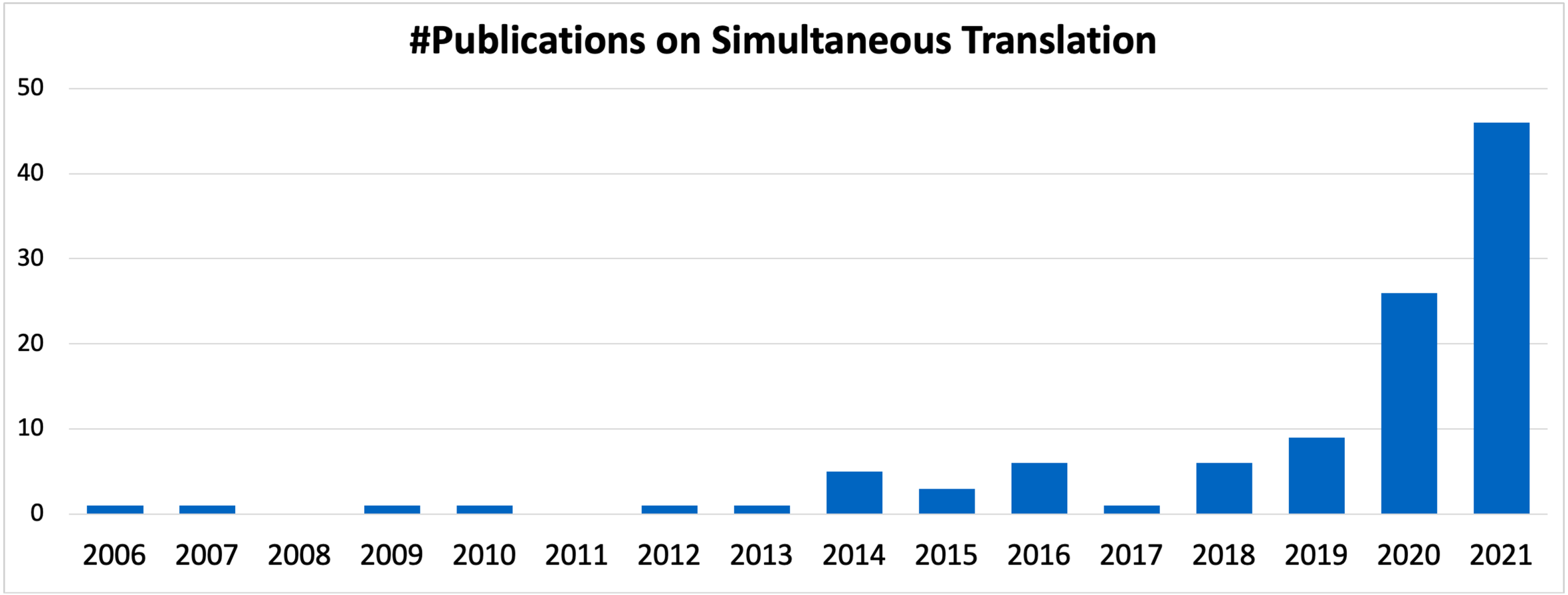
同声传译（Simultaneous Translation）

同声传译（Simultaneous Translation）

同声传译（Simultaneous Translation）

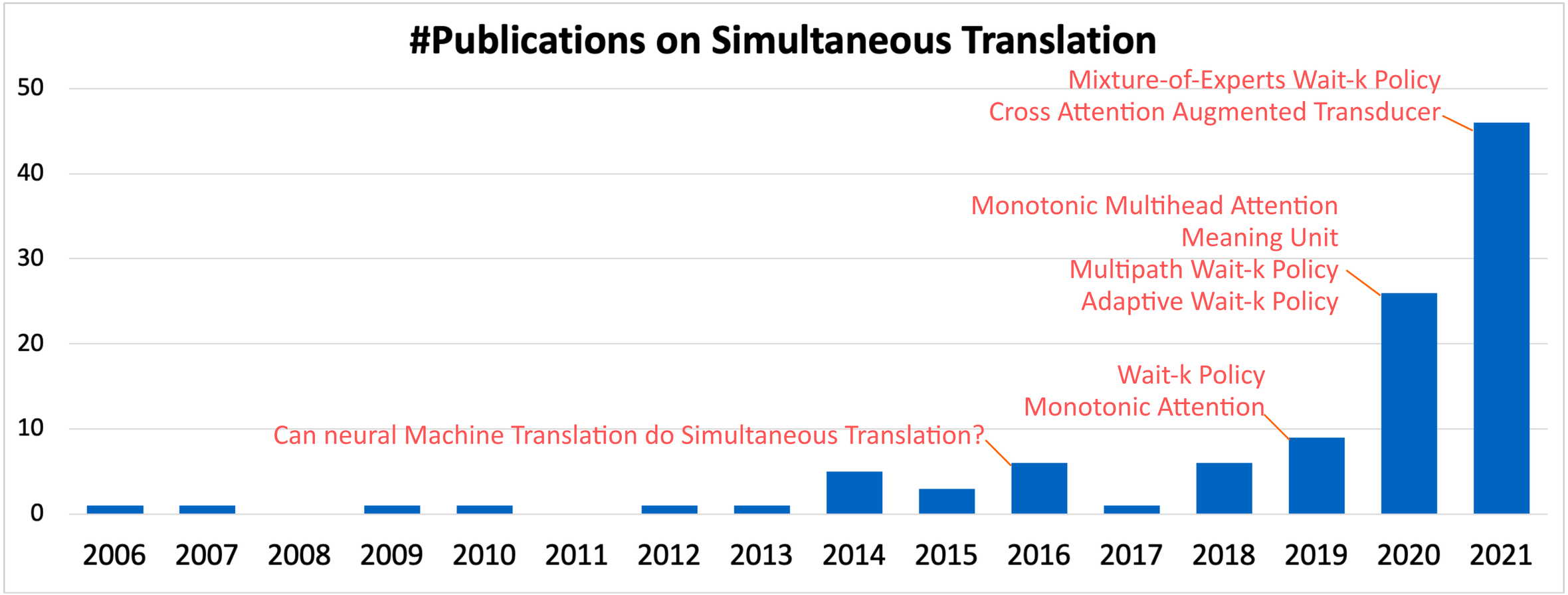
翻译质量 & 翻译延时

同声传译研究发展



https://github.com/Vily1998/Awesome-Simultaneous-Translation

同声传译研究发展



https://github.com/Vily1998/Awesome-Simultaneous-Translation

机器翻译 → 同声传译：差异和挑战

机器翻译

V.S.

