# 机器翻译前沿综述

冯洋
中科院计算所

2019.10.20

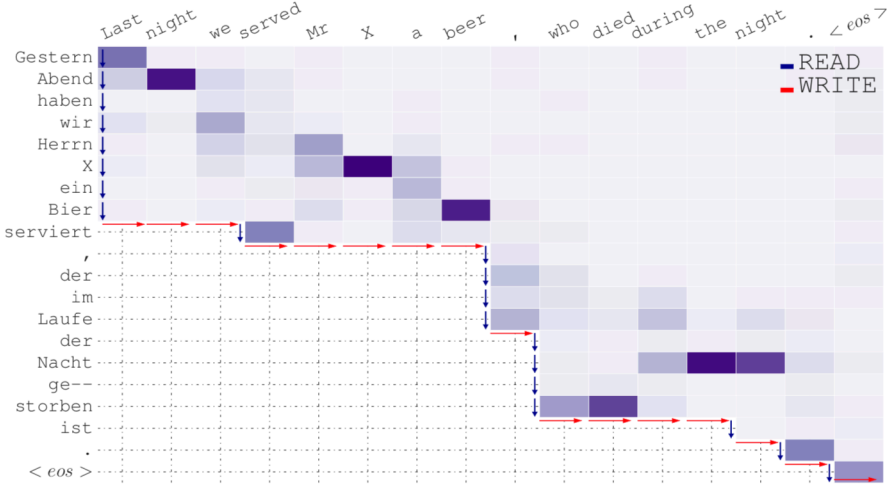
| ACL 2017 | ACL 2018 | ACL 2019 |
|---|---|---|
| • Translation | • Translation | • Neural |
| • Machine | • Machine | • Learning |
| • Neural | • Neural | • Language |
| • **Domain** | • Learning | • Translation |
| • Data | • Attention | • Machine |
| • **Adaptation** | • **Unsupervised** | • Word |
| • Decoding | • Multi | • Sentence |
| • Learning | • Context | • **Embedding** |
| • **Monolingual** | • Model | • Representation |
| • Source | • Evaluation | • Multi |
| • Attention | • **Document** | • **Unsupervised** |
| • Models | • **Non-autoregressive** | • Semantic |
| • Syntax | • Decoder | • Attention |
| • Context | • Search | • **Multilingual** |
| • System | | • **Adversarial** |
| | | • Context |

# 目录

CONTENTS

# 同声传译

- 场景：
  - 在读取源语言输入的同时完成翻译的输出，翻译过程仅有几个单词的延迟。



# 同声传译

- 要解决的问题：
  - 读取源语言输入的过程中，判断是否在当前位置进行输出。
  - SOV语言到SVO语言的翻译，需要读取源语言动词才能进行准确翻译。
  - 如何在训练时模拟测试时的增量环境。

# 同声传译

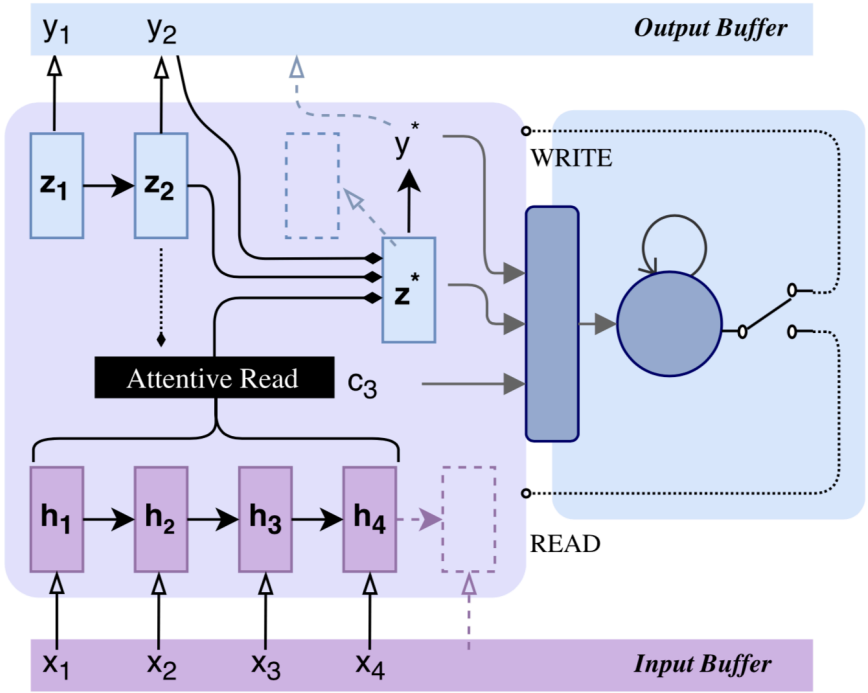
- 解决方案：
  - 更精确的读写策略。

# 同声传译

- 解决方案：
  - 更精确的读写策略。发展分为三个阶段：
    - 基于解码策略和强化学习

# 同声传译

- 解决方案
  - 更精确
  - 基于




段：

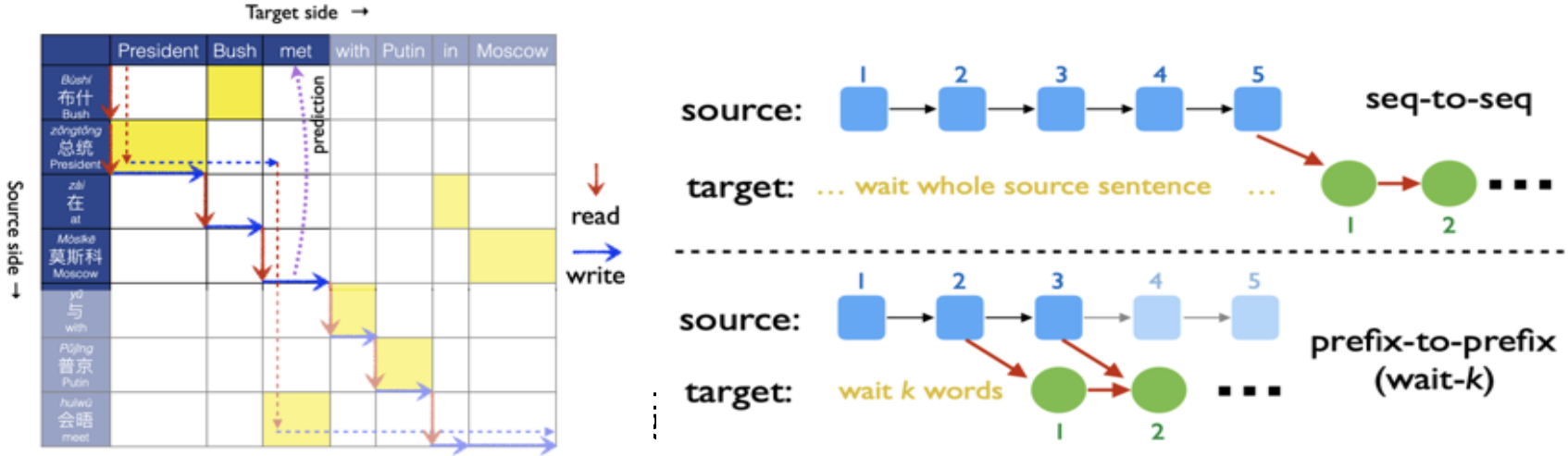
每个输入位置都对应读、写两个动作，为每个动作都定义了reward，通过强化学习来使得NMT系统选择reward最大的动作序列。

