CCL & NLP-NABD 2016
“深度学习与机器翻译” 讲习班

第二部分

神经机器翻译

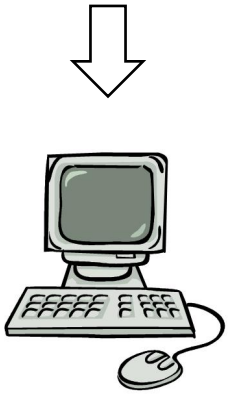
刘洋　　张家俊

2016年10月14日，山东烟台

机器翻译

- 目标：利用计算机实现自然语言的自动翻译

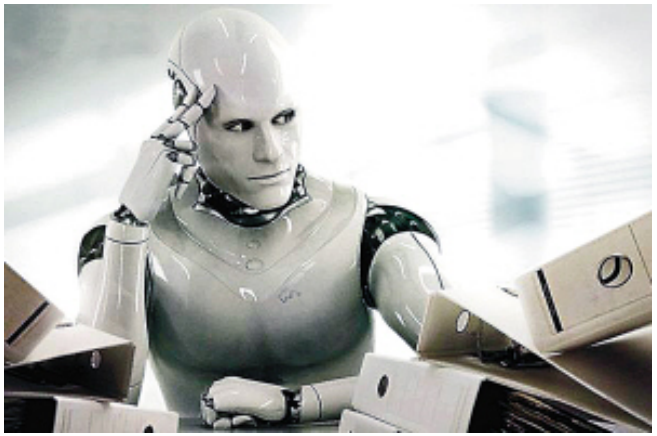
布什 | 与 | 沙龙 | 举行 | 了 | 会谈



Bush | held | a | talk | with | Sharon

