

机器翻译进展综述

冯洋

中国科学院计算技术研究所

2019.5.4



中国科学院计算技术研究所
INSTITUTE OF COMPUTING TECHNOLOGY, CHINESE ACADEMY OF SCIENCES

目录

CONTENTS

一 任务定义

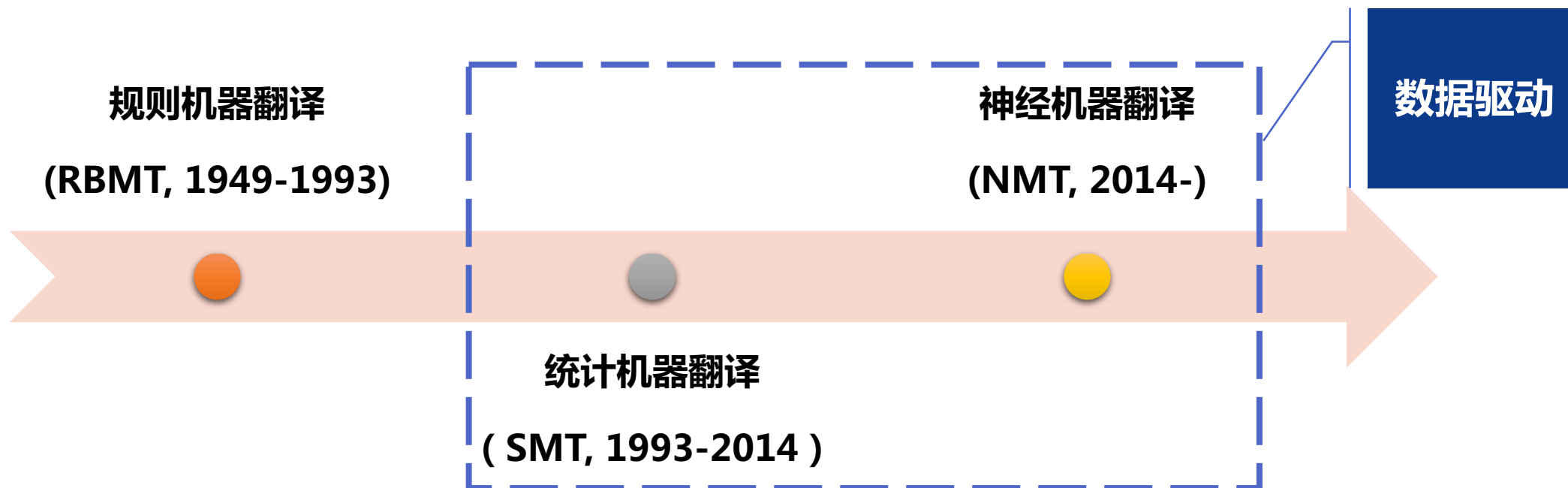
二 研究现状

三 研究热点

四 未来趋势

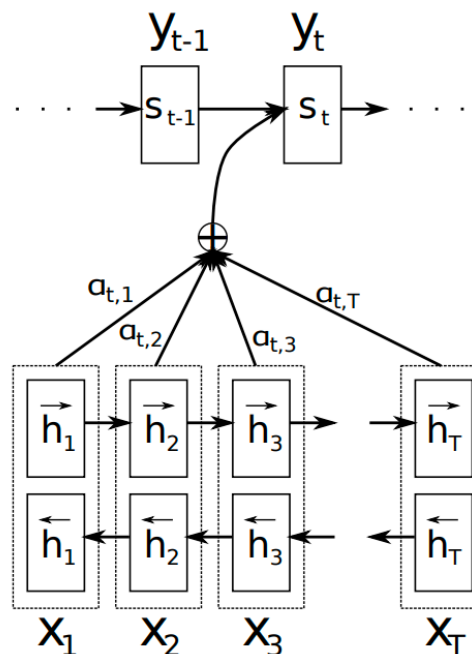
- 机器翻译(Machine translation, MT)

- 自然语言处理的重要研究领域，研究如何利用计算机把一种语言（源语言，source language）翻译成另一种语言（目标语言，target language）