Jiatao Gu, Graham Neubig, Kyunghyun Cho, and Victor O. K. Li. 2017. Learning to translate in real-time with neural machine translation. *In Proceedings of EACL 2017.*

# 同声传译

- 解决方案：
  - 更精确的读写策略。发展分为三个阶段：
    - 基于解码策略和强化学习
    - 固定延迟的读写策略，如 wait-k

# 同声传译




等待输入k个源端单词，然后再进行翻译，即输出始终落后输入k个词。训练时，通过mask源端单词来模拟增量翻译。

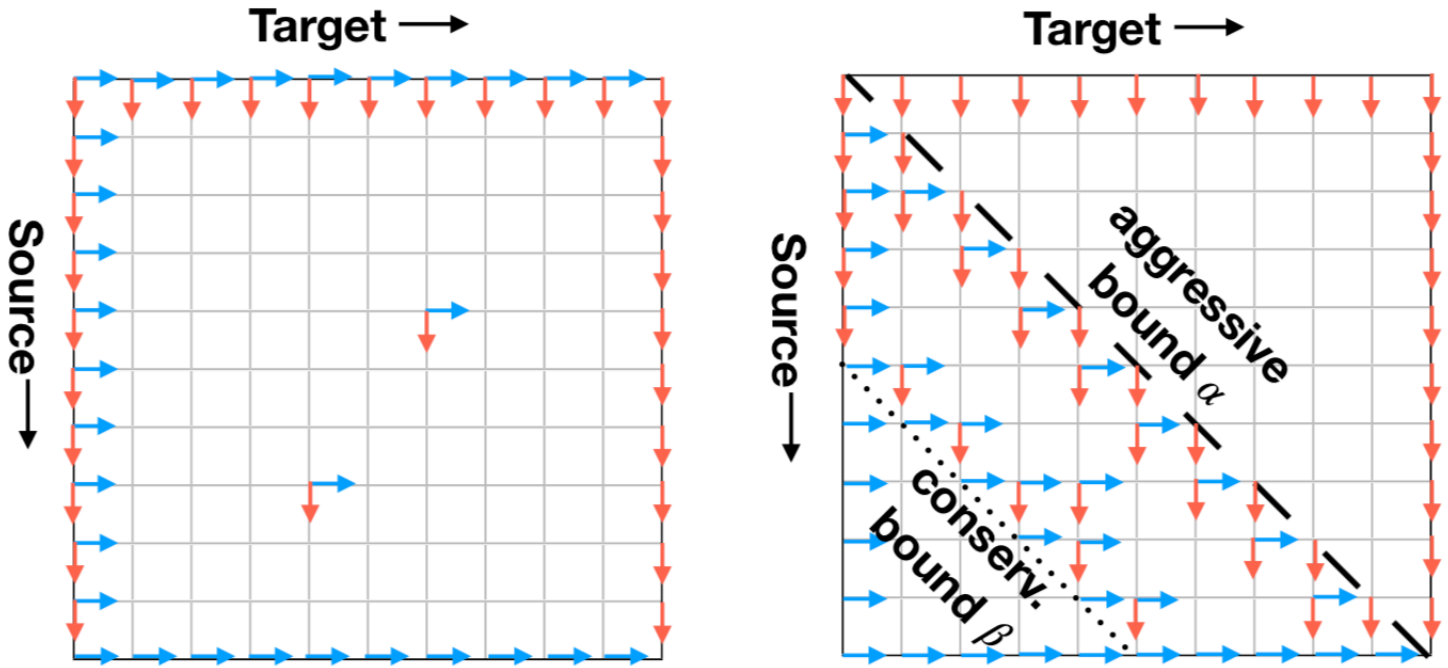
Mingbo Ma, Liang Huang, Hao Xiong, Kaibo Liu, Chuanqiang Zhang, Zhongjun He, Hairong Liu, Xing Li, and Haifeng Wang. 2018. STACL: Simultaneous Translation with Integrated Anticipation and Controllable Latency.

# 同声传译

- 解决方案：
  - 更精确的读写策略。发展分为三个阶段：
    - 策略解码
    - 固定延迟的读写策略，如 wait-k
    - 自适应的读写策略

# 同声传译

- 解
  -




Baigong Zheng, Renjie Zheng, Mingbo Ma, Liang Huang. 2019. Simpler and Faster Learning of Adaptive Policies for Simultaneous Translation. *In Proceedings of EMNLP 2019*.