发展历史

- 趋势：让机器更“自主”地学习如何翻译

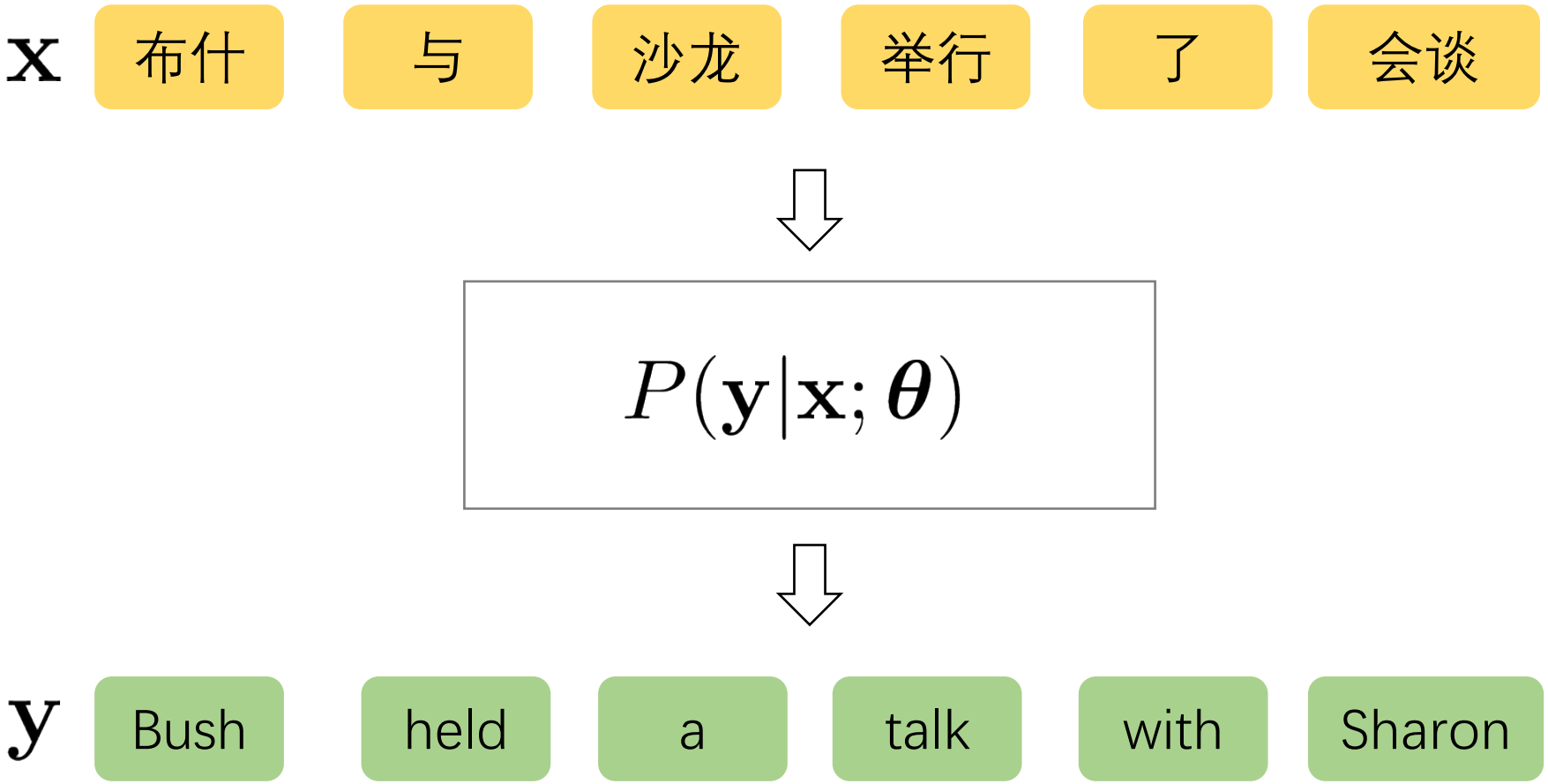


规则
机器翻译
1980

数据驱动
机器翻译
1990

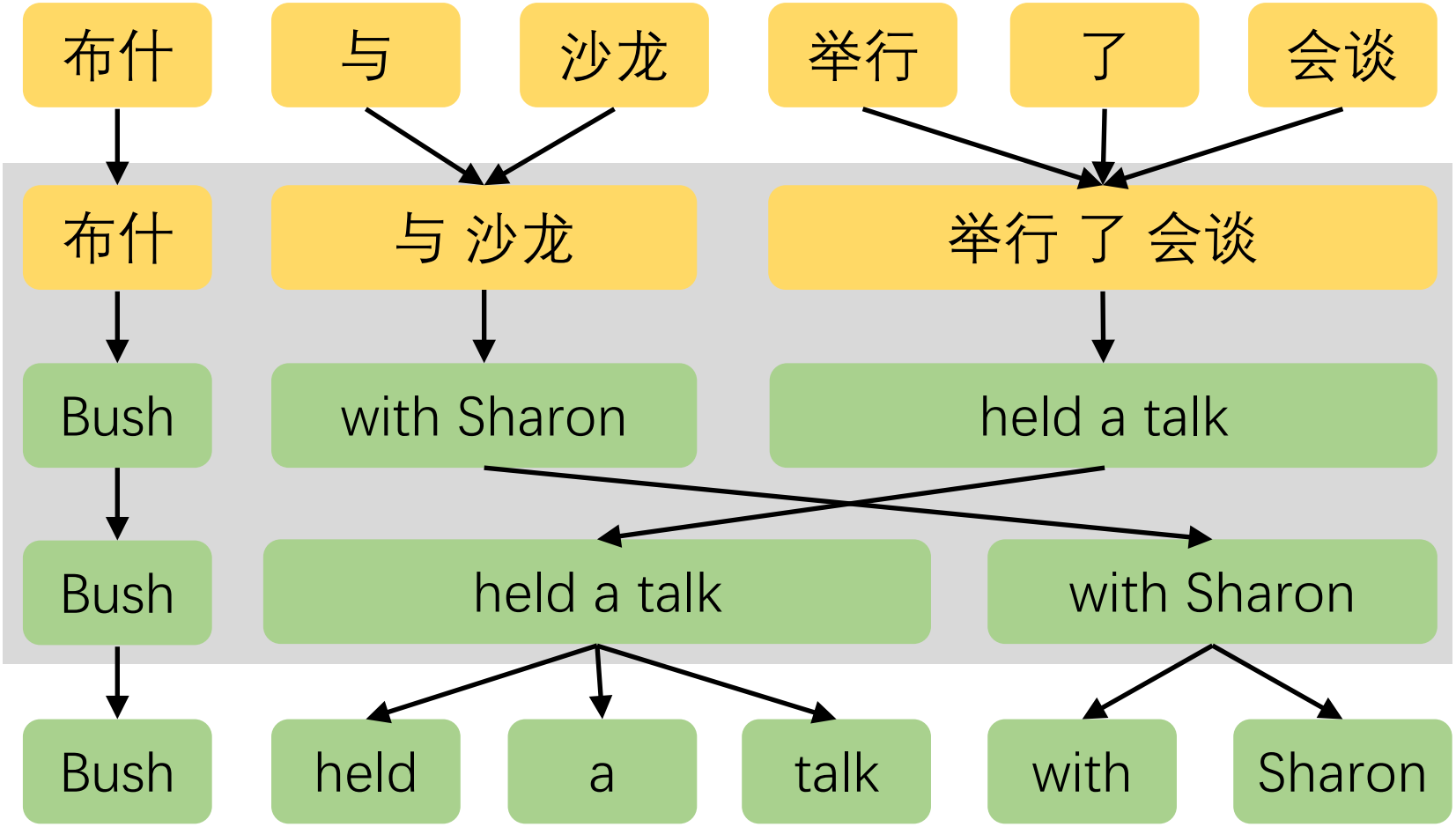
数据驱动的机器翻译

- 核心问题：如何为翻译过程建立概率模型？



基于短语的统计机器翻译

- 短语翻译模型：以隐结构短语为基本翻译单元



(Koehn et al., 2003)