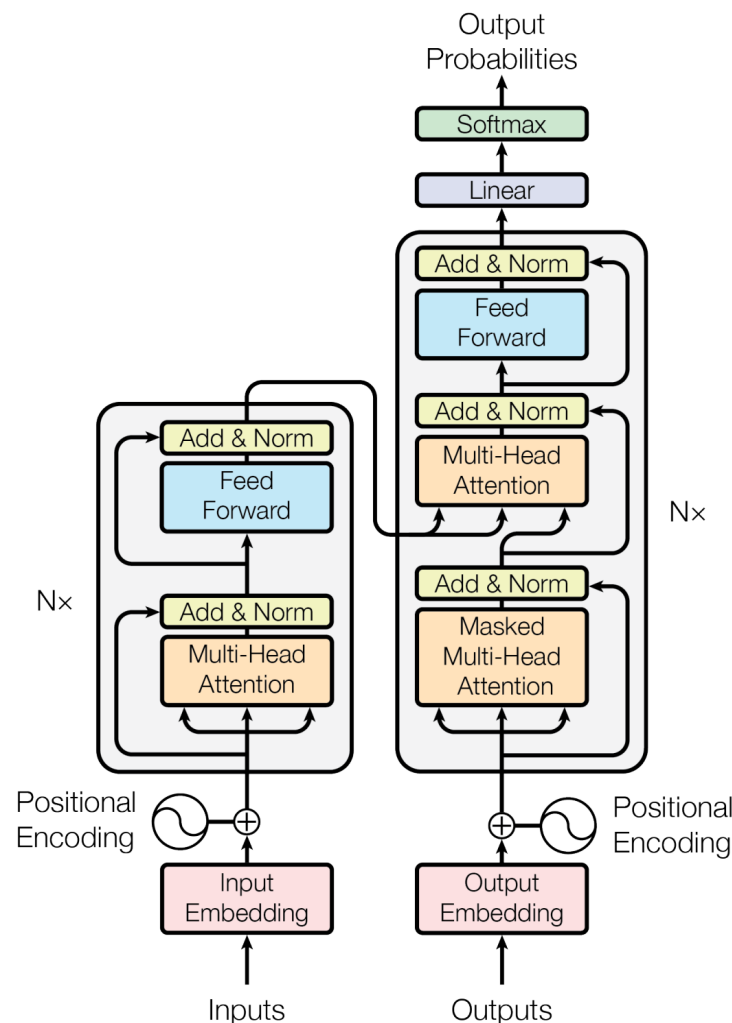


主流模型：带有注意力机制的编码器-解码器结构

- 按照对序列内部建模的方式
 - 基于循环神经网络 (RNN)
 - RNNSearch
 - 基于自注意力机制 (Self-Attention)
 - Transformer
 - 基于卷积网络 (CNN)



RNNSearch



Transformer

研究热点

技术

强化学习

对抗学习

DAE / VAE

应用场景

篇章翻译

语音翻译

低资源语言翻译

领域自适应

强化学习

- **诉求**：序列级训练
- **Rewards**：BLEU/GLEU，分类概率
- **搜索**：MC Search
- **训练**：REINFORCE算法

$$L_{\theta} = - \sum_{\mathbf{x}=x_{1:T}} p(\mathbf{x}) \cdot r(\mathbf{x})$$

$$\nabla_{\theta} L_{\theta} = -\mathbb{E}_{\mathbf{x}} \left[\sum_{t=1}^T \nabla_{\theta} \log(\tilde{p}(x_t)) \cdot r(\mathbf{x}) \right]$$