# 同声传译

- 解决方案：
  - 更精确的读写策略。
  - 增加模型预测能力。
    - 基于语言模型

# 多模态翻译

- 场景：
  - 输入为文本、语音、图像、视频中等，经过机器翻译，输出为文字。

# 多模态机器翻译

- 发展历程：
  - 语音翻译

# 多模态机器翻译

- 发展历程：
  - 语音翻译
    - 管道模型

# 多模态机器翻译

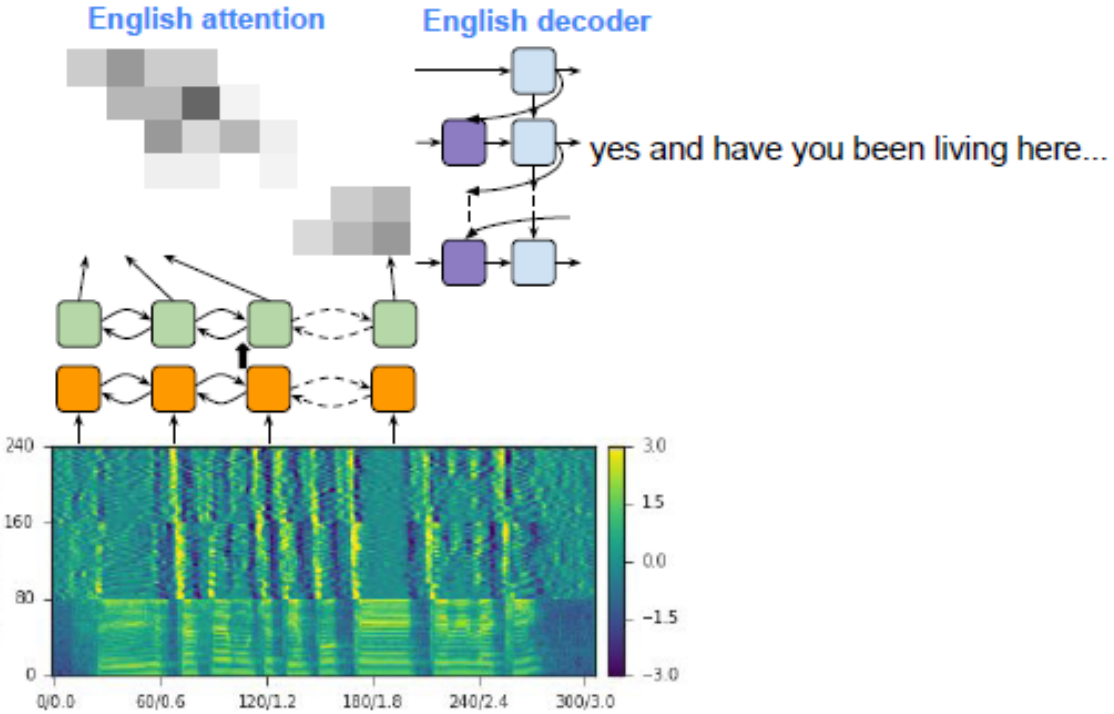
- 发展历程：
  - 语音翻译
    - 管道模型

# 多模态机器翻译

- 发展历程：
  - 语音翻译
    - 管道模型
    - 端到端模型

# 多模态机器翻译

- 发展历程
  - 语音翻
    - 管道
    - 端到




Ron J. Weiss, Jan Chorowski1, Navdeep Jaitly, Yonghui Wu and Zhifeng Chen. 2017. Sequence-to-Sequence Models Can Directly Translate Foreign Speech. *In Proceedings of INTERSPEECH 2017.*

# 多模态机器翻译

- 发展历程：
  - 语音翻译
    - 管道模型
    - 端到端模型
    - 二者融合，双注意力模型

# 多模态机器翻译

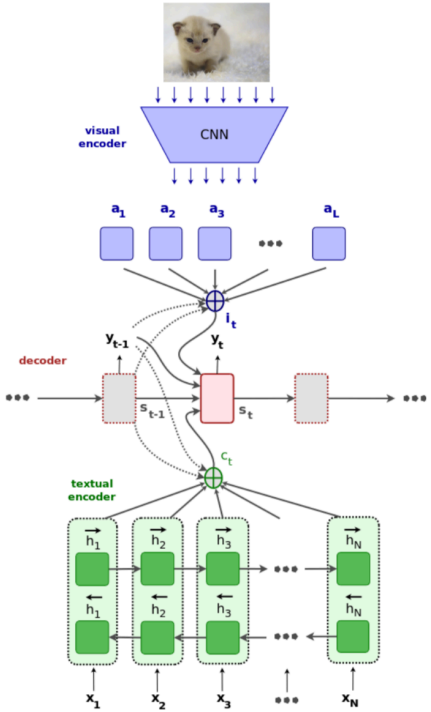
- 发展历程：
  - 语音翻译
  - 多模态翻译

# 多模态机器翻译

- 发展历程：
  - 语音翻译
  - 多模态翻译
    - 直接使用图像信息

# 多模态机器翻译

- 发展历程：
  - 语音翻译
  - 多模态翻译
    - 直接使用图像




Iacer Calixto, Qun Liu, and Nick Campbell. Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In Proceedings of ACL 2017.

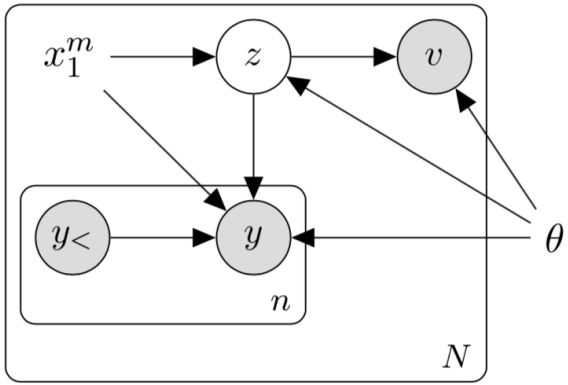
# 多模态机器翻译

- 发展历程：
  - 语音翻译
  - 多模态翻译
    - 直接使用图像信息
    - 使用Caption信息

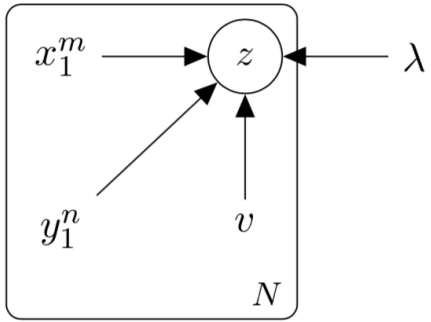
# 多模态机器翻译

- 发展历程：
  - 语音翻译
  - 多模态翻译
    - 直接使用图像信息
    - 使用Caption信息
    - 使用图像语义

# 多模态机器翻译




(a) VMMT$_C$: given the source text $x_1^m$, we model the joint likelihood of the translation $y_1^n$, the image (features) $v$, and a stochastic embedding $z$ sampled from a conditional latent Gaussian model. Note that the stochastic embedding is the sole responsible for assigning a probability to the observation $v$, and it helps assign a probability to the translation.




(b) Inference model for VMMT$_C$: to approximate the true posterior we have access to both modalities (text $x_1^m$, $y_1^n$ and image $v$).