统计机器翻译

- 隐变量对数线性模型：在隐式语言结构上设计特征

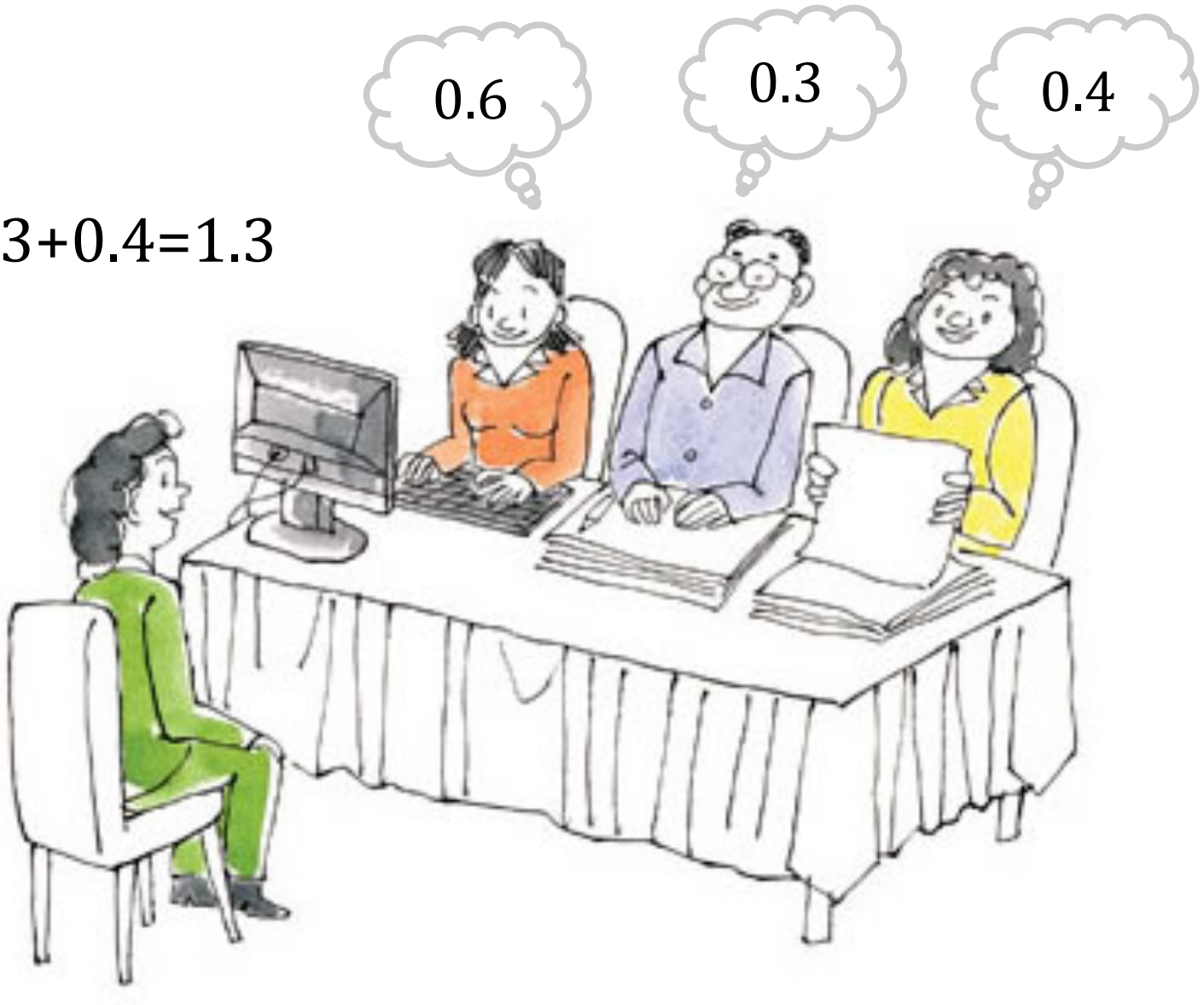
$\mathbf{x}$ 布什 与 沙龙 举行 了 会谈

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \sum_{\mathbf{z}} \frac{\exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}, \mathbf{z}))}{\sum_{\mathbf{y}'} \sum_{\mathbf{z}'} \exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\mathbf{x}, \mathbf{y}', \mathbf{z}'))}$$

(Och and Ney, 2002)