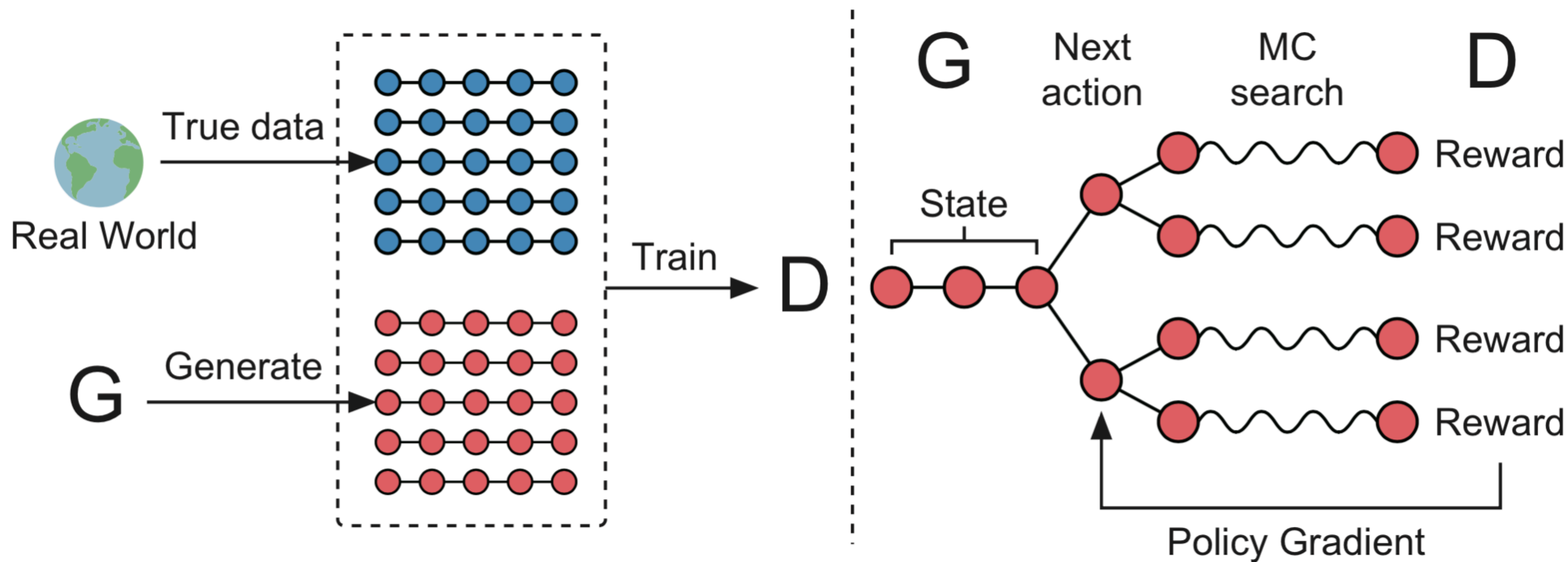
- **难点**：如何选取rewards的baseline以减小方差

对抗学习

- **特点**：生成器和判别器互相对抗，共同促进
- **狭义**：生成对抗网络GAN
- **广义**：所有的MinMax问题
- **难点**：如何达到纳什平衡
- **训练**：
 - 分别交替训
 - Joint training