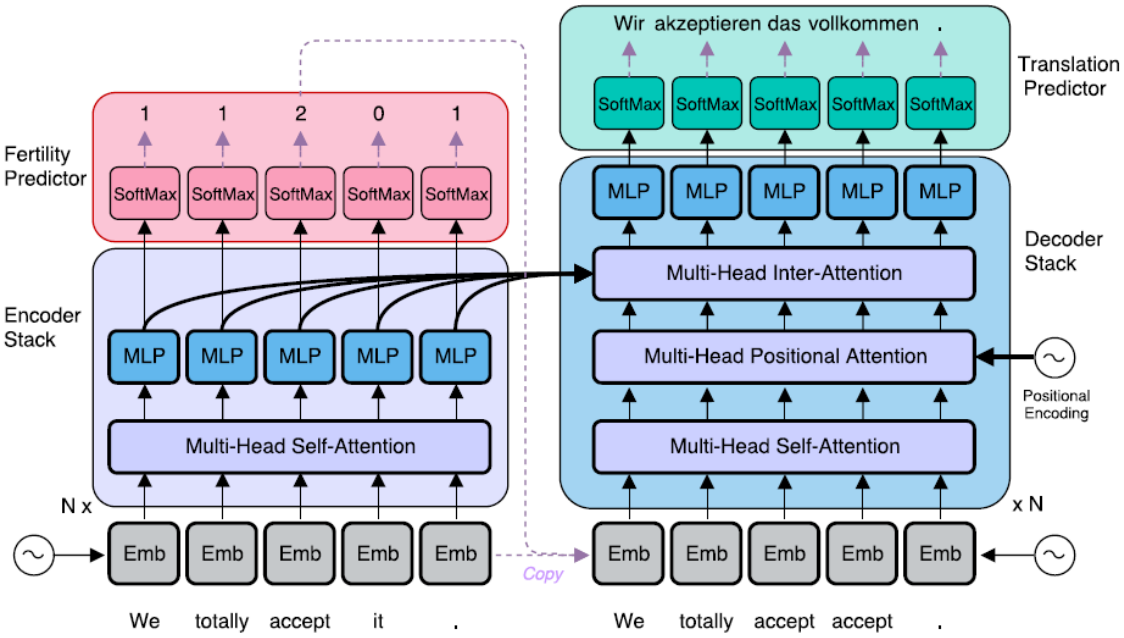
Iacer Calixto Miguel Rios Wilker Aziz. Latent Visual Clues for Neural Machine Translation. In proceedings of ACL 2019.

# 非自回归模型

- 场景：
  - 所有目标词同时生成，无需将前一个生成词作为输入

# 非自回归模型

## 通用模型：



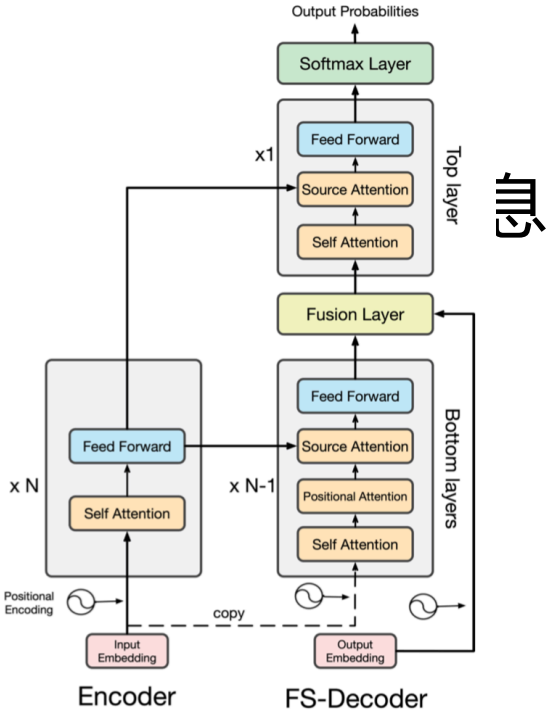

Jiatao Gu, James Bradbury, Caiming Xiong, Victor OK Li, and Richard Socher. 2017. Non- autoregressive neural machine translation.

# 非自回归模型

- 发展趋势：添加序列信息
  - 方式一：在模型中添加序列信息

# 非自回归模型

- 发展趋势：添加
  - 方式一：在模 息




Chenze Shao, Yang Feng, Qun Liu. Retrieving Sequential Information for Non-Autoregressive Neural Machine Translation. In Proceedings of ACL 2019.

# 非自回归模型

- 发展趋势：添加序列信息
  - 方式一：在模型中添加序列信息
  - 方式二：序列级训练，如强化学习

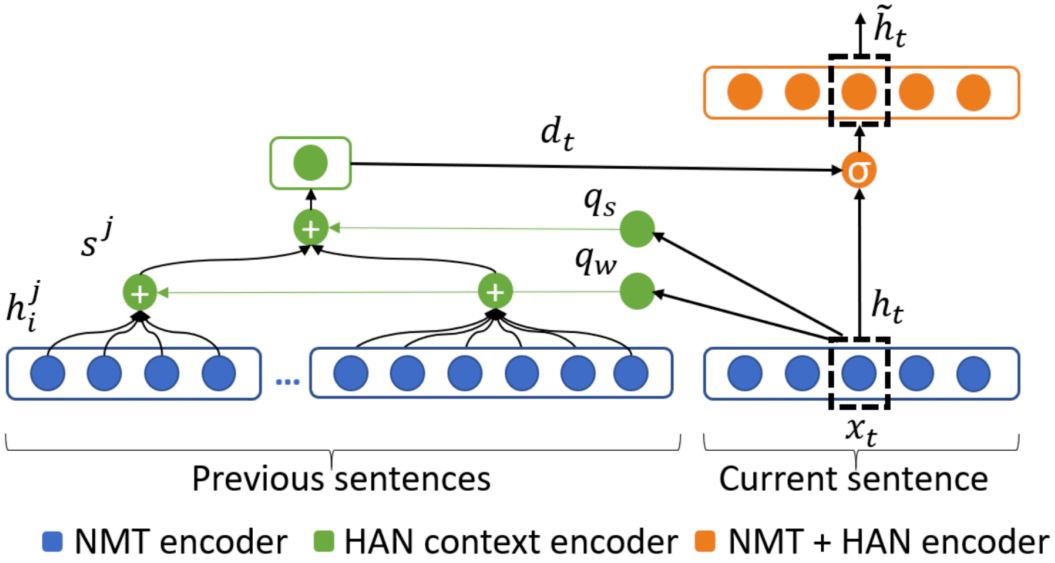
Chenze Shao, Yang Feng, Qun Liu. Retrieving Sequential Information for Non-Autoregressive Neural Machine Translation. In Proceedings of ACL 2019.