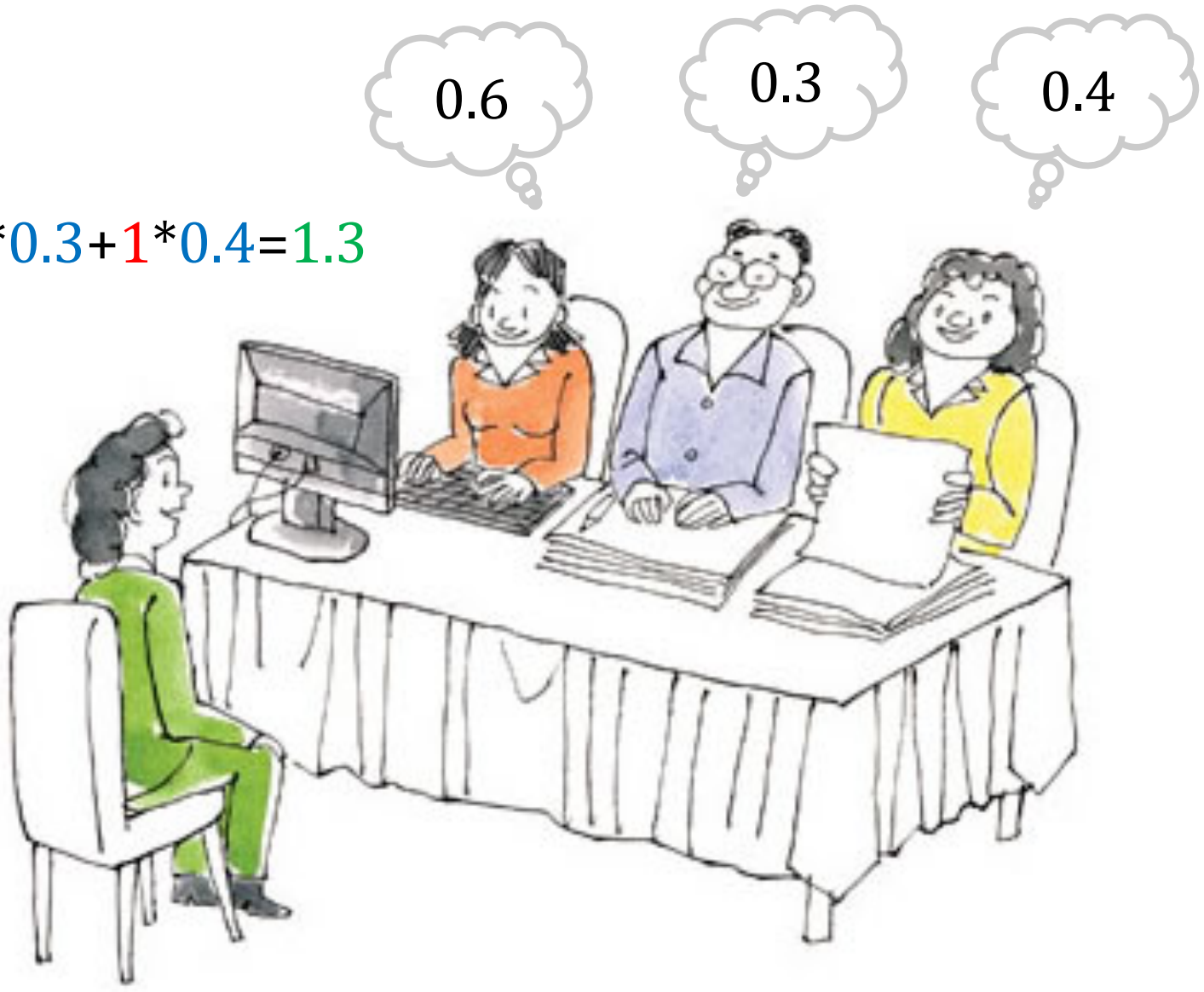
$\mathbf{y}$ Bush held a talk with Sharon