同声传译

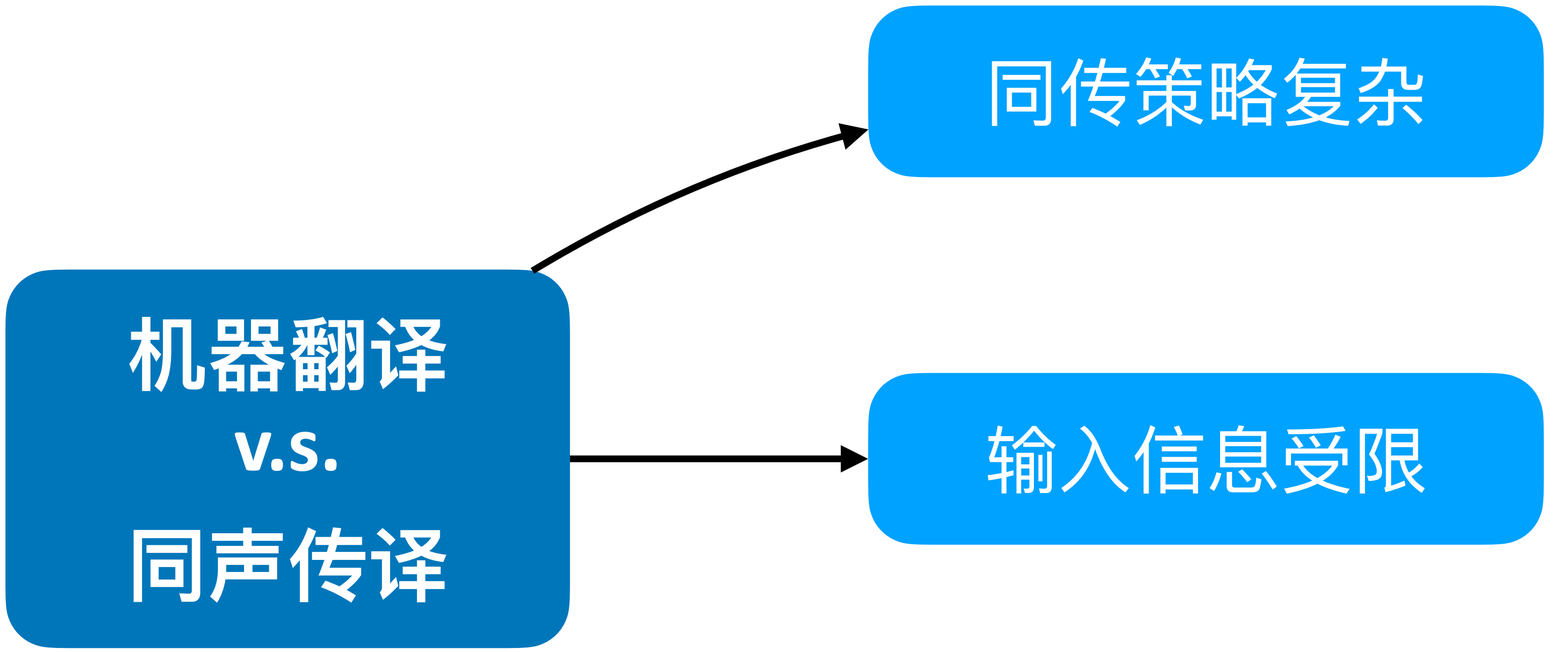
机器翻译 → 同声传译：差异和挑战

机器翻译
v.s.
同声传译

同传策略复杂

输出翻译 or 等待输入？

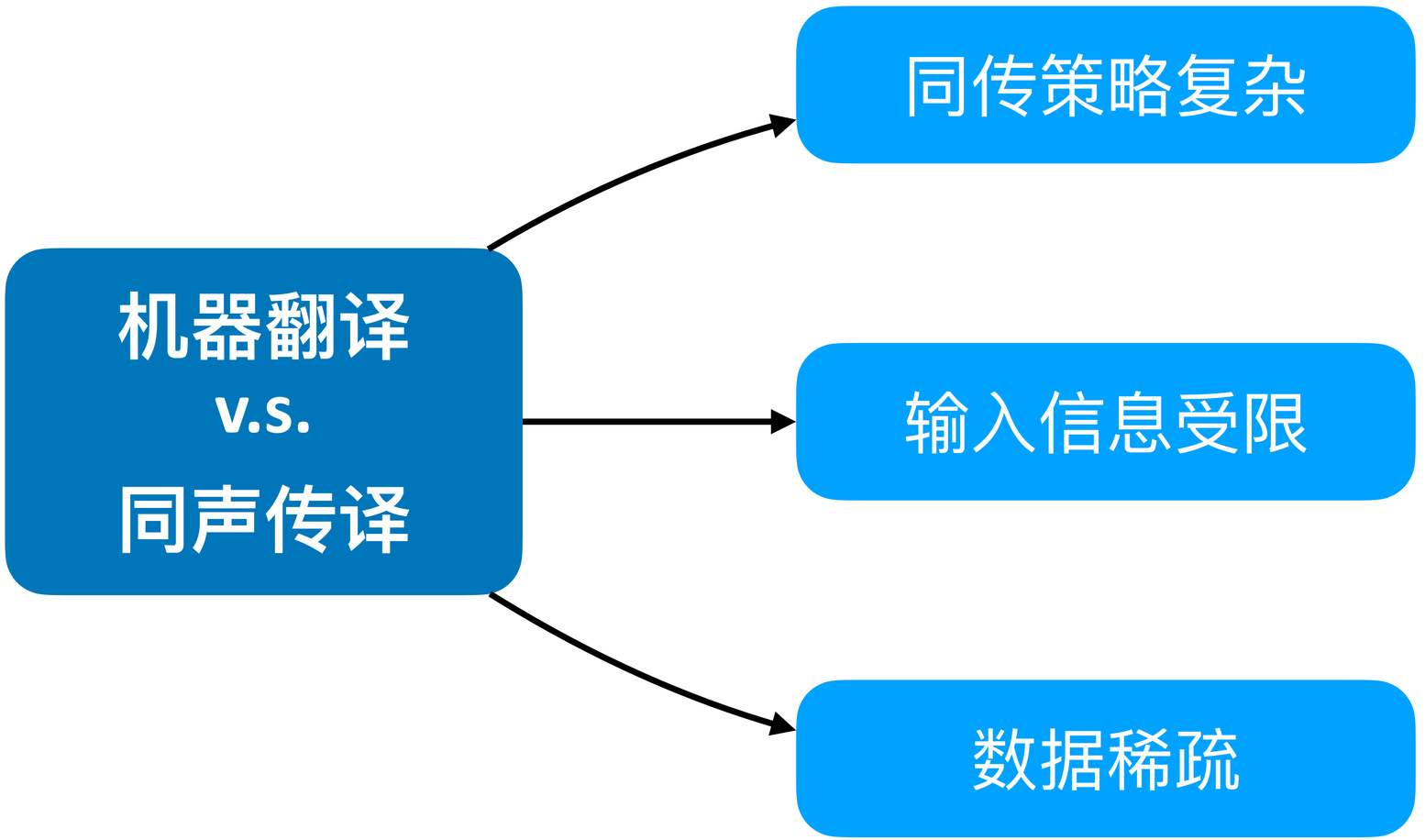
机器翻译 → 同声传译：差异和挑战



输出翻译 or 等待输入？

仅接受部分源信息时，
如何生成正确翻译?

机器翻译 → 同声传译：差异和挑战



输出翻译 or 等待输入？

仅接受部分源信息时，
如何生成正确翻译?

同传数据存在领域差异且稀疏，如何训练?

同传策略：读 or 写?

- 同传模型需要在每个时刻，决定一个读/写操作（action）
  - 读：继续等待下一个源输入
  - 写：翻译并输出一个目标词
  - 直接决定延时 & 翻译质量

Decoding Step

Source sentence	鲍威尔	12日	与	沙龙	举行	了	会谈	。							
Full-sentence NMT	'WAIT' for the whole source sentence								Powell	held	talks	with	Sharon	on	12	.
Simultaneous Translation	'WAIT'	'WAIT'	Powell	held	'WAIT'	talks	with	Sharon	on	12	.					

starts translations synchronously while reading source sentences

同传策略：读 or 写?

- 同传模型需要在每个时刻，决定一个读/写操作（action）
  - 读：继续等待下一个源输入
  - 写：翻译并输出一个目标词
  - 直接决定延时 & 翻译质量

- **固定策略**
- **自适应策略**
- **分段策略**

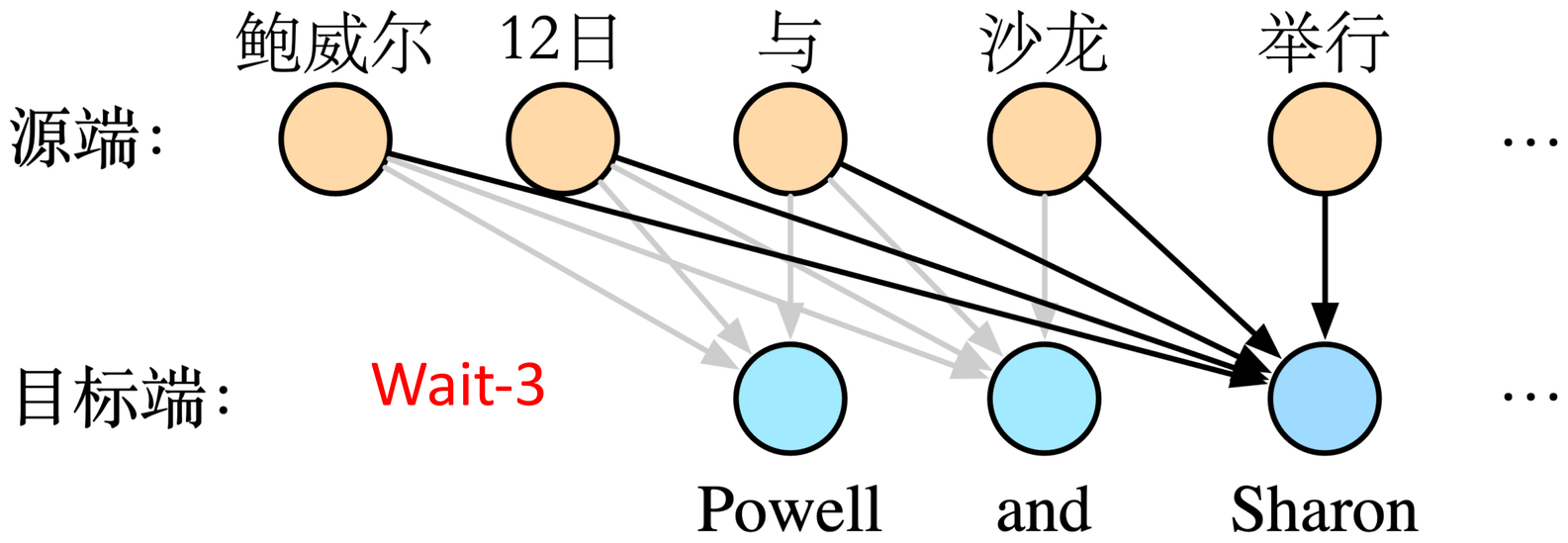
Decoding Step

Source sentence	鲍威尔	12日	与	沙龙	举行	了	会谈	。								
Full-sentence NMT	'WAIT' for the whole source sentence								Powell	held	talks	with	Sharon	on	12	.
Simultaneous Translation	'WAIT'	'WAIT'	Powell	held	'WAIT'	talks	with	Sharon	on	12	.					

starts translations synchronously while reading source sentences

固定策略：简单、易训练

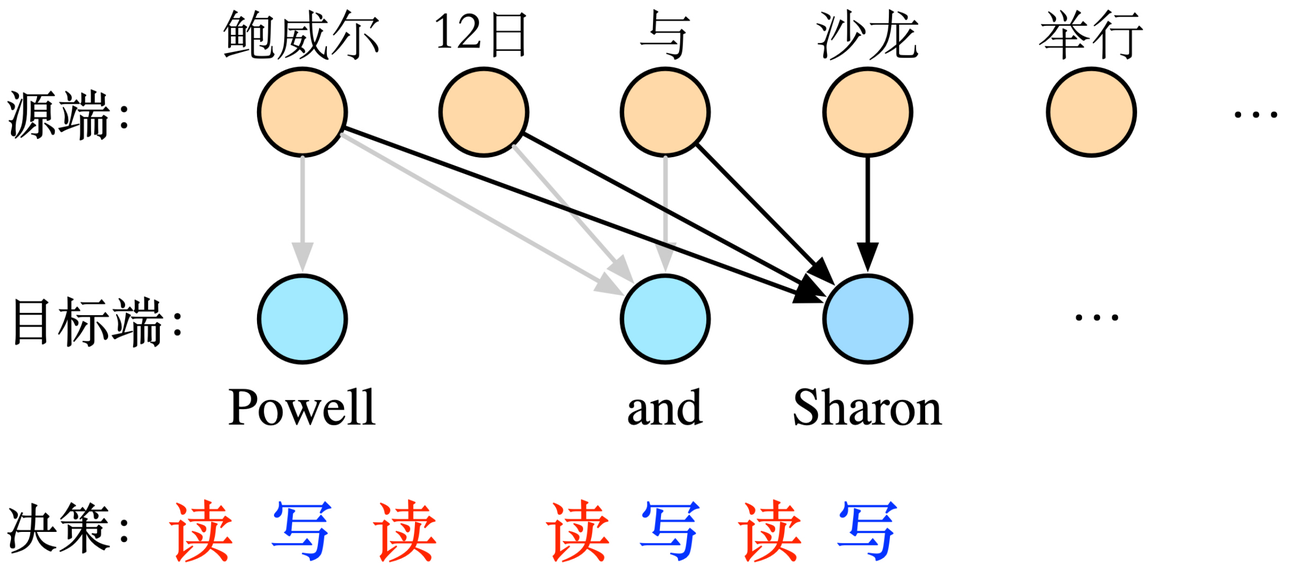
- 策略由人为预先设置好，在翻译过程中固定不变。
  - **Wait-k Policy**：翻译滞后输入 k 个词

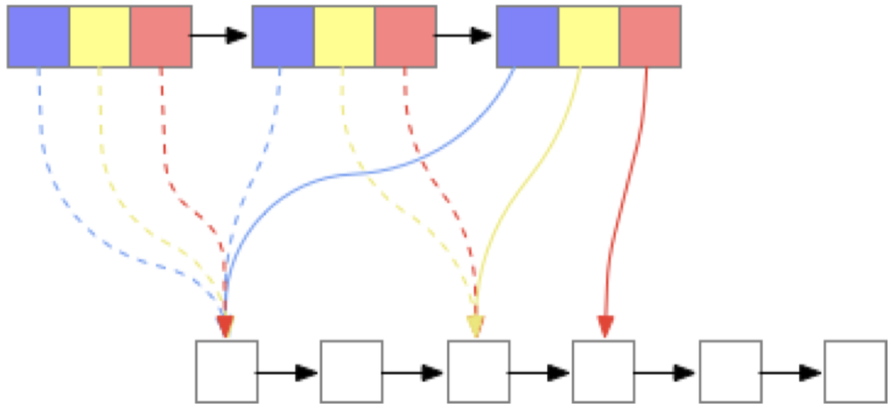


- STACL: Simultaneous Translation with Implicit Anticipation and Controllable Latency using Prefix-to-Prefix Framework. *ACL 2019*.
- Incremental Decoding and Training Methods for Simultaneous Translation in Neural Machine Translation. *NAACL 2018*.