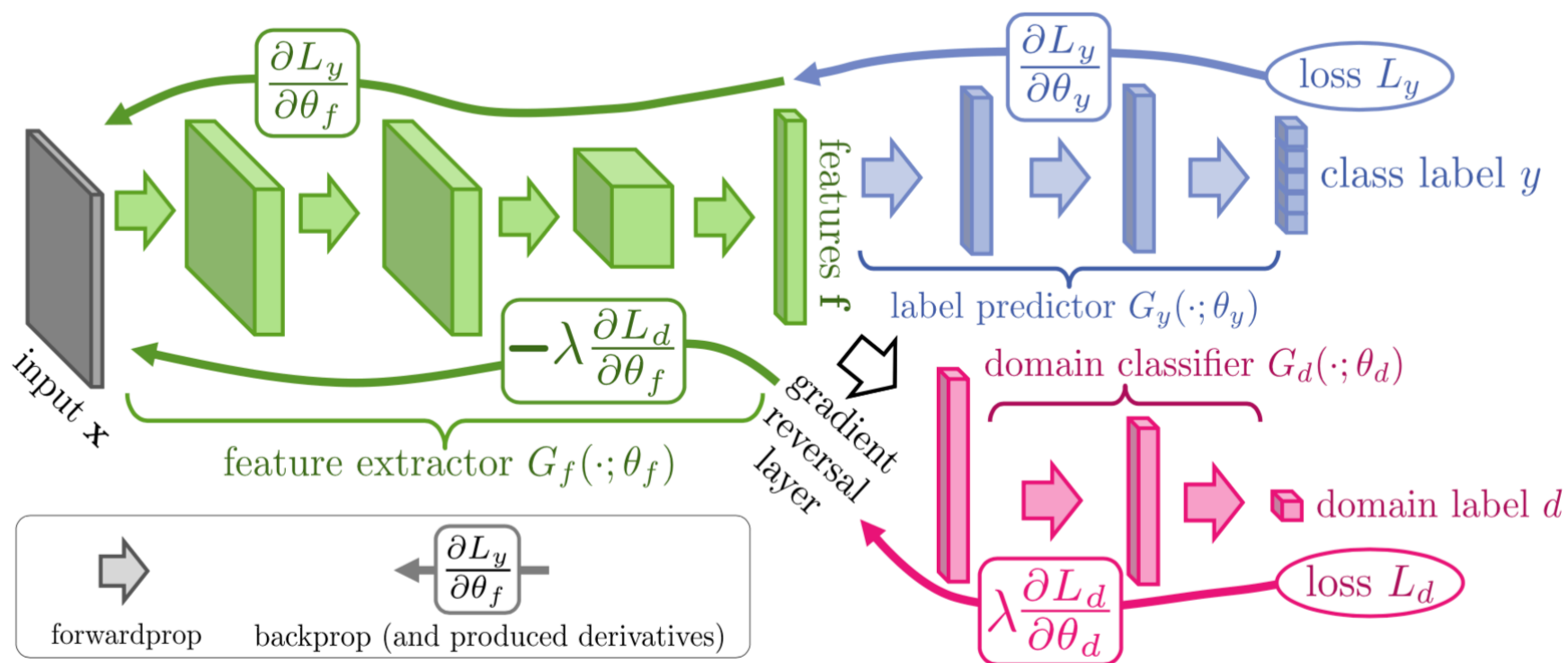
强化学习+对抗学习实例：SeqGan

- 生成网络和识别网络轮流训练



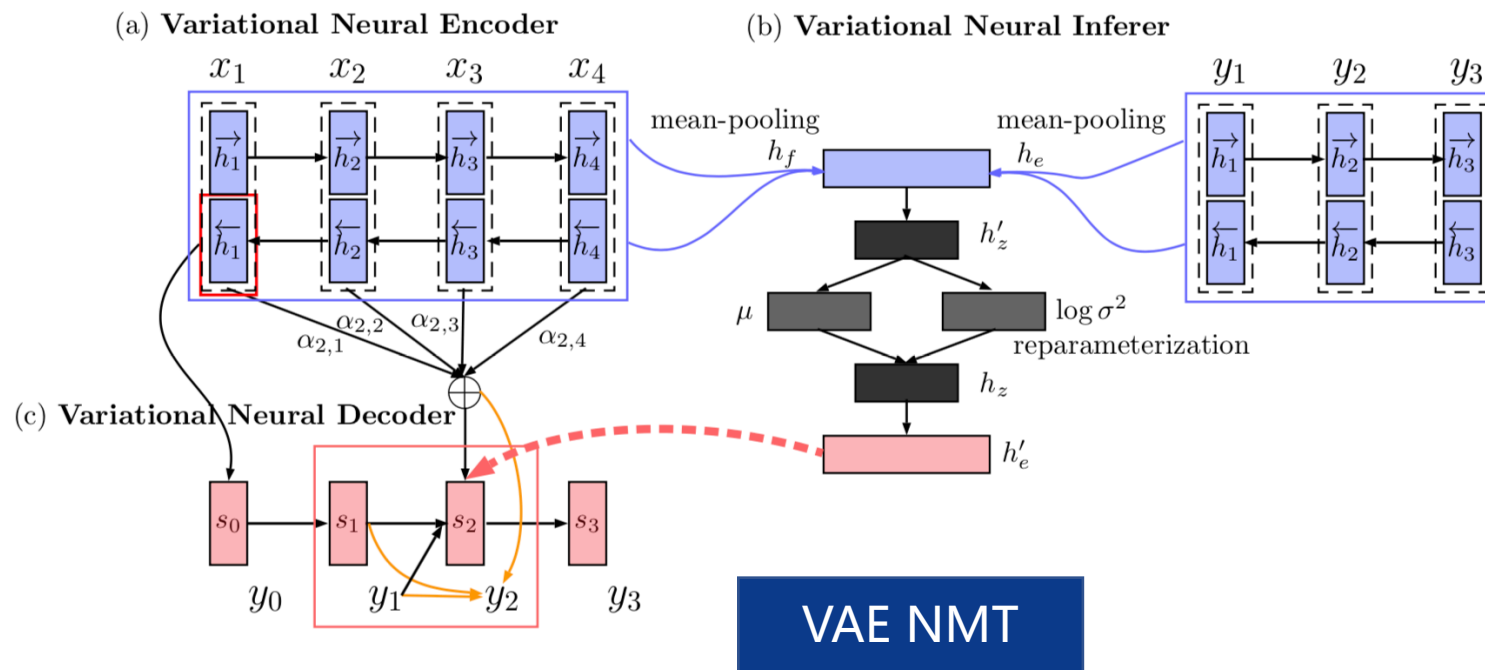
对抗学习实例：对抗编码生成

- Joint training