# 篇章翻译

- 场景：
  - 对一个段落进行翻译，需考虑上下文的一致性以及指代问题

# 篇章翻译

- 通用模型：



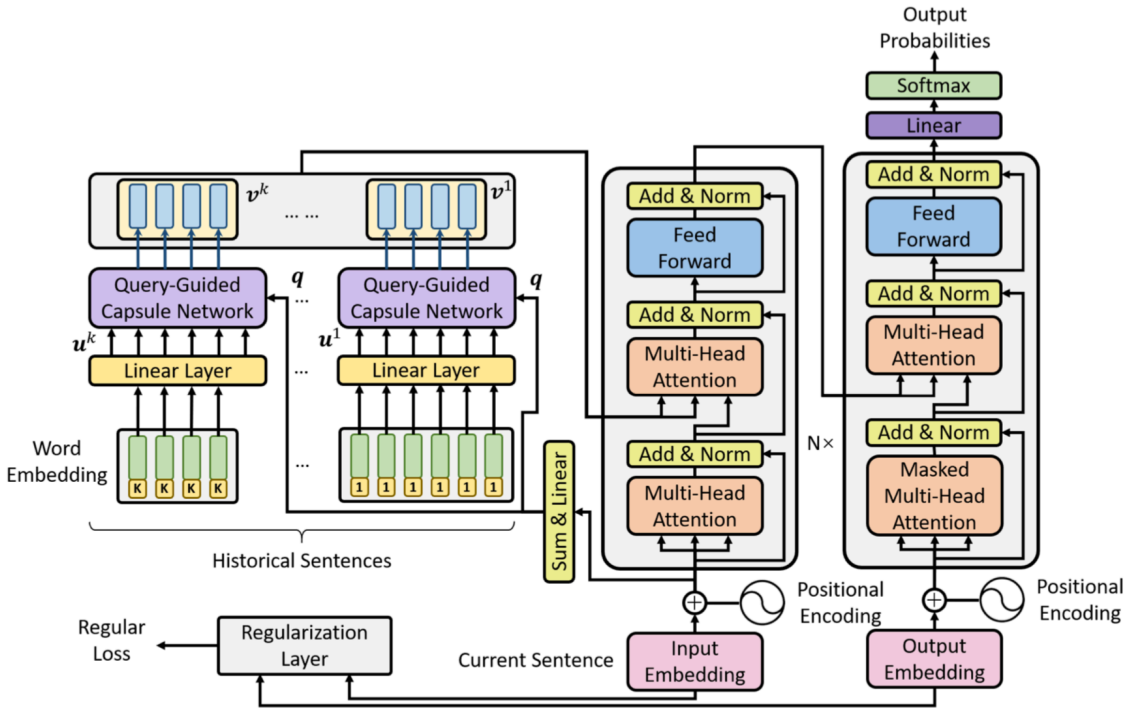

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. Document-level neural machine translation with hierarchical attention networks. In Proceedings of EMNLP 2018.

# 篇章翻译

- 发展趋势：
  - 改进注意力机制

# 篇章翻译

- 发展趋势
  - 改进注




Zhengxin Yang, Jinchao Zhang, Fandong Meng, Shuhao Gu, Yang Feng, Jie Zhou. Enhancing Context Modeling with a Query-Guided Capsule Network for Document-level Translation. In Proceedings of EMNLP 2019.

# 篇章翻译

- 发展趋势：
  - 改进注意力机制
  - 学习更好的表示 (参照多轮对话)