自适应策略：更好的翻译质量

- 策略由模型生成，在翻译过程中动态调整。
  - **Monotonic attention**：预测伯努利变量0/1来表示读/写



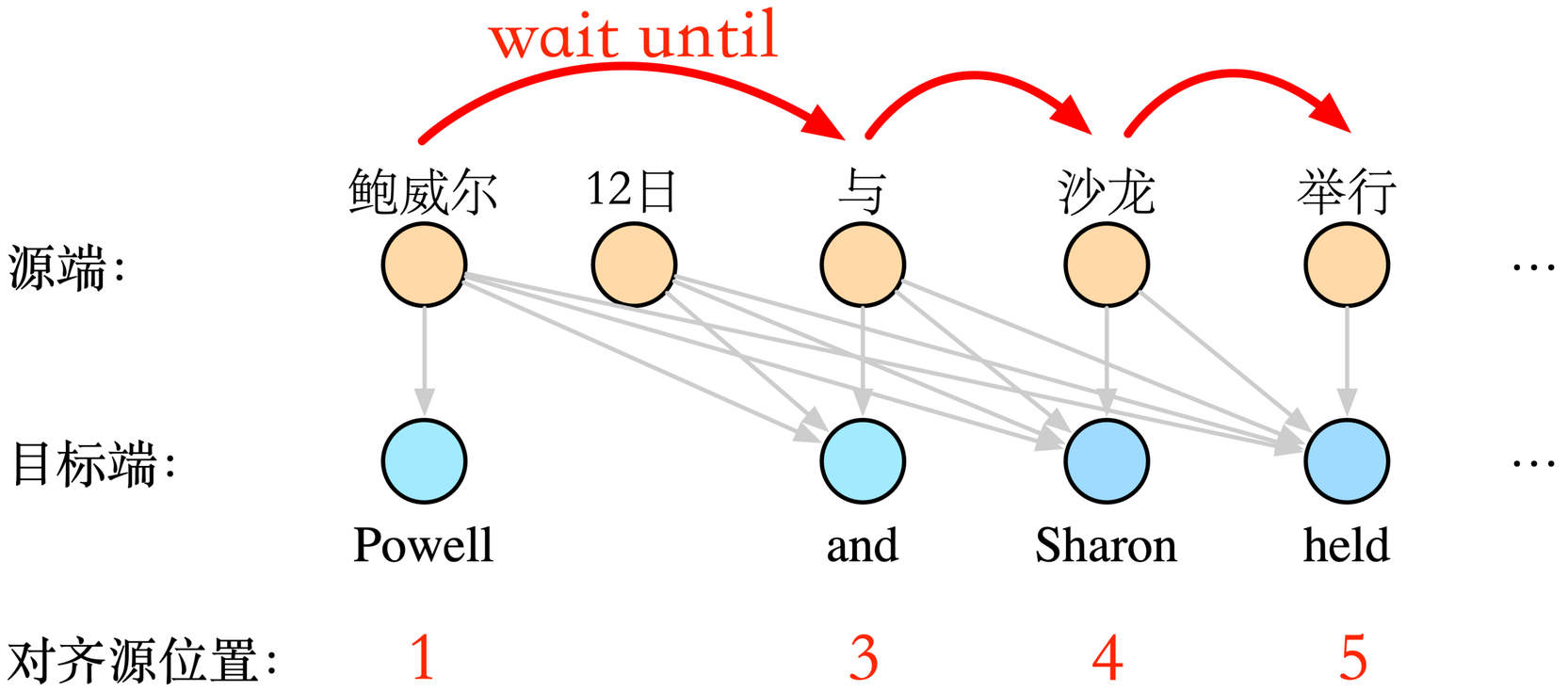


monotonic multihead attention

- Online and Linear-Time Attention by Enforcing Monotonic Alignments. *ICML 2017*.
- Monotonic Chunkwise Attention. *ICLR 2018*.
- Monotonic Infinite Lookback Attention for Simultaneous Machine Translation. *ACL 2019*.
- Monotonic Multihead Attention. *ICLR 2020*.

自适应策略：更好的可解释性

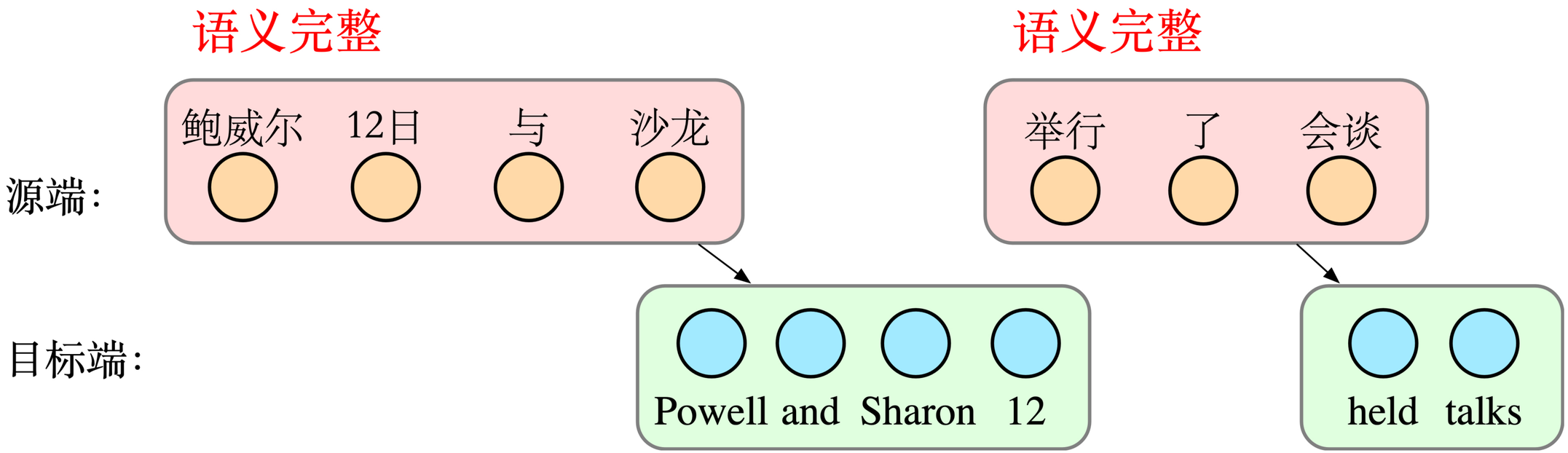
- 策略由模型生成，在翻译过程中动态调整。
  - **Gaussian multihead attention**：预测对齐位置，读到对齐位置之后开始写。



• Gaussian Multi-head Attention for Simultaneous Machine Translation. *ACL 2022 findings*.

分段策略：基于整句翻译模型

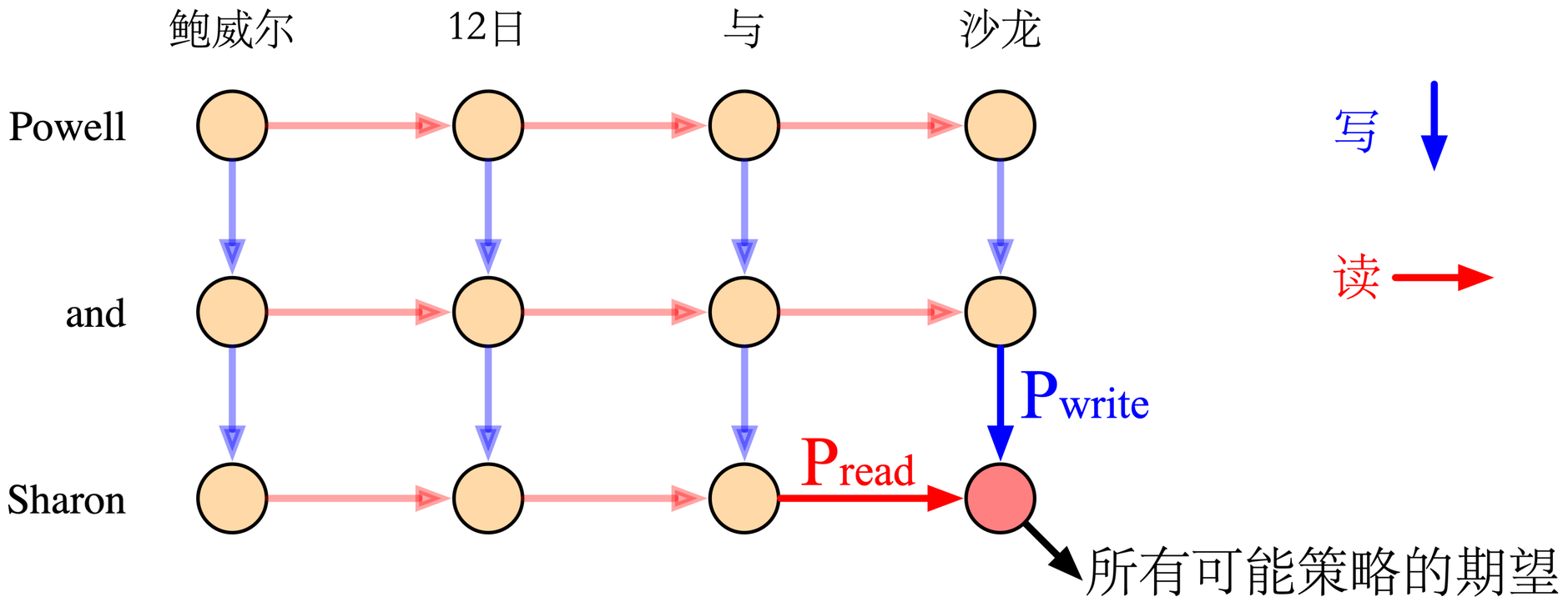
- 将输入划分为若干片段，在每个片段上执行整句翻译。
  - **Meaning Unit**：按照源语义是否完整，划分片段



- Learning Adaptive Segmentation Policy for Simultaneous Translation. *EMNLP 2020*.
- Translation-based Supervision for Policy Generation in Simultaneous Neural Machine Translation. *EMNLP 2021*.