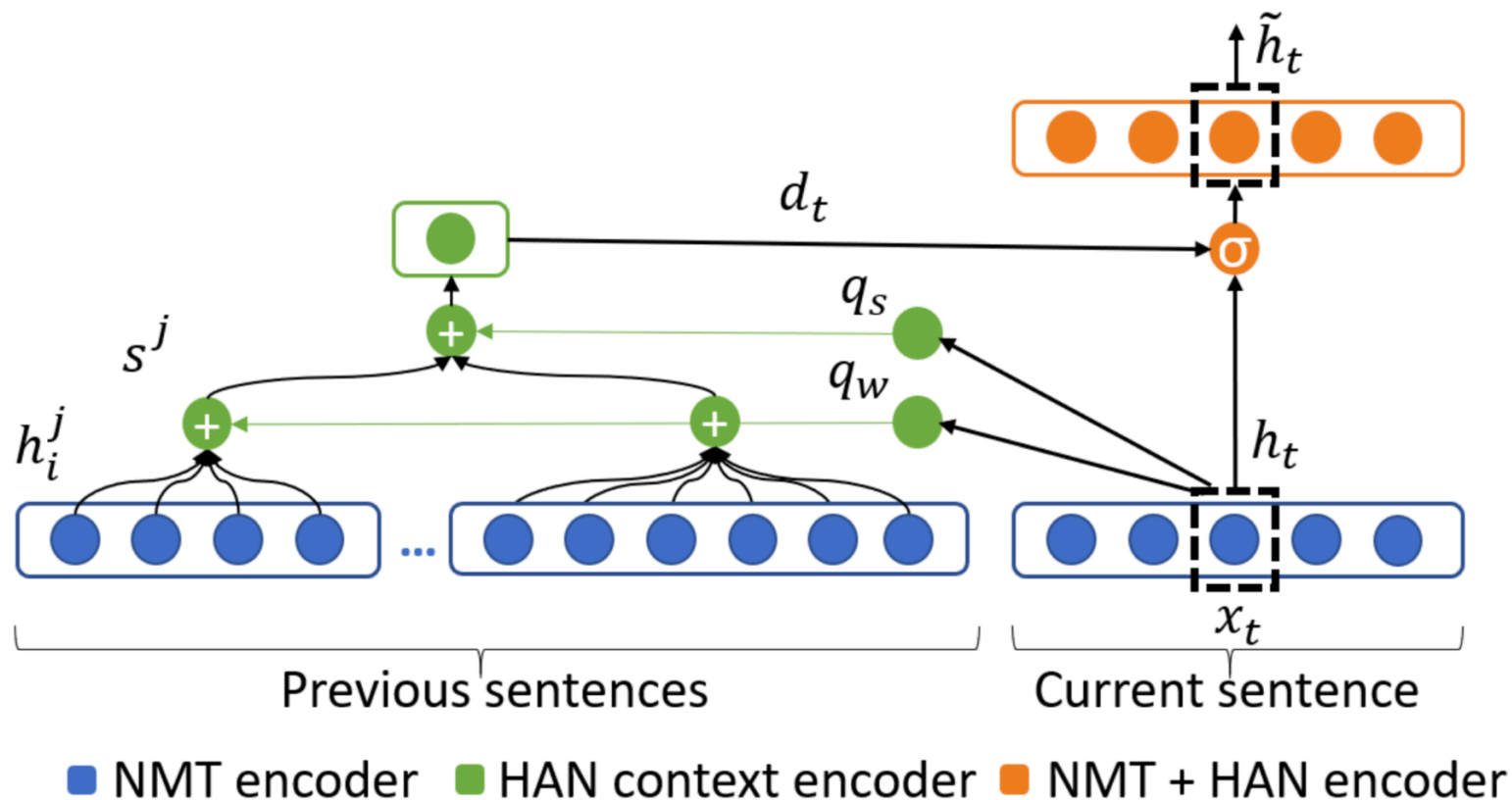
DAE / VAE

- 特点：重构输入
- DAE 多用于无监督学习
- VAE 引入新的变量，多用于提升多样性和鲁棒性



篇章翻译

- 层次化attention