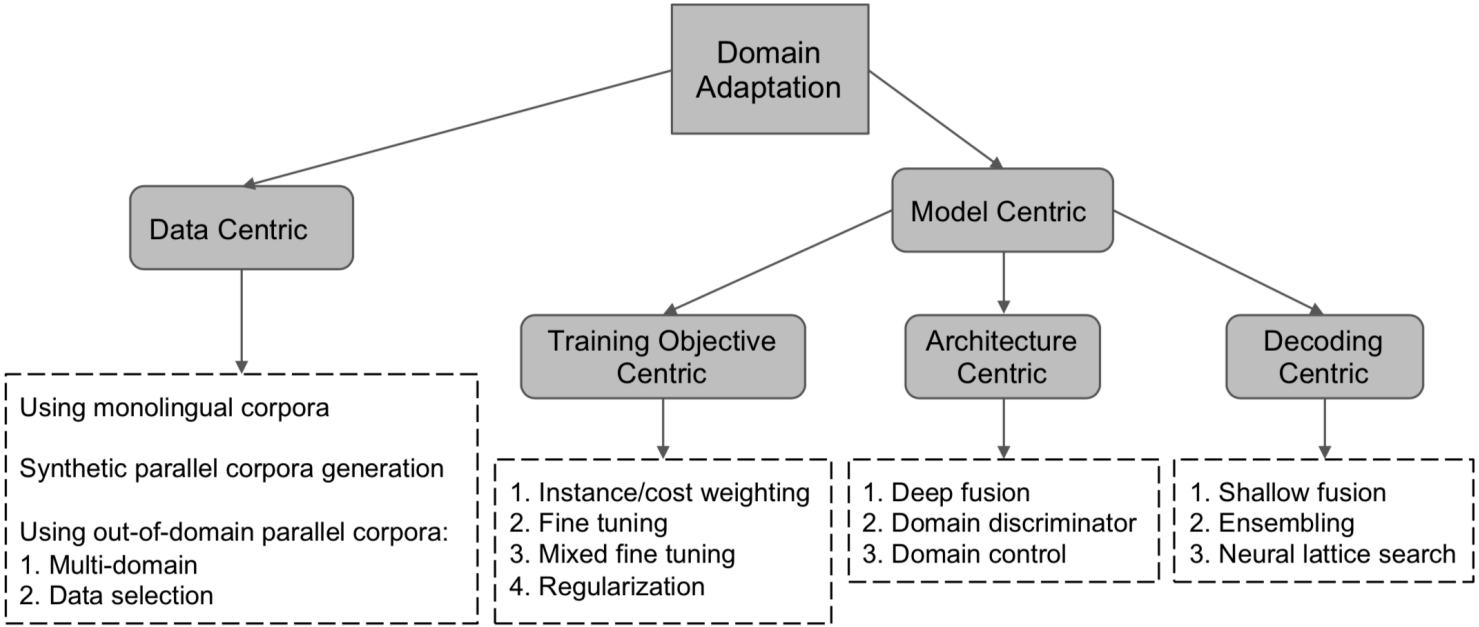
# 篇章翻译

- 发展趋势：
  - 改进注意力机制
  - 学习更好的表示
  - 使用更长的历史

# 领域自适应

- 场景：
  - 大量的out-of-domain或者通用领域数据，少量的in-domain数据，需要对in-domain的句子进行翻译。

# 领域自适应

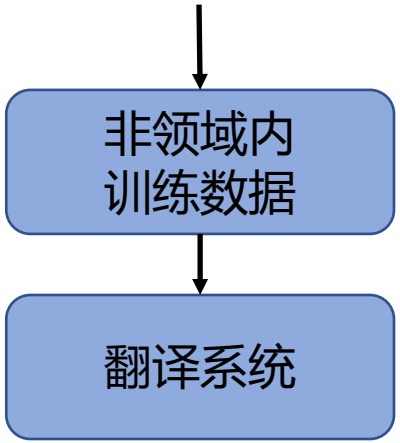



Chenhui Chu and Rui Wang. 2018. A Survey of Domain Adaptation for Neural Machine Translation. *In Proceedings of COLING 2018*.

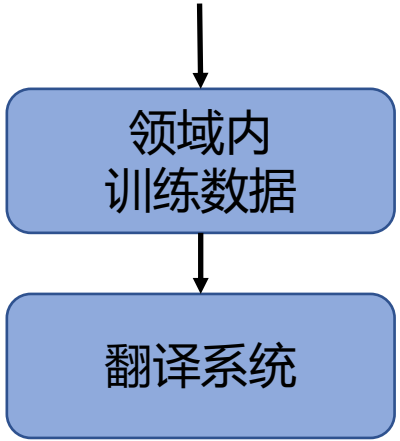
# 领域自适应

- 通用模型：Fine tune

第一阶段：




第二阶段：



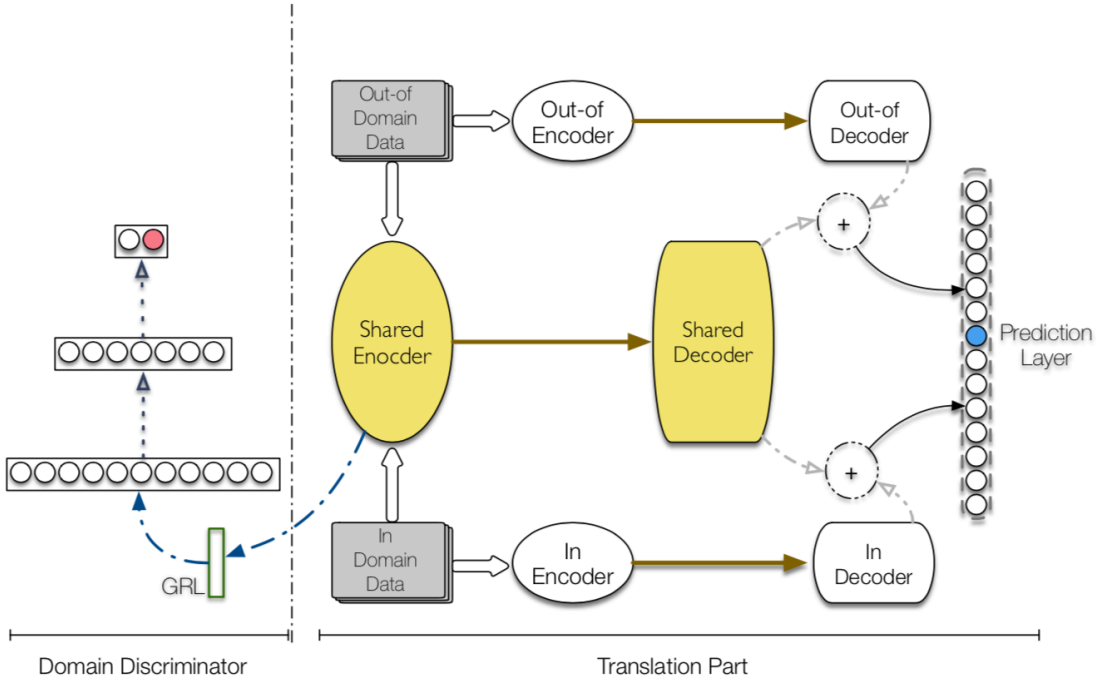

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. In Proceedings of the International Workshop on Spoken Language Translation.

# 领域自适应

- 发展趋势：
  - 采用对抗学习抽取特征

# 领域自适应

- 发展趋
  - 采用




Shuhao Gu, Yang Feng, Qun Liu. Improving Domain Adaptation Translation with Domain Invariant and Specific Information. *In Proceedings of NAACL 2019*.