增强的同传策略：单一策略能力过弱

- 对于自适应策略：利用动态规划算法，计算翻译当前词之前采取的所有可能策略的期望。



- Cross Attention Augmented Transducer Networks for Simultaneous Translation. *EMNLP 2021*.
- A Generative Framework for Simultaneous Machine Translation. *EMNLP 2021*.

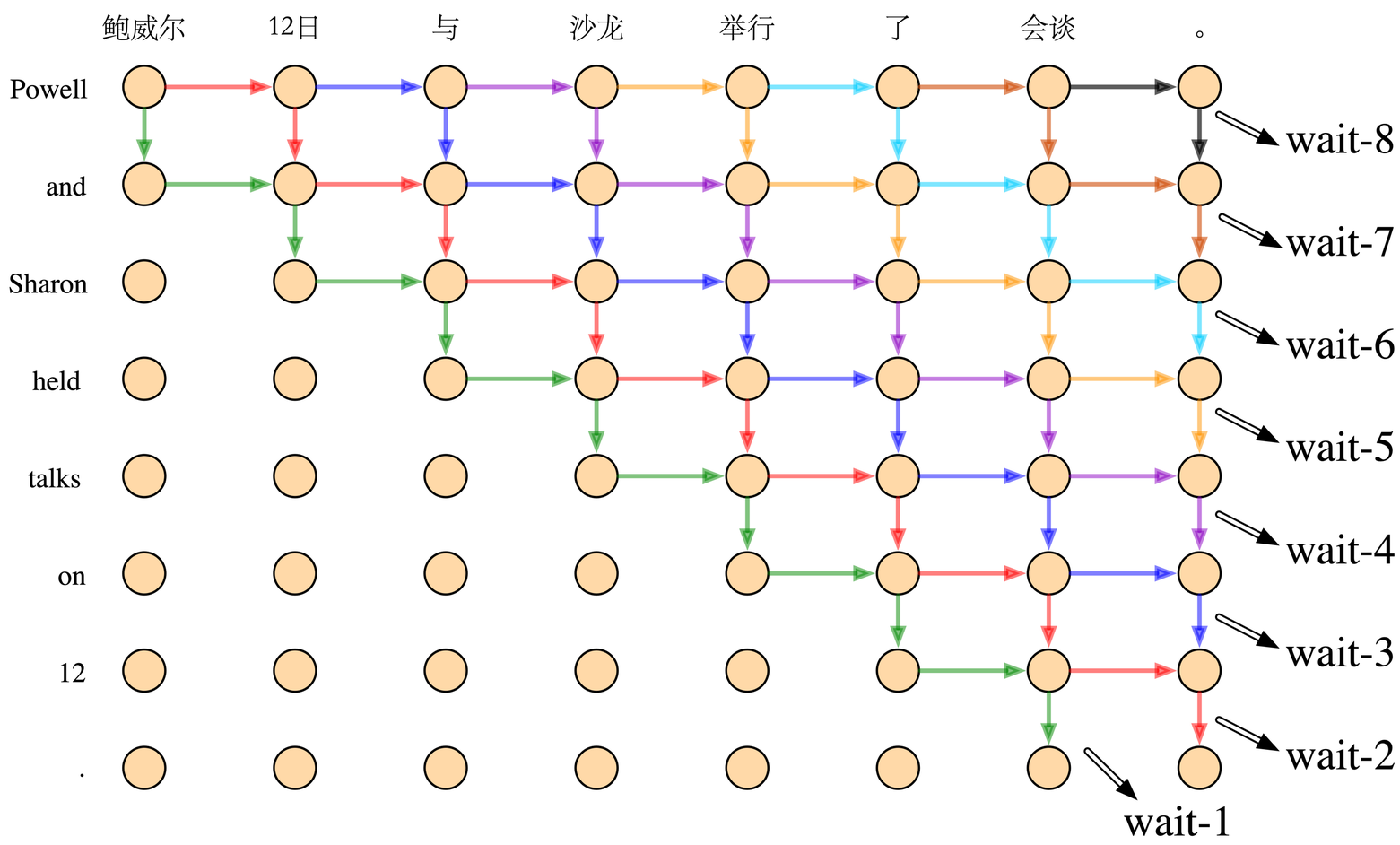
增强的同传策略：单一策略能力过弱

- 对于固定策略：利用混合专家模型，每个专家学习一种策略，多个策略共同决定最终翻译。

MoE Wait-k Policy

8个attention head分别学习

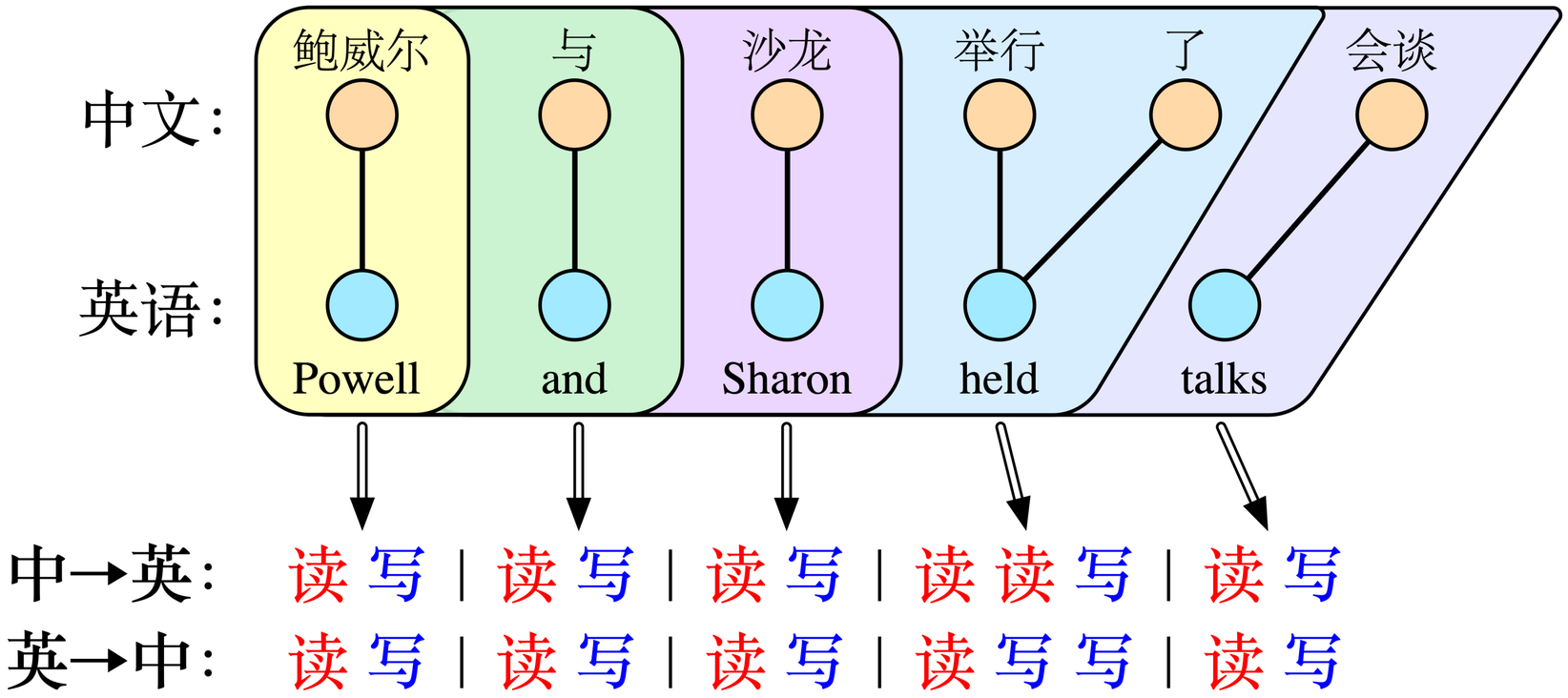
wait-1, wait-2, …, wait-8



- Universal Simultaneous Machine Translation with Mixture-of-Experts Wait-k Policy. *EMNLP 2021*.
- Efficient Wait-k Models for Simultaneous Machine Translation. *InterSpeech 2020*.