例子



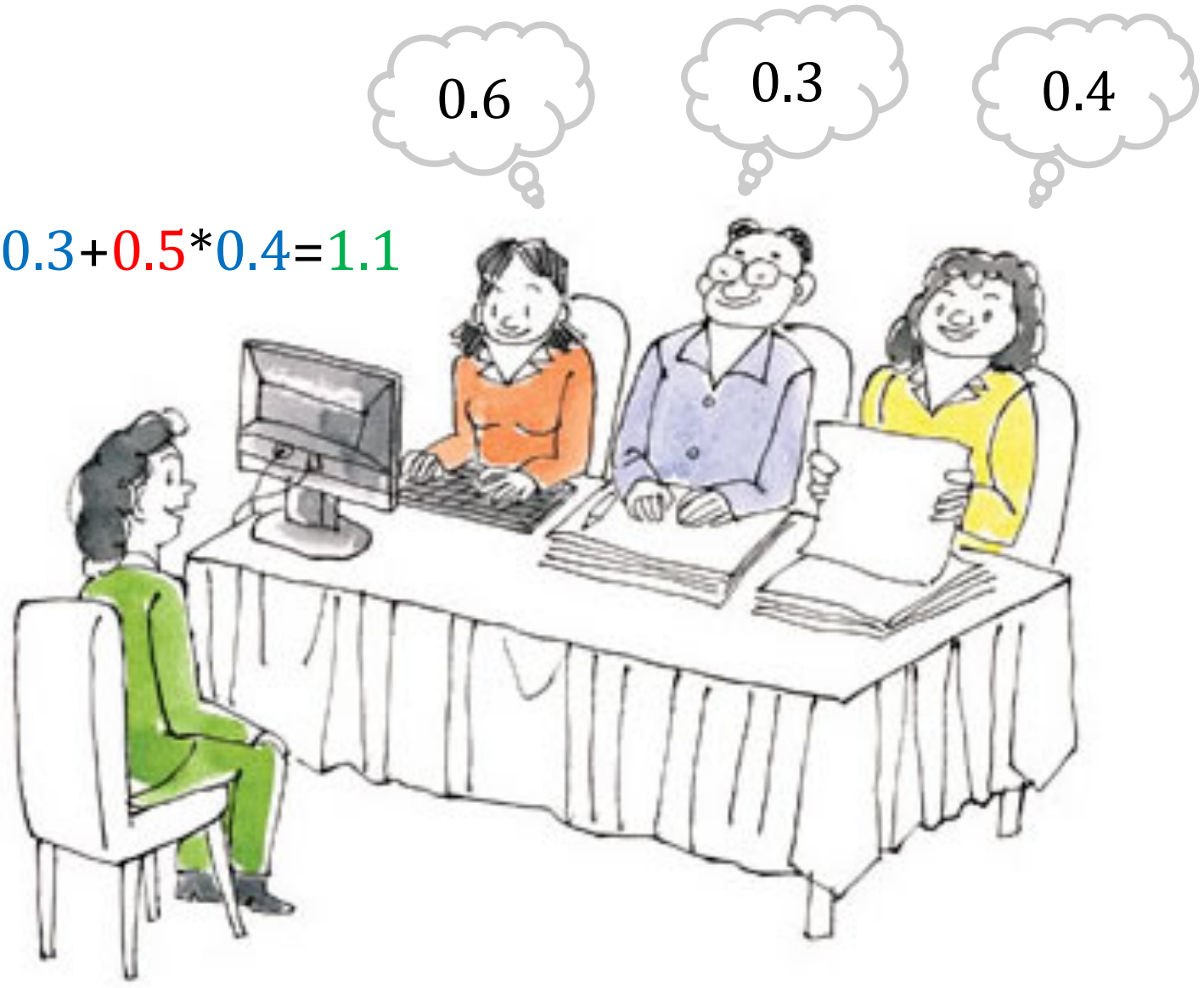
0.6+0.3+0.4=1.3

例子



1*0.6+1*0.3+1*0.4=1.3

例子

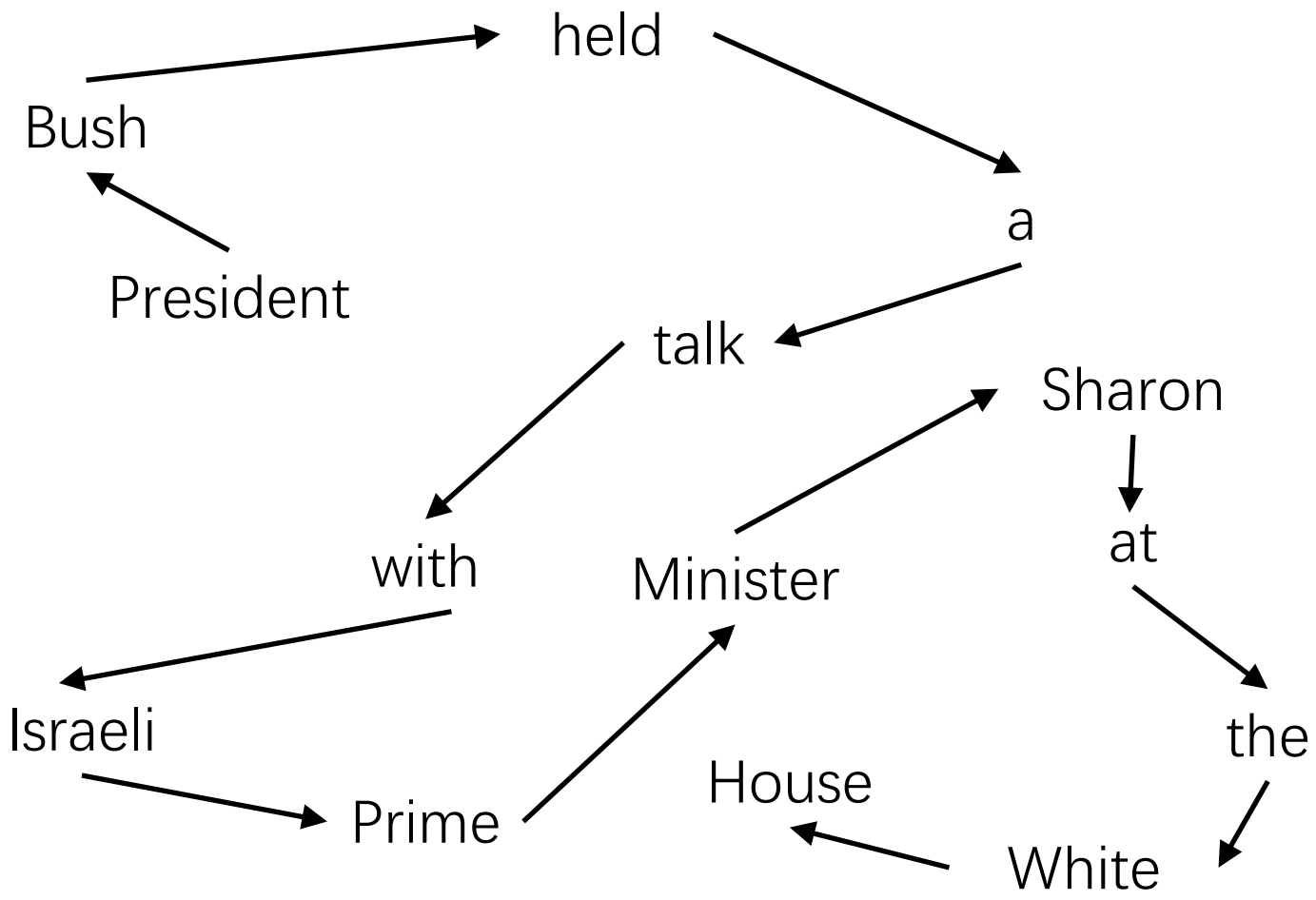


0.5*0.6+2*0.3+0.5*0.4=1.1

统计机器翻译的优缺点

- 优点
  - 隐结构可解释性高
  - 利用局部特征和动态规划处理指数级结构空间
- 缺点
  - 线性模型难以处理高维空间中线性不可分的情况
  - 需要人类专家设计隐式结构及相应的翻译过程
  - 需要人类专家设计特征
  - 离散表示带来严重的数据稀疏问题
  - 难以处理长距离依赖

难点：长距离调序



如何用上述词语拼成合理的译文？

统计机器翻译示例

Chinese

美国总统布什昨天在白宫与以色列总理沙龙就中东局势举行了一个小时的会谈。

English

Yesterday, U.S. President George W. Bush at the White House with Israeli Prime Minister Ariel Sharon on the situation in the Middle East held a one-hour talks.

深度学习带来新思路

nature

full stop is chosen[17,72,76]. Overall, this process generates sequences of French words according to a probability distribution that depends on the English sentence. This rather naive way of performing machine translation has quickly become competitive with the state-of-the-art, and this raises serious doubts about whether understanding a sentence requires anything like the internal symbolic expressions that are manipulated by using inference rules. It is more compatible with the