# 领域自适应

- 发展趋势：
  - 采用对抗学习抽取特征
  - 无监督领域自适应

# 领域自适应

- 发展趋势：
  - 采用对抗学习抽取特征
  - 无监督领域自适应
  - 课程学习进行数据选择

# 多语言翻译

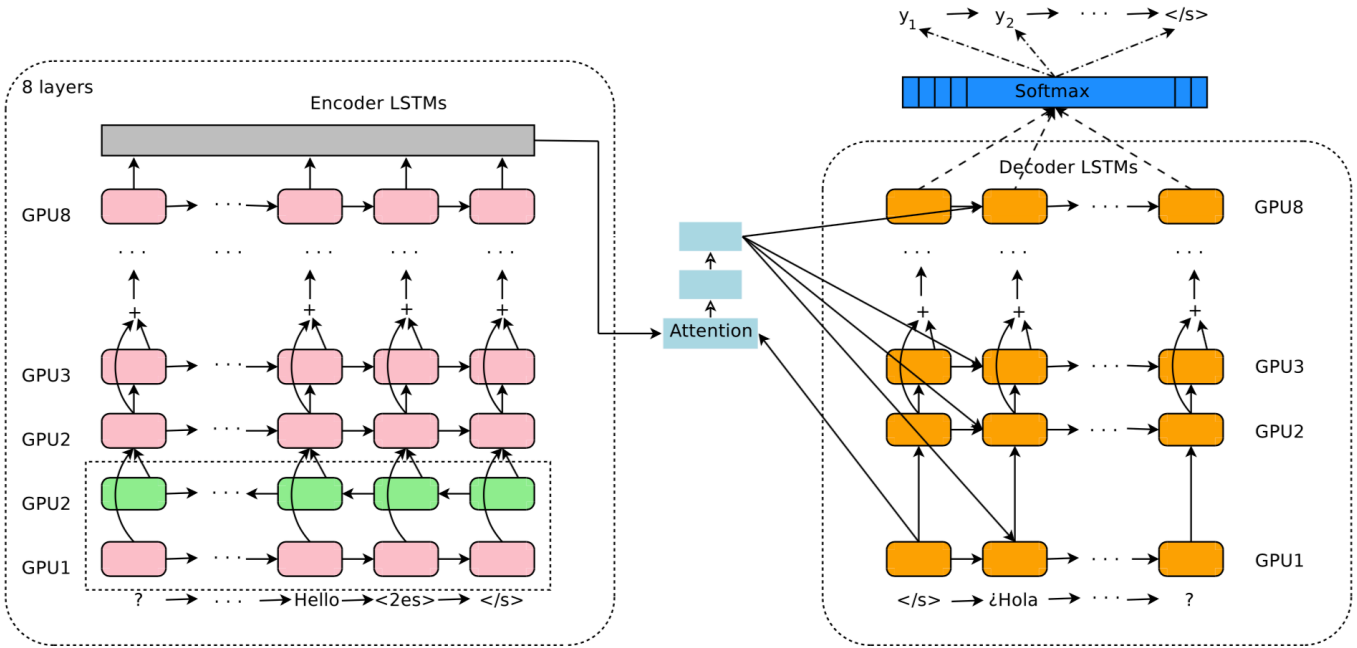
- 场景：
  - 一对多
  - 多对一
  - 多对多

# 多语言翻译

- 发展趋势：
  - 共享结构

# 多语言翻译

- 发展
  - 共



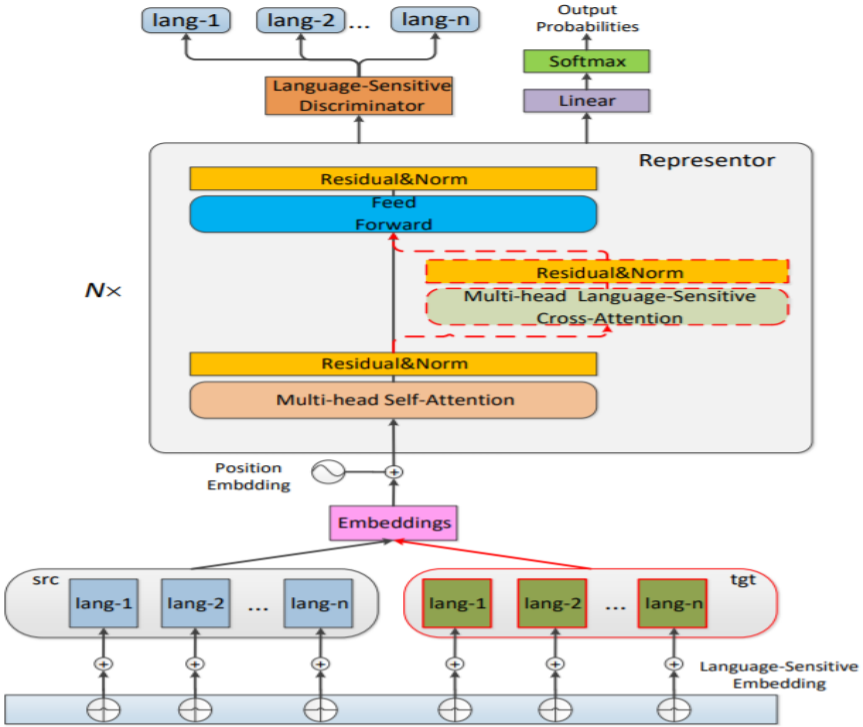

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Vi´egas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: Enabling zero-shot translation.

# 多语言翻译

- 发展趋势：
  - 共享编码器-解码器
  - 区分语言特征

# 多语言翻译

- 发展趋势：
  - 共享编码器
  - 区分语言特



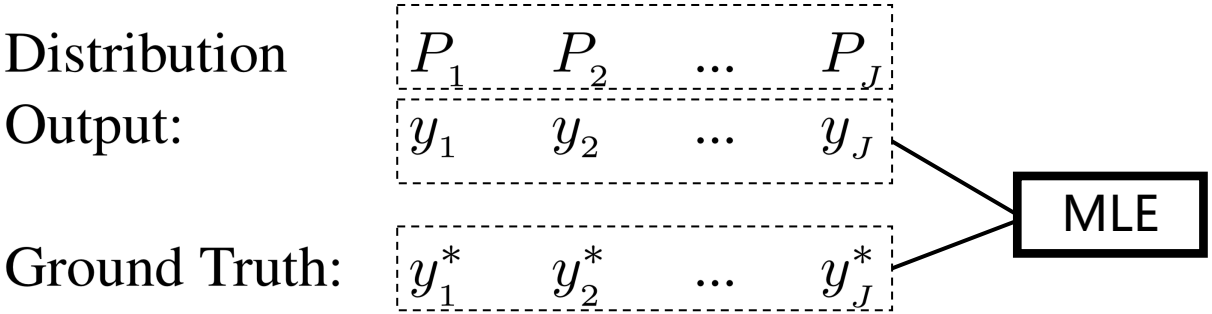

Yining Wang, Long Zhou, Jiajun Zhang, Feifei Zhai, Jingfang Xu and Chengqing Zong. A Compact and Language-Sensitive Multilingual Translation Method . *In Proceedings of ACL 2019.*