增强的同传策略：双向策略互相监督

✤ **Dual-Path**：基于同传策略在两个方向间对偶性约束，构造监督信号。



同传策略基于片段对的划分

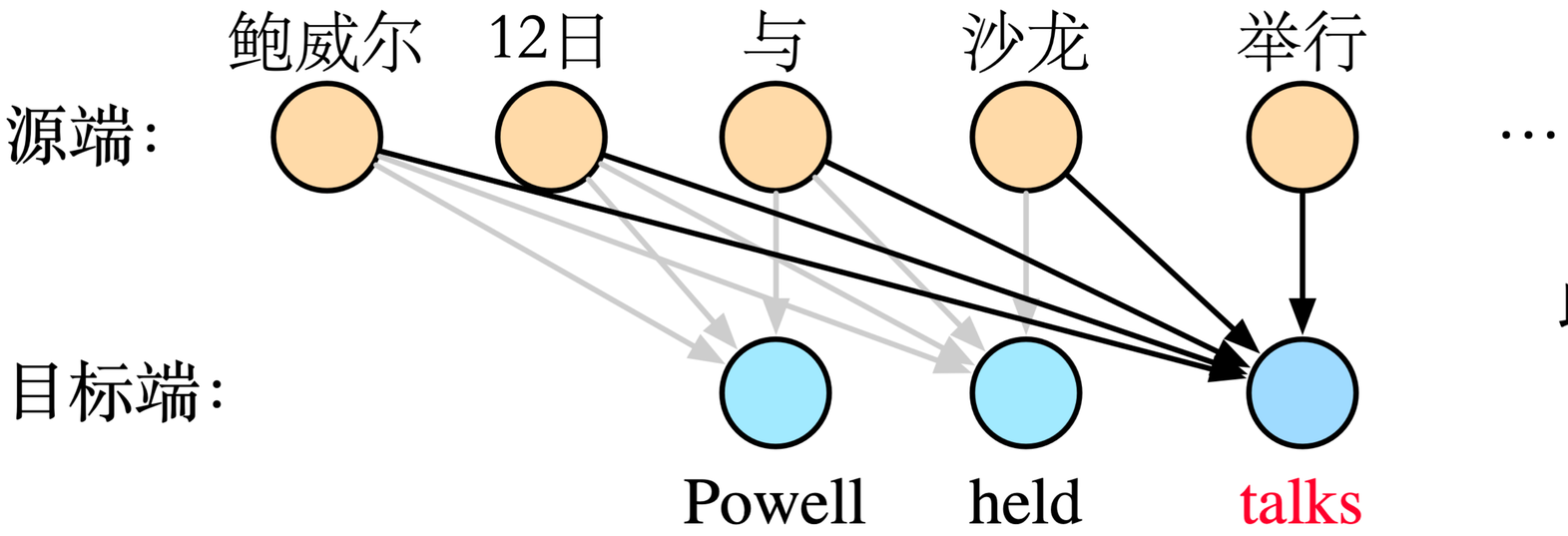
⇓

两个方向间的策略满足对偶形式

• Modeling Dual Read/Write Paths for Simultaneous Machine Translation. *ACL 2022*.

输入信息受限

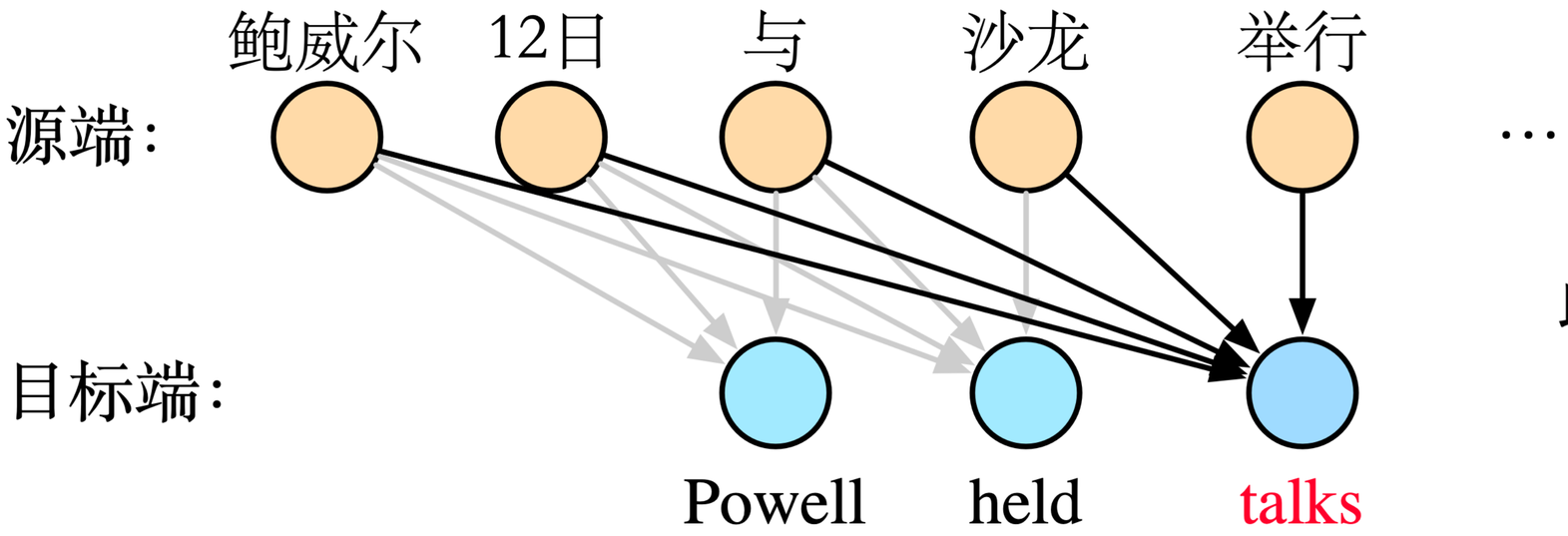
- 模型需要**基于不完整源端进行翻译**的能力



此时“talks”对应的源端词“会谈”还未读入

输入信息受限

- 模型需要**基于不完整源端进行翻译**的能力



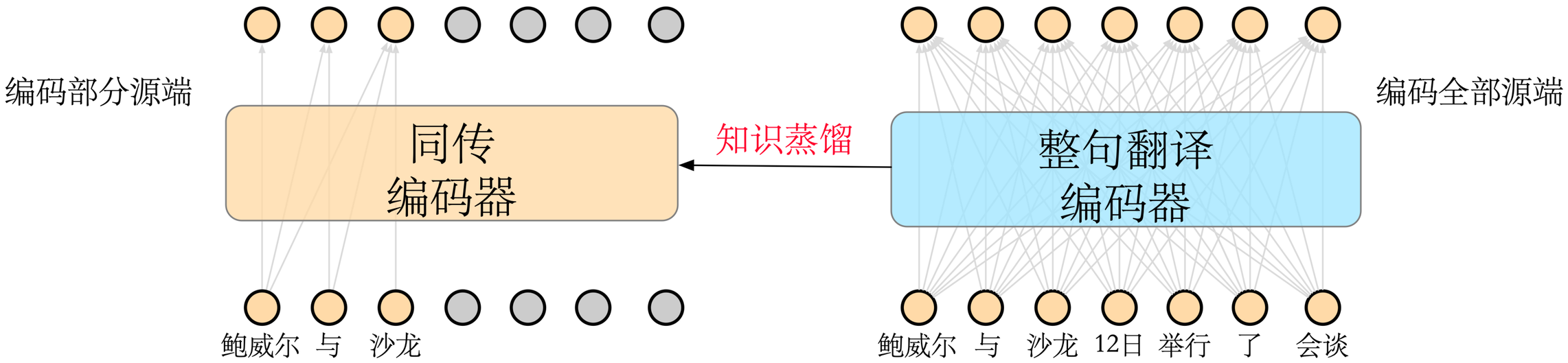
此时“talks”对应的源端词“会谈”还未读入

需要引入额外信息

隐式 or 显式

输入信息受限：引入未来信息

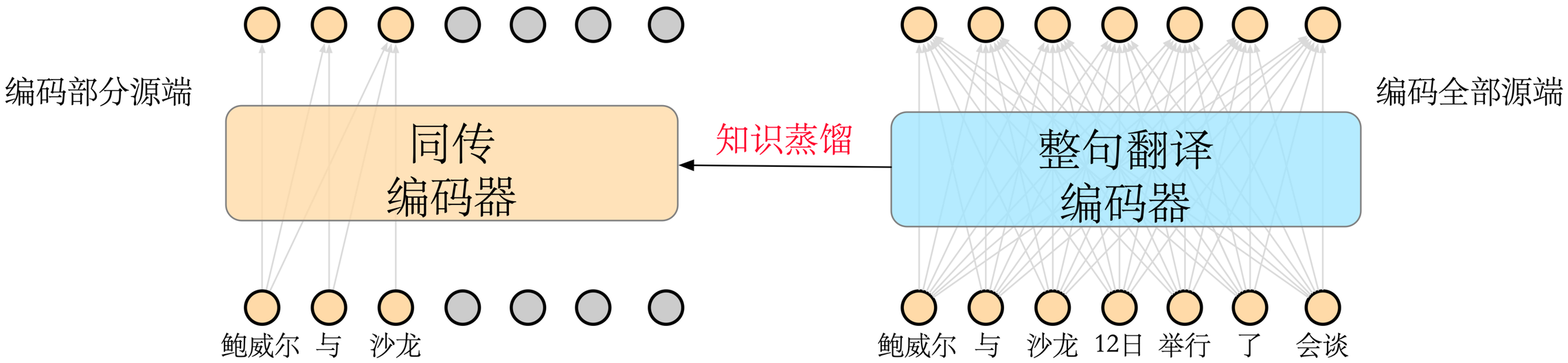
✤ **Future-Guided**：利用整句翻译指导同传编码器训练。



• Future-Guided Incremental Transformer for Simultaneous Translation. *AAAI 2021*.