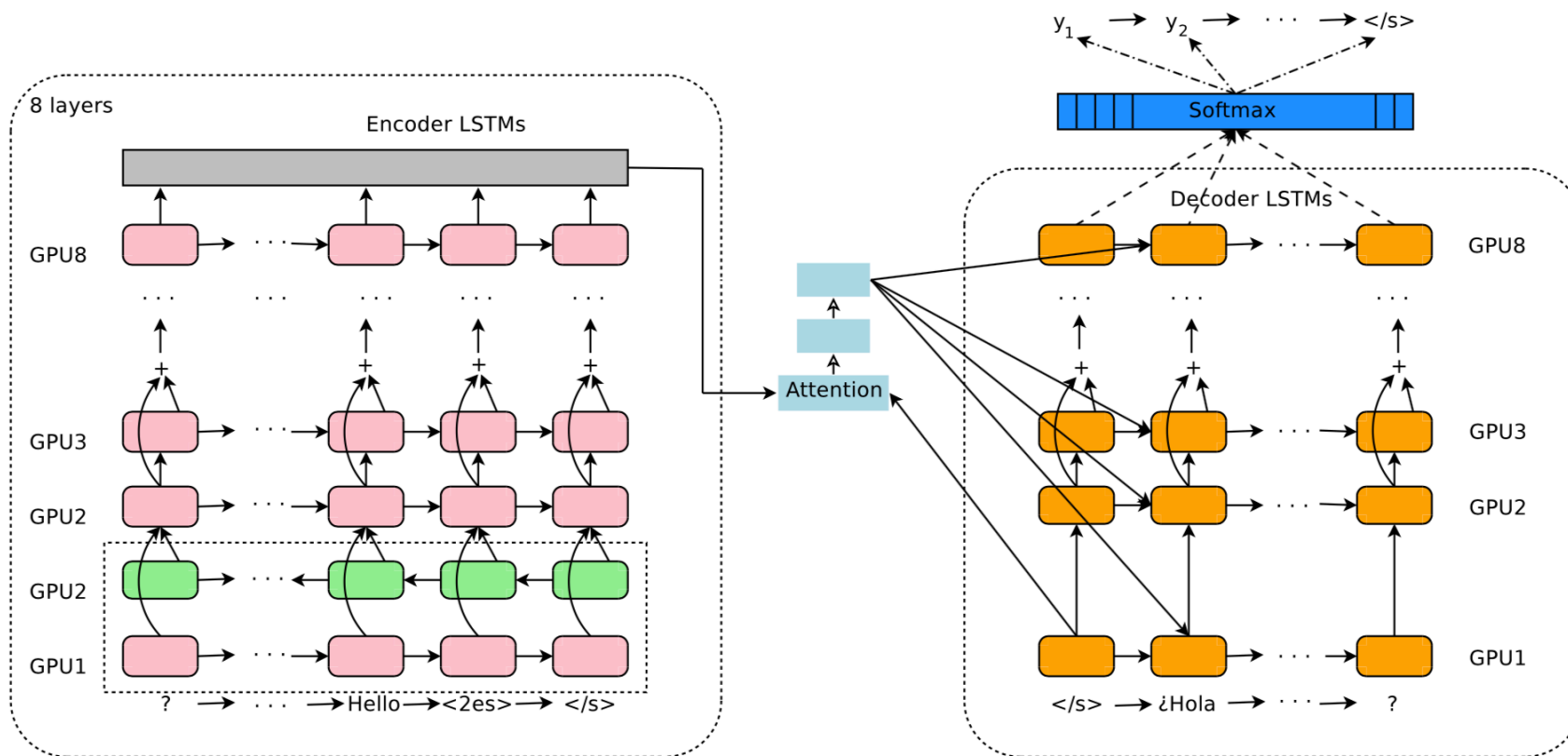


语音翻译

- **场景定义**：输入是语音，输出翻译的文本
- **现有方法**：先ASR，再翻译
 - 干净语料引入人工噪音
 - 引入拼音信息
 - 针对特定语言现象进行处理
- CCMT语音翻译评测
- End-to-end翻译