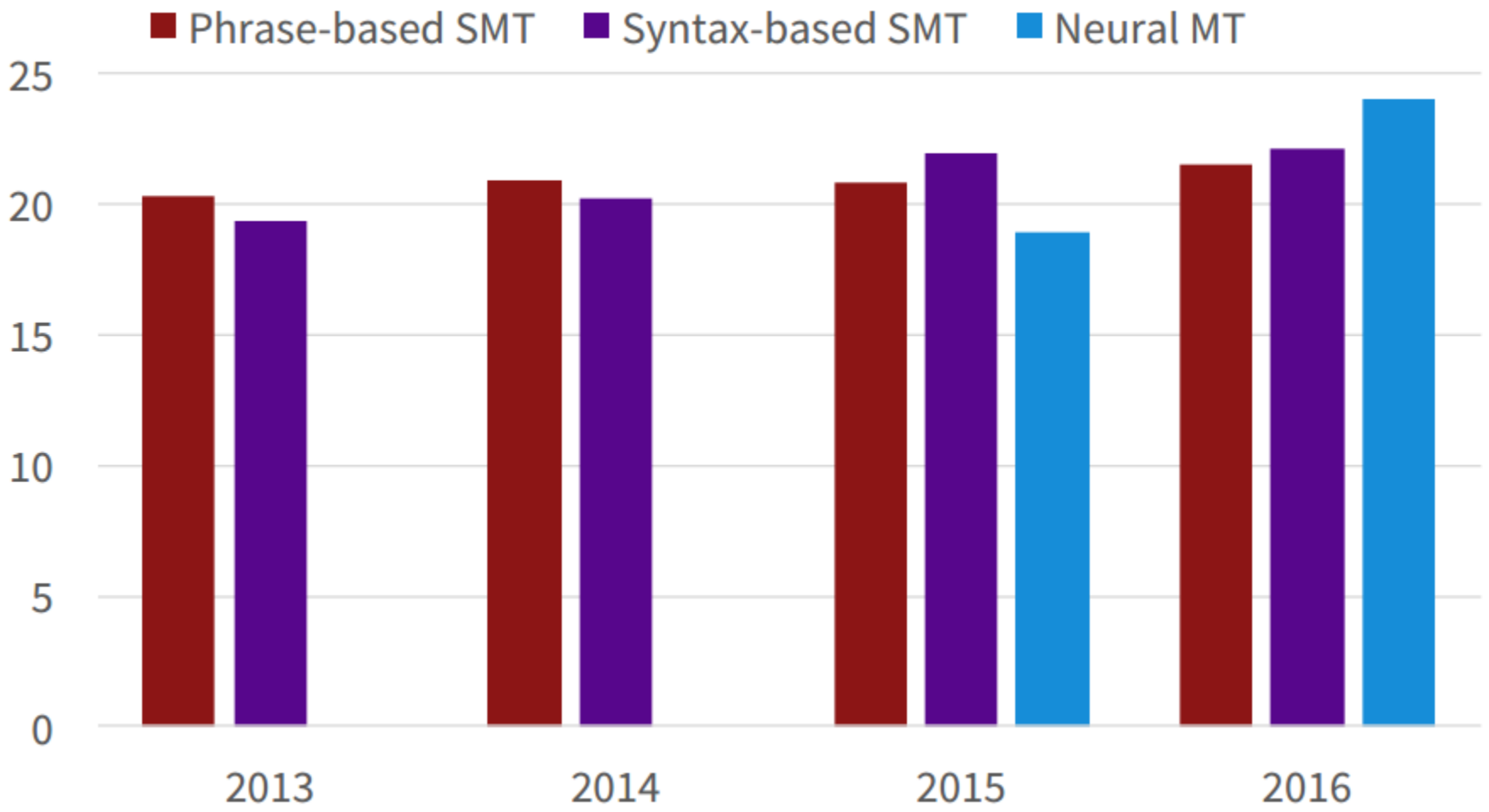
Yann LeCun

Yoshua Bengio

Geoffrey Hinton

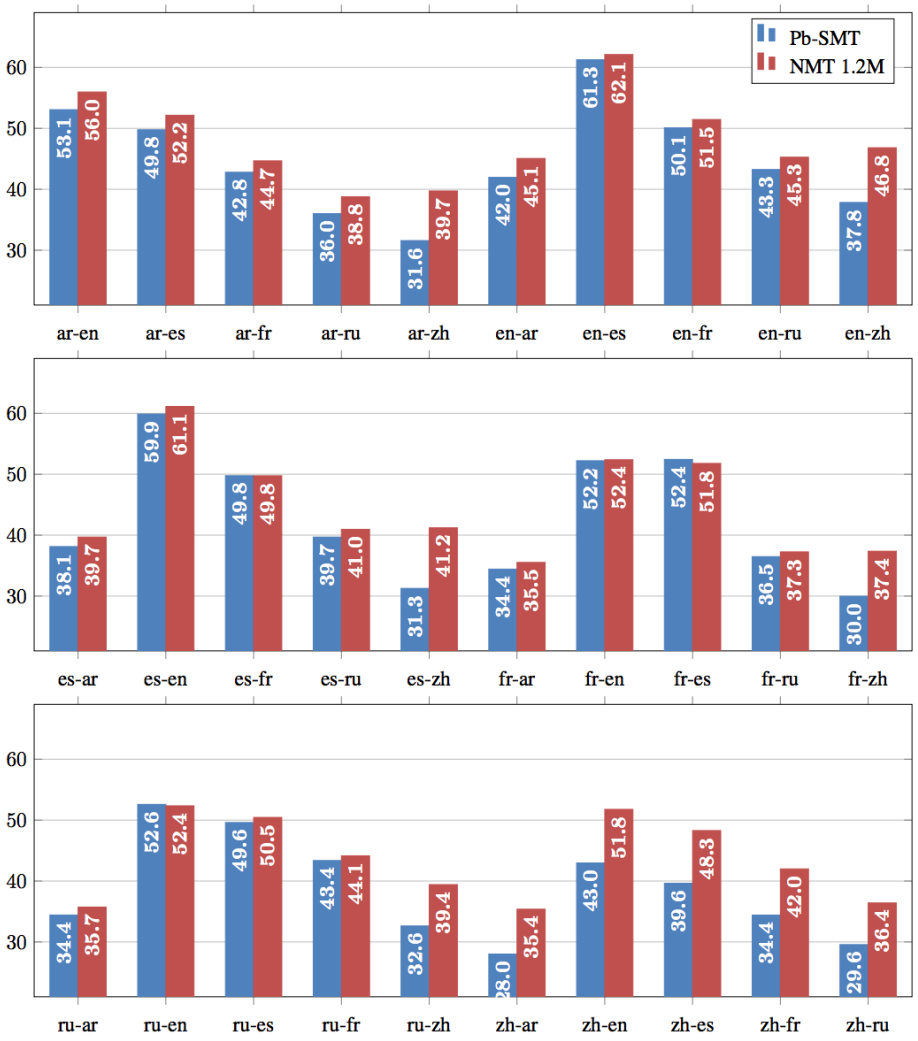
(LeCun et al, 2015)

机器翻译方法对比



英国爱丁堡大学在WMT英德评测数据上的BLEU值。NMT 2015年结果来自蒙特利尔大学。来源：Rico Sennrich报告和斯坦福ACL 2016 Tutorial。