输入信息受限：引入未来信息

✤ **Future-Guided**：利用整句翻译指导同传编码器训练。

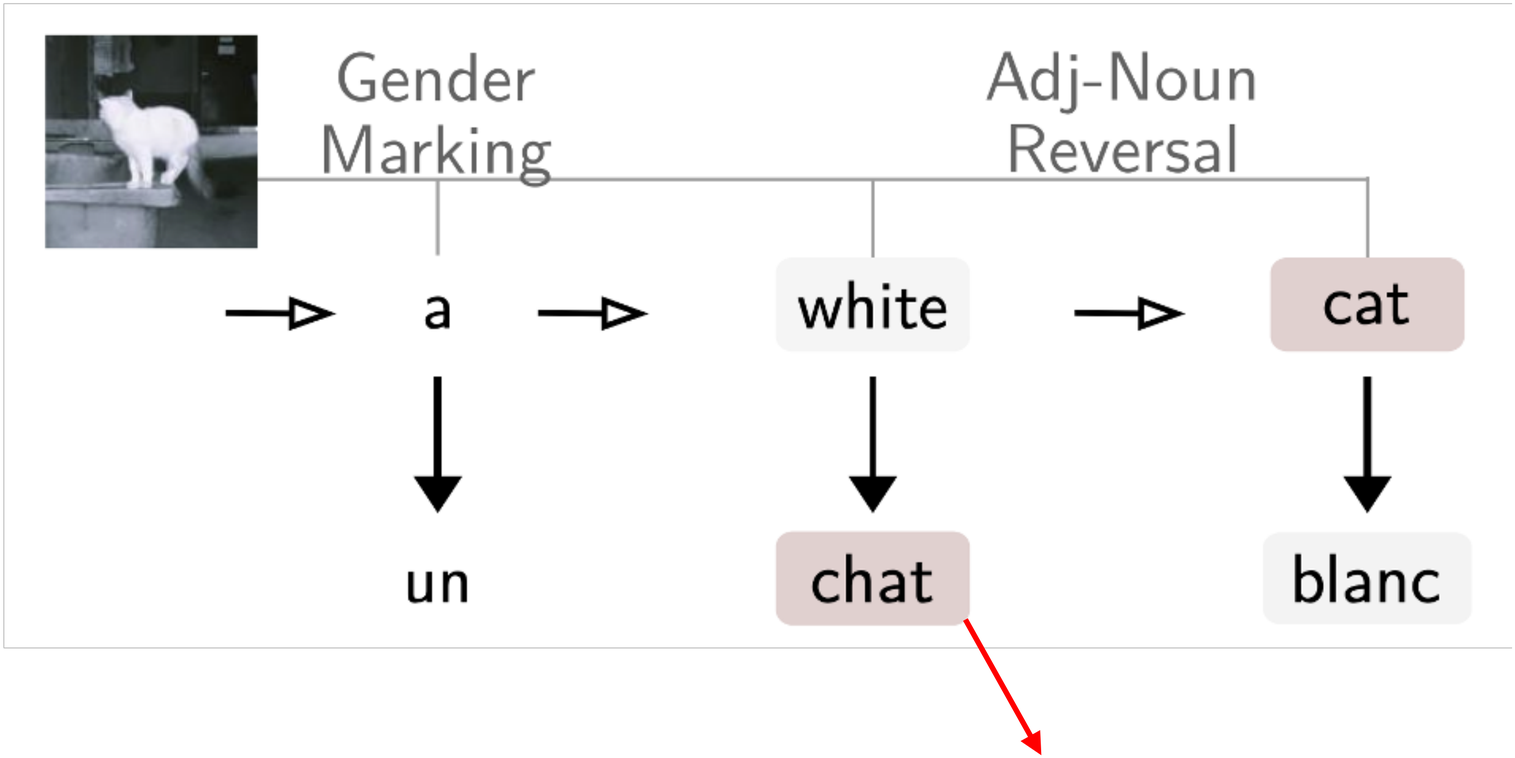


隐式嵌入未来信息，在输入信息不完整时建模全局信息。

• Future-Guided Incremental Transformer for Simultaneous Translation. *AAAI 2021*.

输入信息受限：引入多模态信息

- 同声传译+**图片翻译**：

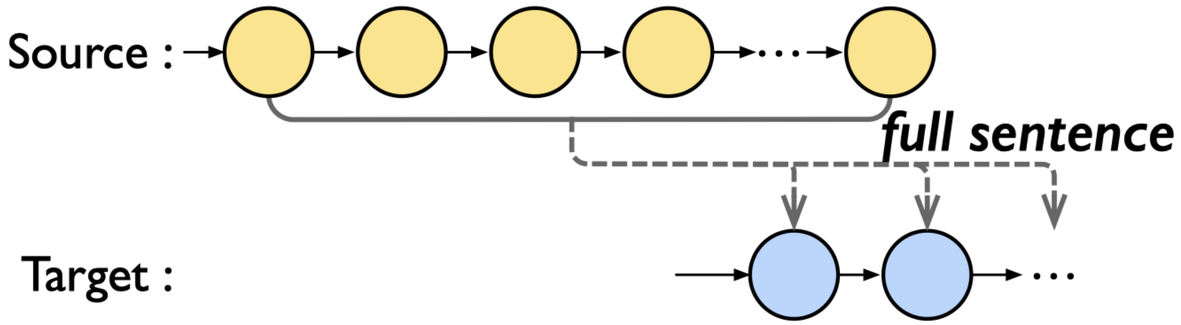


在源端信息受限的情况下，图片信息能提供有效的补充

- Simultaneous Machine Translation with Visual Context. *EMNLP 2020*.
- Exploiting Multimodal Reinforcement Learning for Simultaneous Machine Translation. *EACL 2021*.

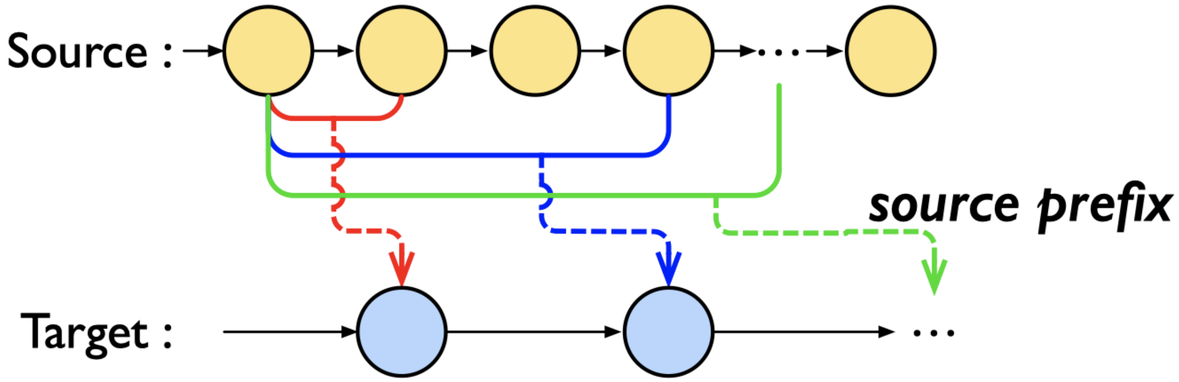
输入信息受限：增强全局规划

- **seq-to-seq框架 v.s. prefix-to-prefix框架：**



整句翻译　seq-to-seq

每个目标词都能关注完整的源句子



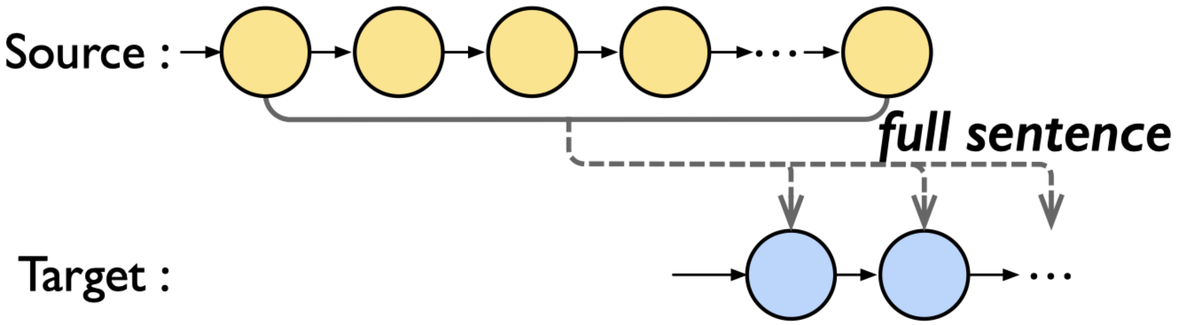
同声传译　prefix-to-prefix

强制每个目标词只能关注部分前缀

• Reducing Position Bias in Simultaneous Machine Translation with Length-Aware Framework. *ACL 2022*.

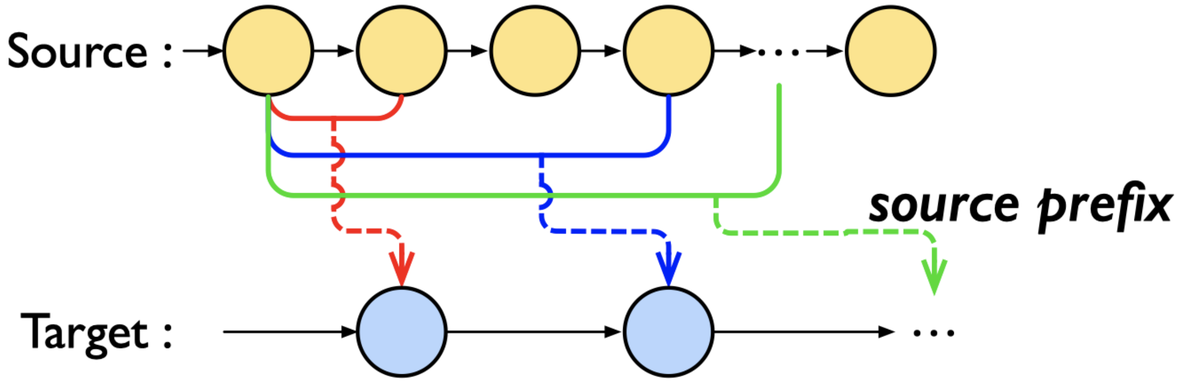
输入信息受限：增强全局规划

✤ **seq-to-seq框架 v.s. prefix-to-prefix框架：**



整句翻译　seq-to-seq

每个目标词都能关注完整的源句子



同声传译　prefix-to-prefix

强制每个目标词只能关注部分前缀

- Prefix无相关信息，也被强制关注
- 过于关注靠前词，缺乏全局规划

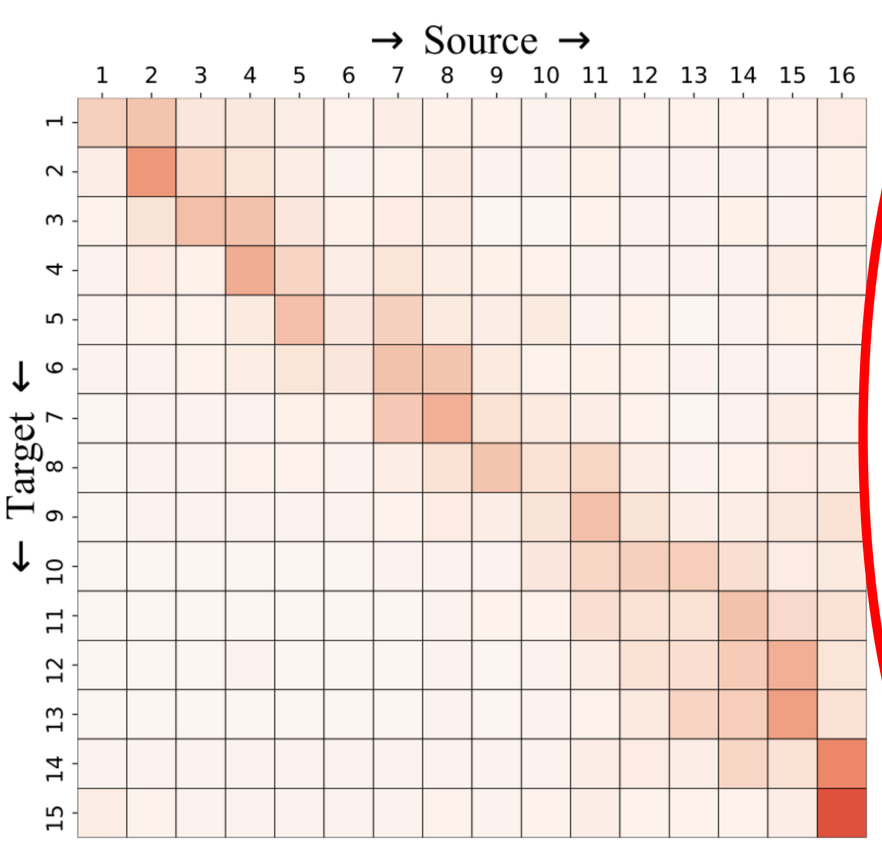
• Reducing Position Bias in Simultaneous Machine Translation with Length-Aware Framework. *ACL 2022*.