低资源语言翻译

- 场景：多种语言对翻译
- 方法：Google' s Multilingual Machine Translation System



低资源语言翻译

- 场景：有大量的单语资源
- 方法：Back translation

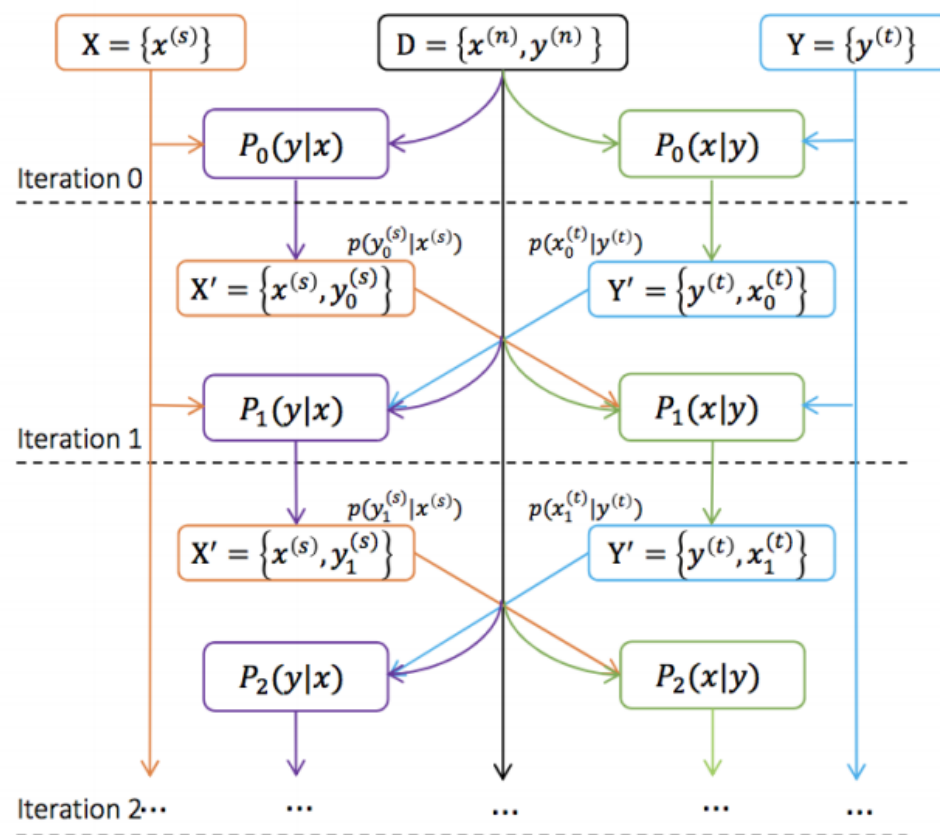
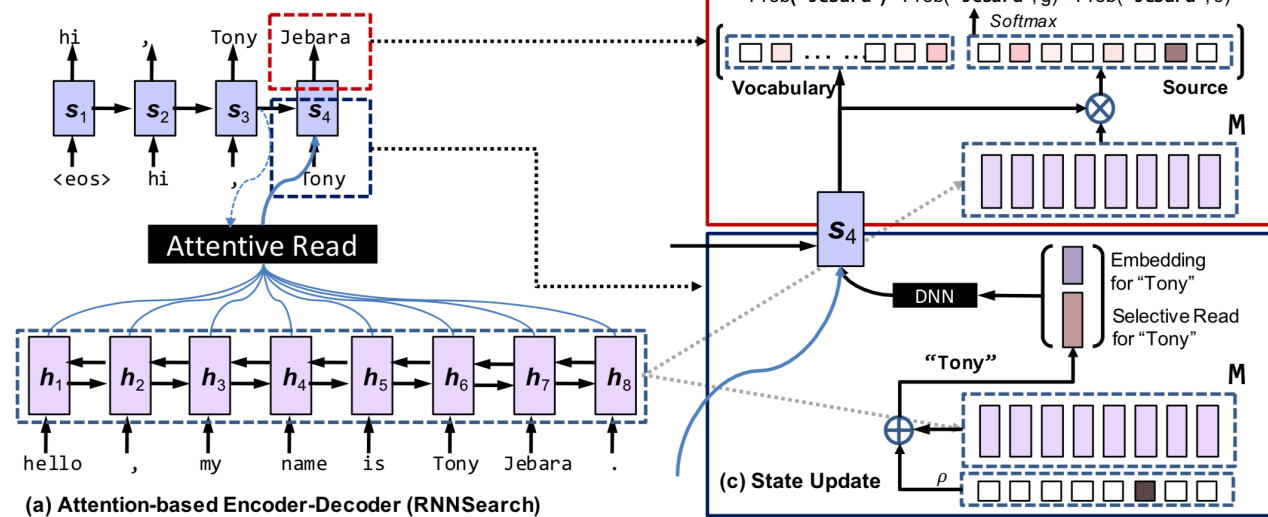
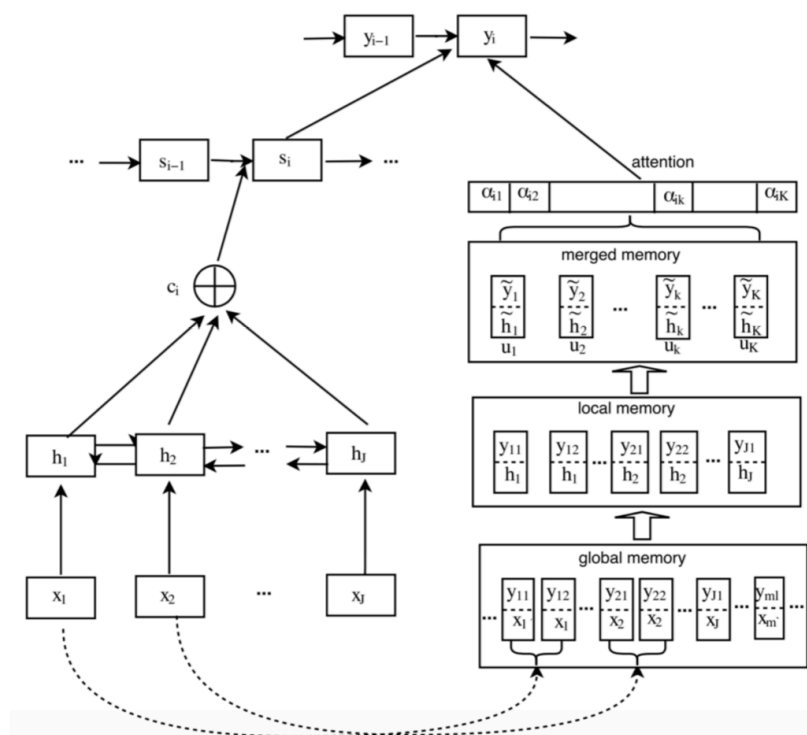