# 多语言翻译

- 发展趋势：
  - 共享编码器-解码器
  - 区分语言特征
  - 改善低资源和零资源翻译

# 模型训练

- 通用框架：Teacher Forcing



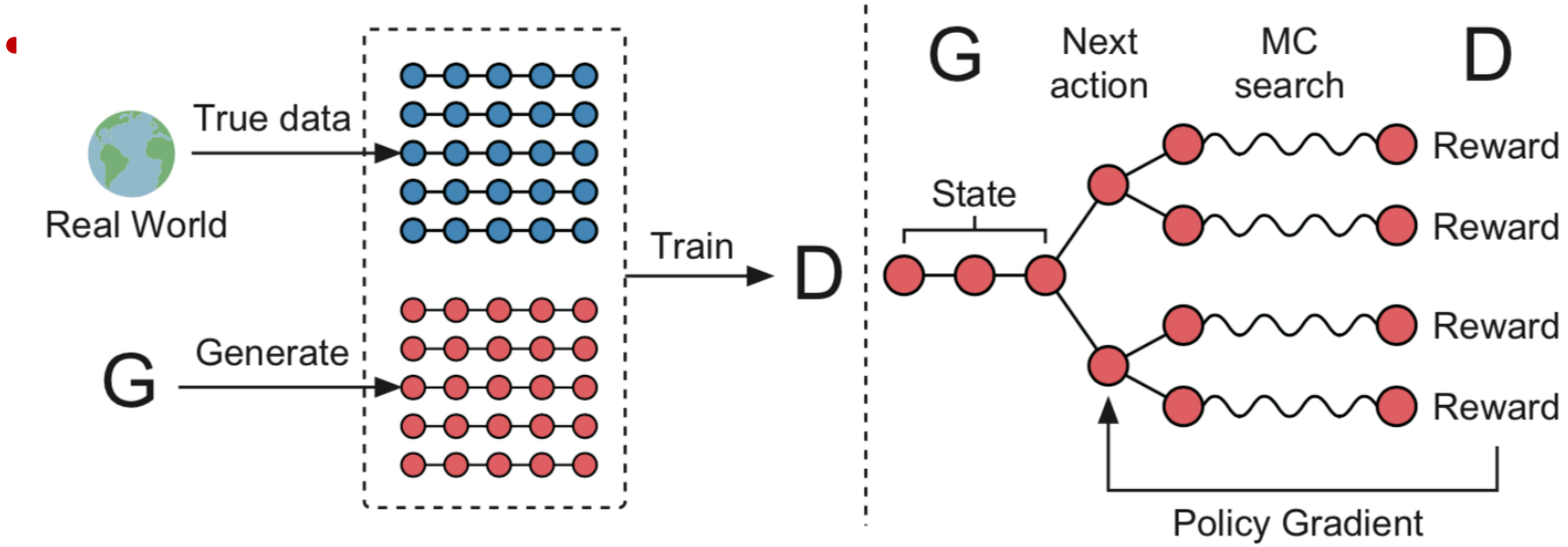

Loss: $$\mathcal{L}_{\text{MLE}} = -\sum_{j=1}^{J} \log P_j(y_j^*)$$

# 模型训练

- 发展趋势：
  - 基于强化学习的序列级训练

# 模型训练

- 发展趋势：



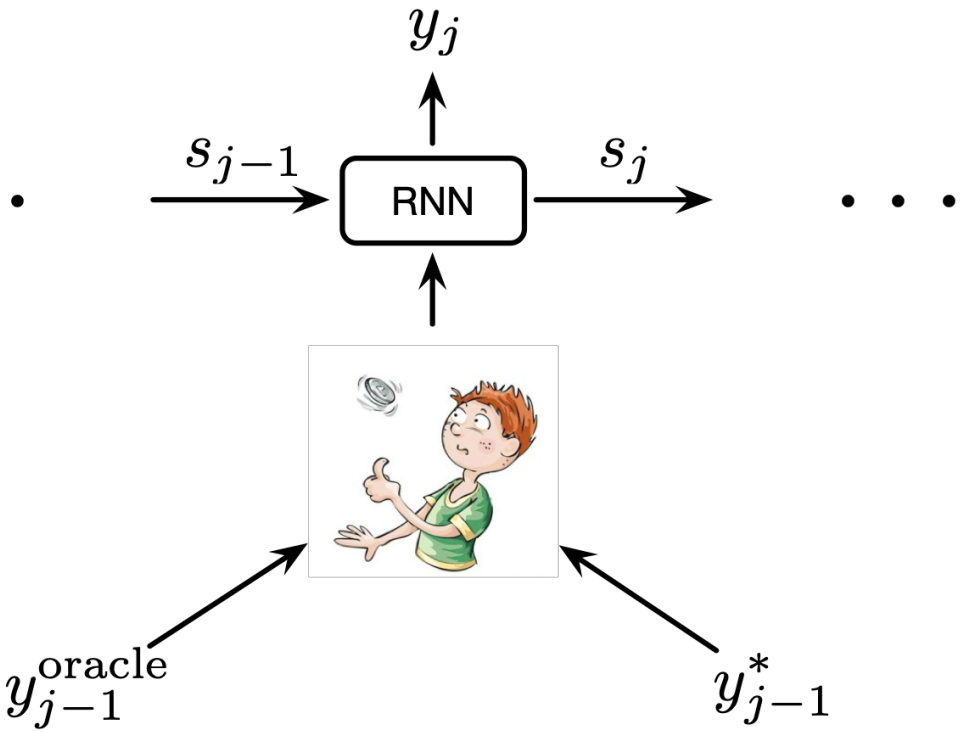

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. *In AAAI*.

# 模型训练

- 发展趋势：
  - 基于强化学习的序列级训练
  - 训练和测试一致性

# 模型训练

- 发展趋势
  - 基于强
  - 训练和




Wen Zhang, Jiawei Hu, Yang Feng, and Qun Liu. Refining source representations with relation networks for neural machine translation. In Proceedings of *ACL 2019*.

# 模型训练

- 发展趋势：
  - 基于强化学习的序列级训练
  - 训练和测试一致性
  - 可导的序列级训练

Chenze Shao, Xilin Chen, and Yang Feng. 2018. Greedy search with probabilistic n-gram matching for neural machine translation.. In Proceedings of *EMNLP 2018*.

# 低资源翻译

- 通用模型：数据增强，如Back Translation
- 趋势： 仍以数据增强为主。
- 最新论文：
  - ACL 2019：
    - Revisiting Low-Resource Neural Machine Translation: A Case Study
    - Generalized Data Augmentation for Low-Resource Translation（短文）
  - EMNLP 2019:
    - Handling Syntactic Divergence in Low-resource Machine Translation
    - Exploiting Multilingualism through Multistage Fine-Tuning for Low-Resource Neural Machine Translation

# Pre-train

- 在机器翻译上没有显著提升
  - 机器翻译语料规模已足够？

# 目录

CONTENTS

# Thanks for your attention!