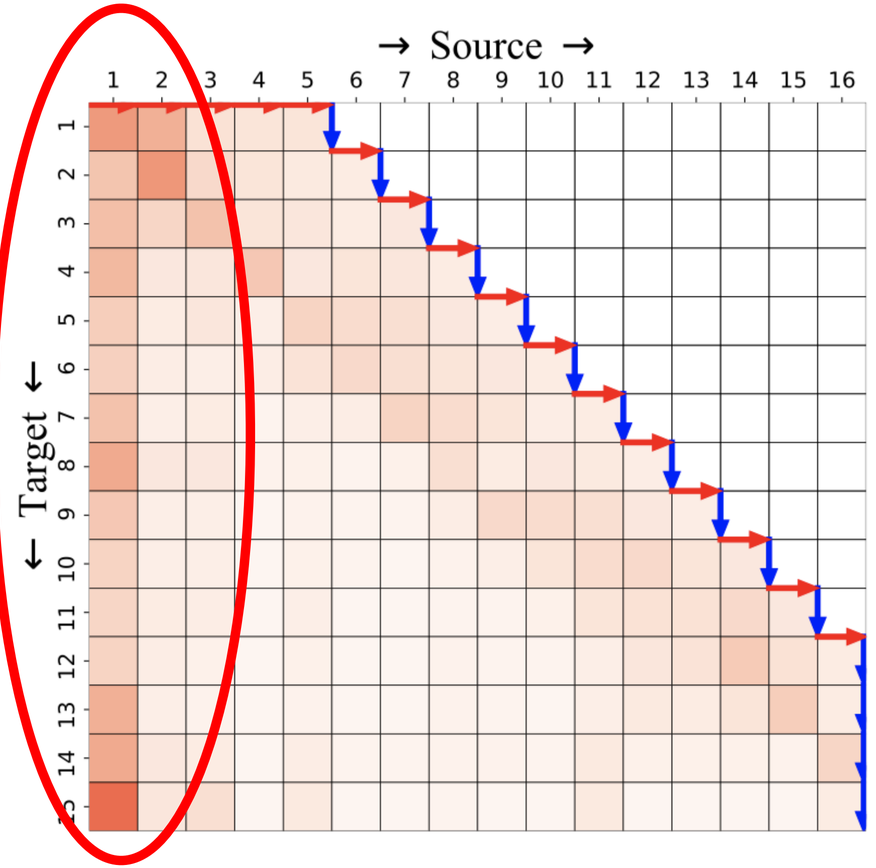
输入信息受限：增强全局规划

位置偏差（Position Bias）

- 靠前的源位置获得更多的attention权重

整句翻译





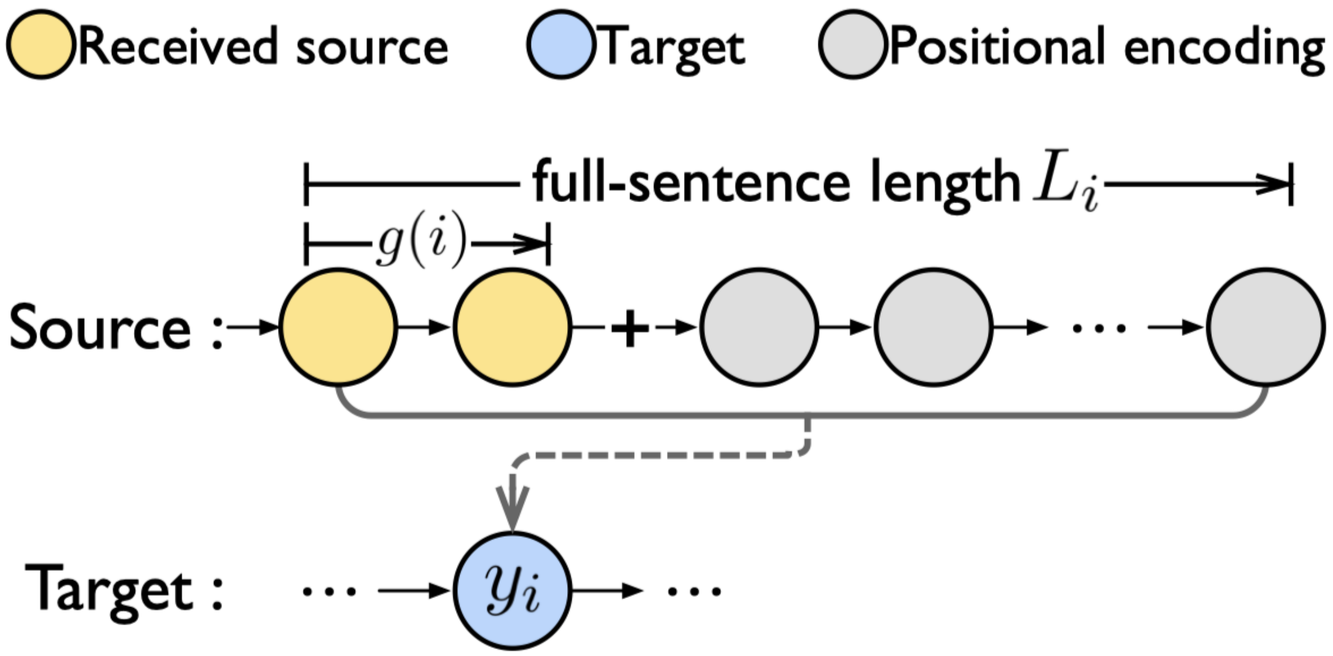
同声传译

• Reducing Position Bias in Simultaneous Machine Translation with Length-Aware Framework. *ACL 2022*.

输入信息受限：增强全局规划

✤ 长度感知框架（Length-Aware Framework）

❖ 构造伪整句：预测整句长度 ⟹ 用位置编码填充未来位置



提供全局规划

避免position bias

• Reducing Position Bias in Simultaneous Machine Translation with Length-Aware Framework. *ACL 2022*.

数据稀疏

- 现有常用数据集：
  - 整句文本⇒文本翻译数据集： WMT、IWSLT…
  - 整句语音⇒文本翻译数据集： MuST-C…
  - 同声传译数据集： NAIST-SIC 英-日、BSTC 中-英

数据稀疏

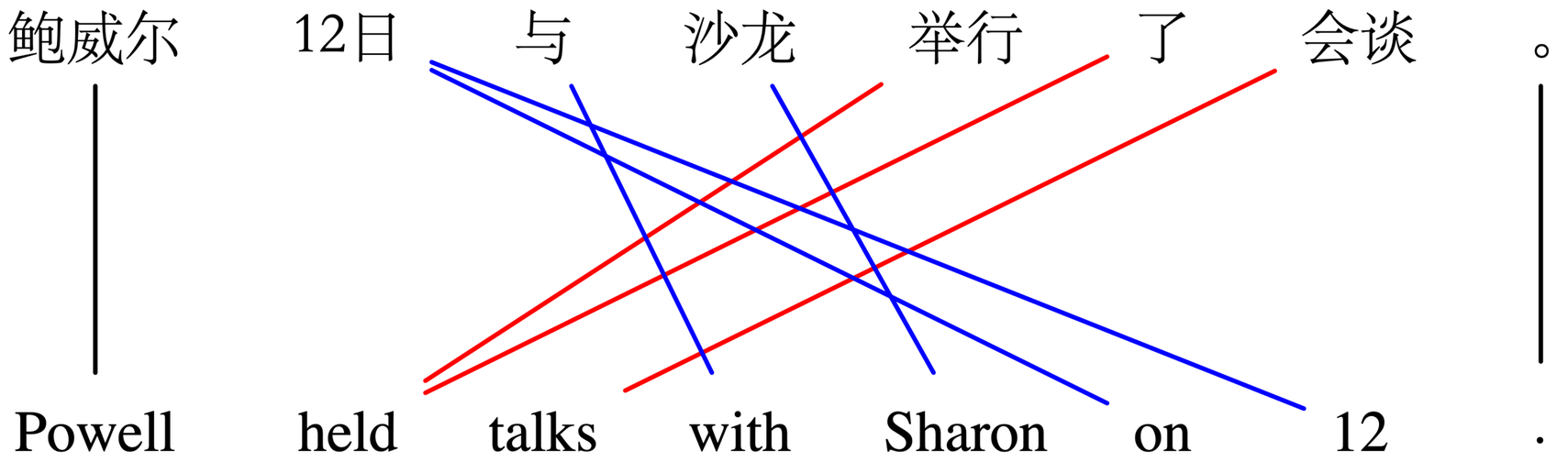
- 现有常用数据集：
  - 整句文本⇒文本翻译数据集： WMT、IWSLT…
  - 整句语音⇒文本翻译数据集： MuST-C…
  - 同声传译数据集： NAIST-SIC 英-日、BSTC 中-英