Figure 1: Illustration of joint training: S2T $p(\mathbf{y}|\mathbf{x})$ and T2S $p(\mathbf{x}|\mathbf{y})$

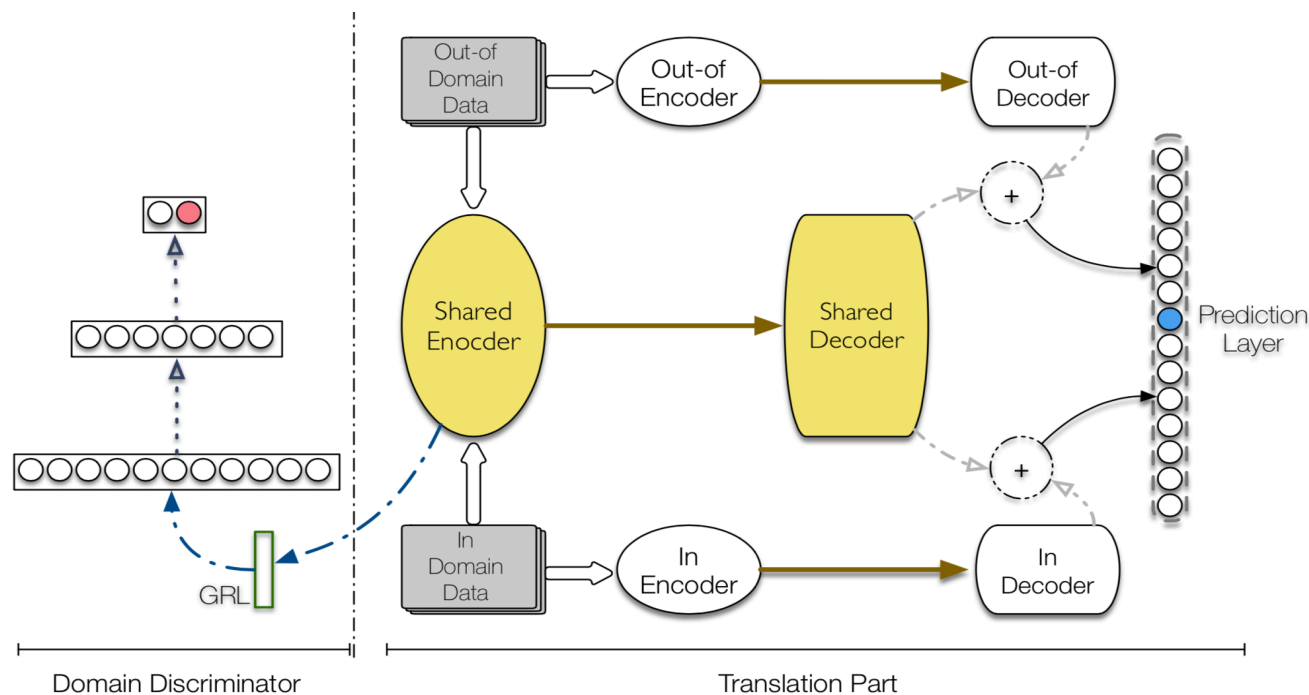
低资源语言翻译

- 场景：有大量的Rare words
- 方法：Memory networks 或 COPYNET



领域自适应

- **场景**：训练语料是通用领域，测试句子为某个特定领域
- **方法**：
 - 最简单有效 -- Fine tune
 - 特征分类



未来研究热点：

- 迁移学习
 - 从一种语言到另一种语言
- 元学习
 - 自动调参
- 同声传译
 - 如何更好的识别语音