语言对少、数据量小

数据稀疏

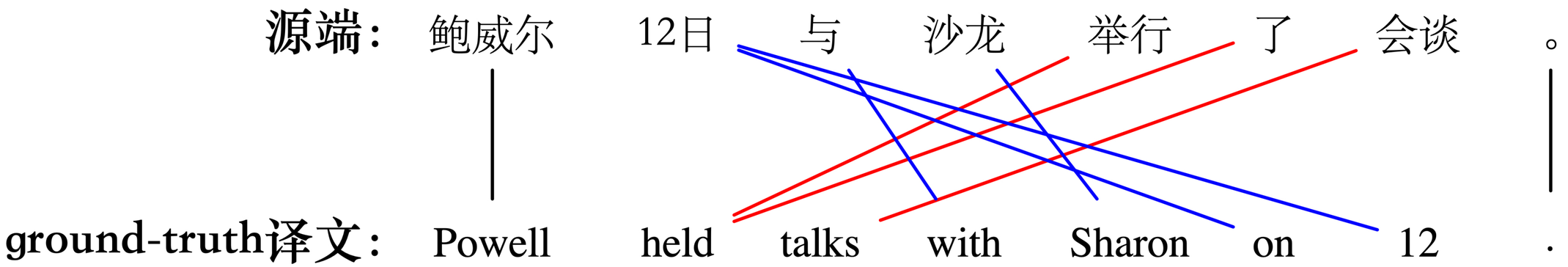
- 现有常用数据集：
  - 整句文本⇒文本翻译数据集： WMT、IWSLT…
  - 整句语音⇒文本翻译数据集： MuST-C…
  - 同声传译数据集： NAIST-SIC 英-日、BSTC 中-英

和同传数据存在领域差异，对模型的学习不利



增广数据：构造单调对齐译文

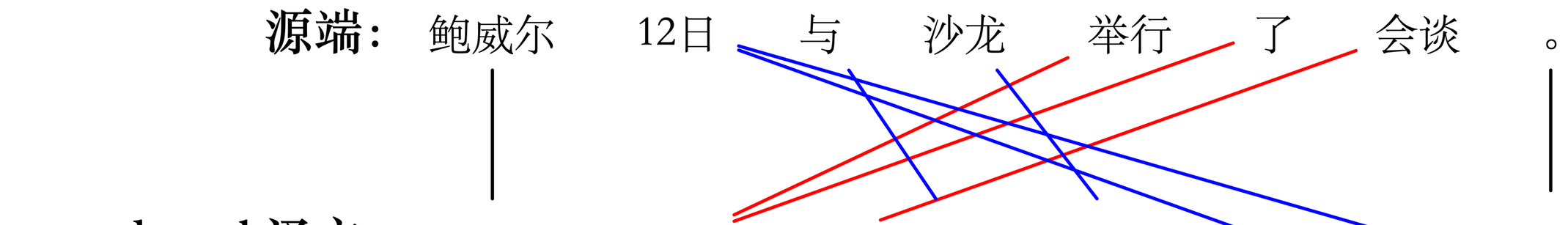
- 关注整句翻译数据和同传数据语序差异



• Improving Simultaneous Translation by Incorporating Pseudo-References with Fewer Reorderings. *EMNLP 2021*.

增广数据：构造单调对齐译文

- 关注整句翻译数据和同传数据语序差异



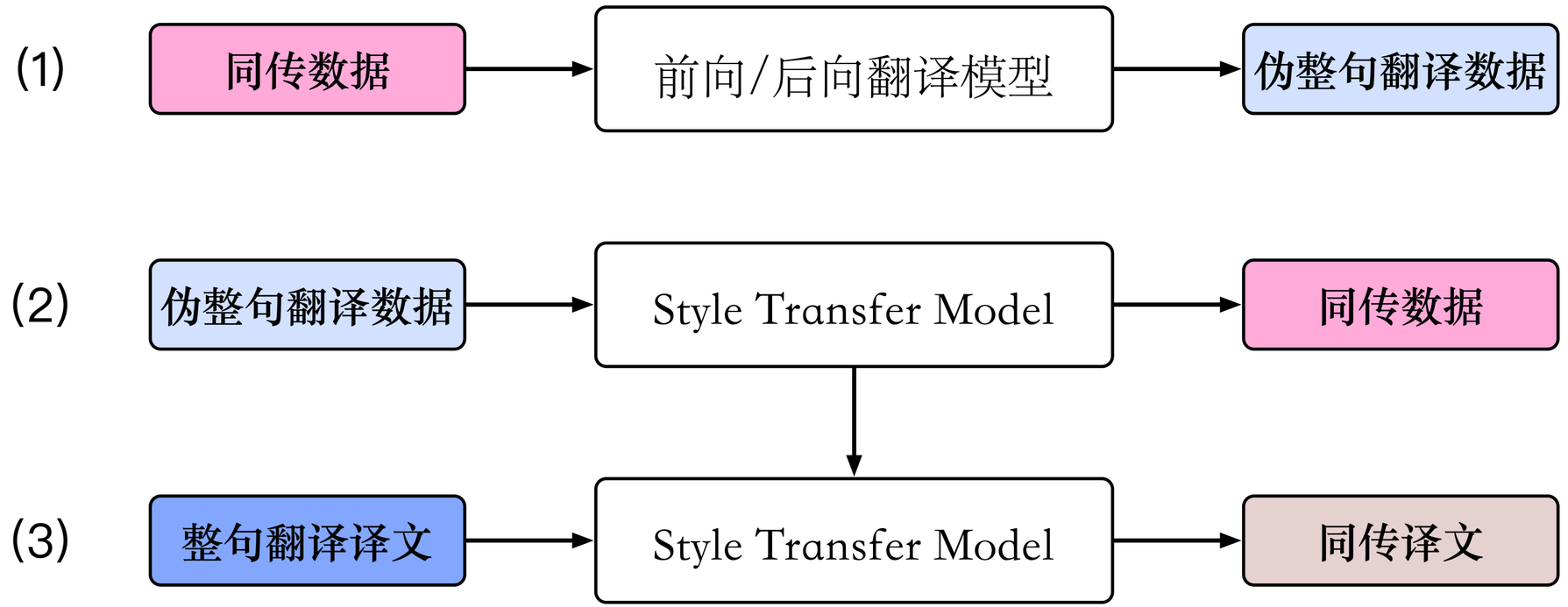
伪单调译文：

Powell	12	and	Sharon	held	talks	.
Powell	and	Sharon	held	talks	on	12 .
on	12	Powell	and	Sharon	held	talks .

• Improving Simultaneous Translation by Incorporating Pseudo-References with Fewer Reorderings. *EMNLP 2021*.

增广数据：生成同传风格译文

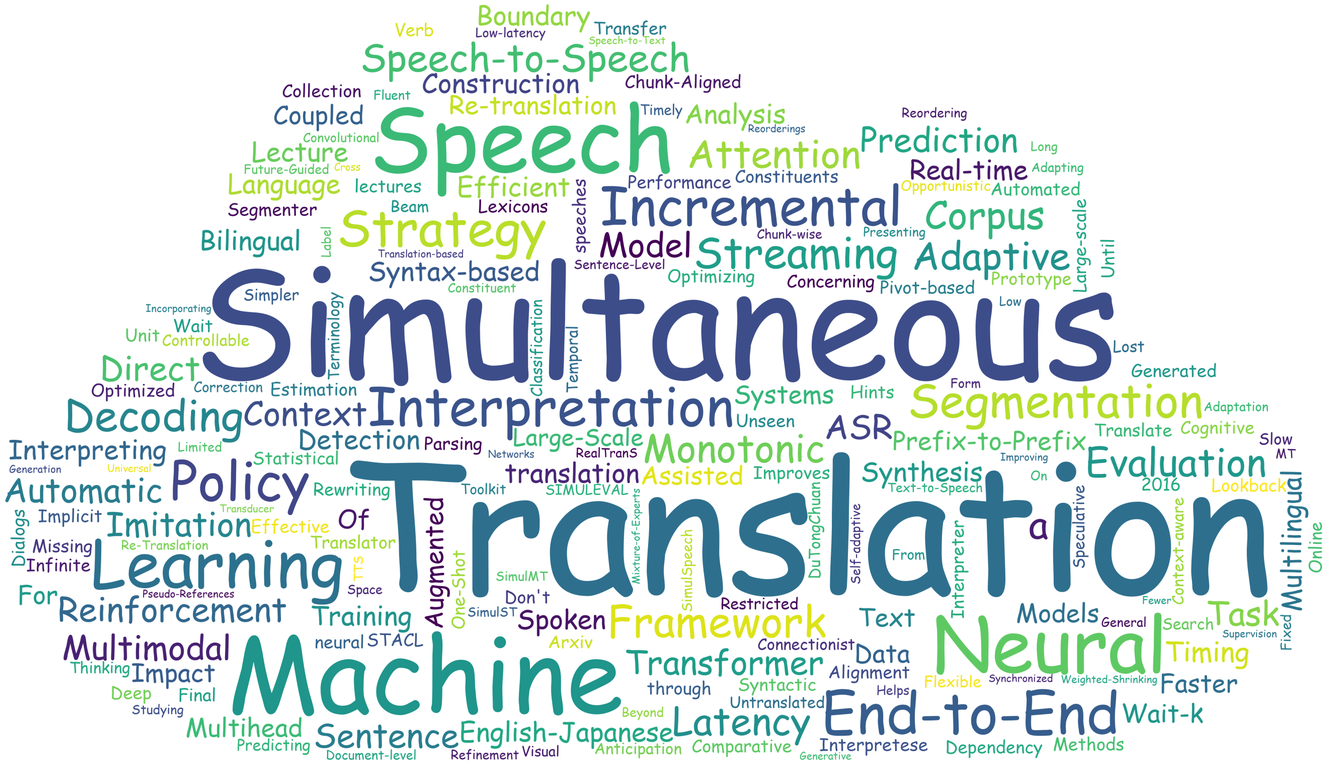
- 基于少量标注同传数据训练 **Style Transfer Model**
  - 大量整句翻译译文 ⇒ 同传译文



• It Is Not As Good As You Think! Evaluating Simultaneous Machine Translation on Interpretation Data. *EMNLP 2021*.

未来研究

- **精确的读/写策略：**
  - 可解释性、训练/测试匹配程度、模仿整句翻译
- **引入额外信息：**
  - 建模未来信息+全局规划
  - 引入多模态信息：语音、图片...
- **对同传友好的训练数据：**
  - 构造更单调的、更分段的译文
  - 收集同传数据集



Thanks!

Shaolei Zhang

Email: zhangshaolei20z@ict.ac.cn

Ref: github.com/Vily1998/Awesome-Simultaneous-Translation