机器翻译方法对比



Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions

(Junczys-Dowmunt et al, 2016)

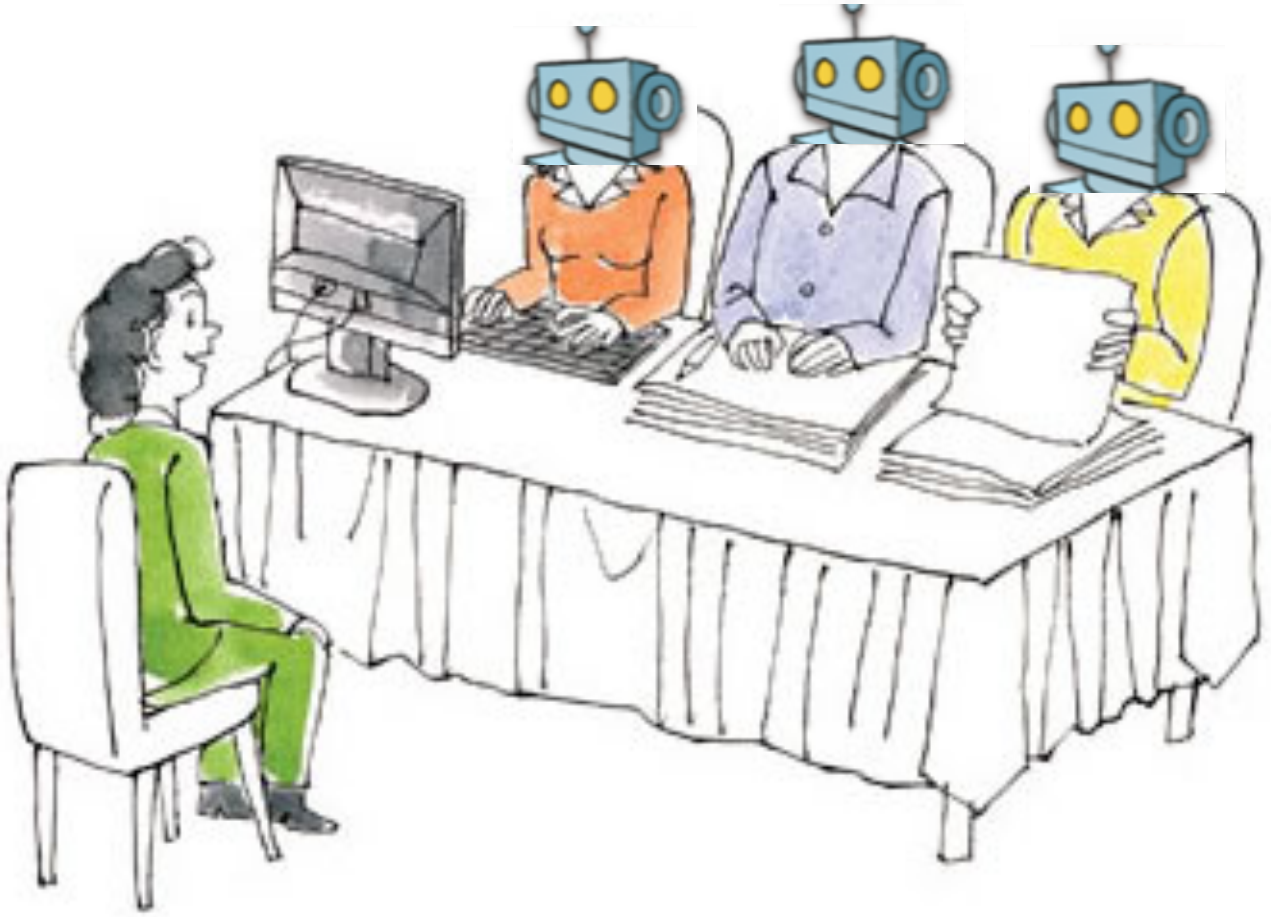
例子

1.1 • • •

• • •

例子

1.1



神经机器翻译

- 利用神经网络实现自然语言的映射

$\mathbf{x}$ 布什 与 沙龙 举行 了 会谈

$$P(\mathbf{y}|\mathbf{x};\boldsymbol{\theta}) = \prod_{n=1}^{N} P(\mathbf{y}_n|\mathbf{x}, \mathbf{y}_{<n};\boldsymbol{\theta})$$

(Sutskever et al, 2014)

$\mathbf{y}$ Bush held a talk with Sharon

如何对条件概率建模？

$$P(\mathbf{y}_n|\mathbf{x}, \mathbf{y}_{<n}; \boldsymbol{\theta})$$

$\mathbf{y}_n$	当前目标语言词
$\mathbf{x}$	源语言句子
$\mathbf{y}_{<n}$	已经生成的目标语言句子

挑战：数据稀疏

思路：用连续表示来代替离散表示

如何对条件概率建模？

$$P(\mathbf{y}_n|\mathbf{x}, \mathbf{y}_{<n}; \boldsymbol{\theta})$$

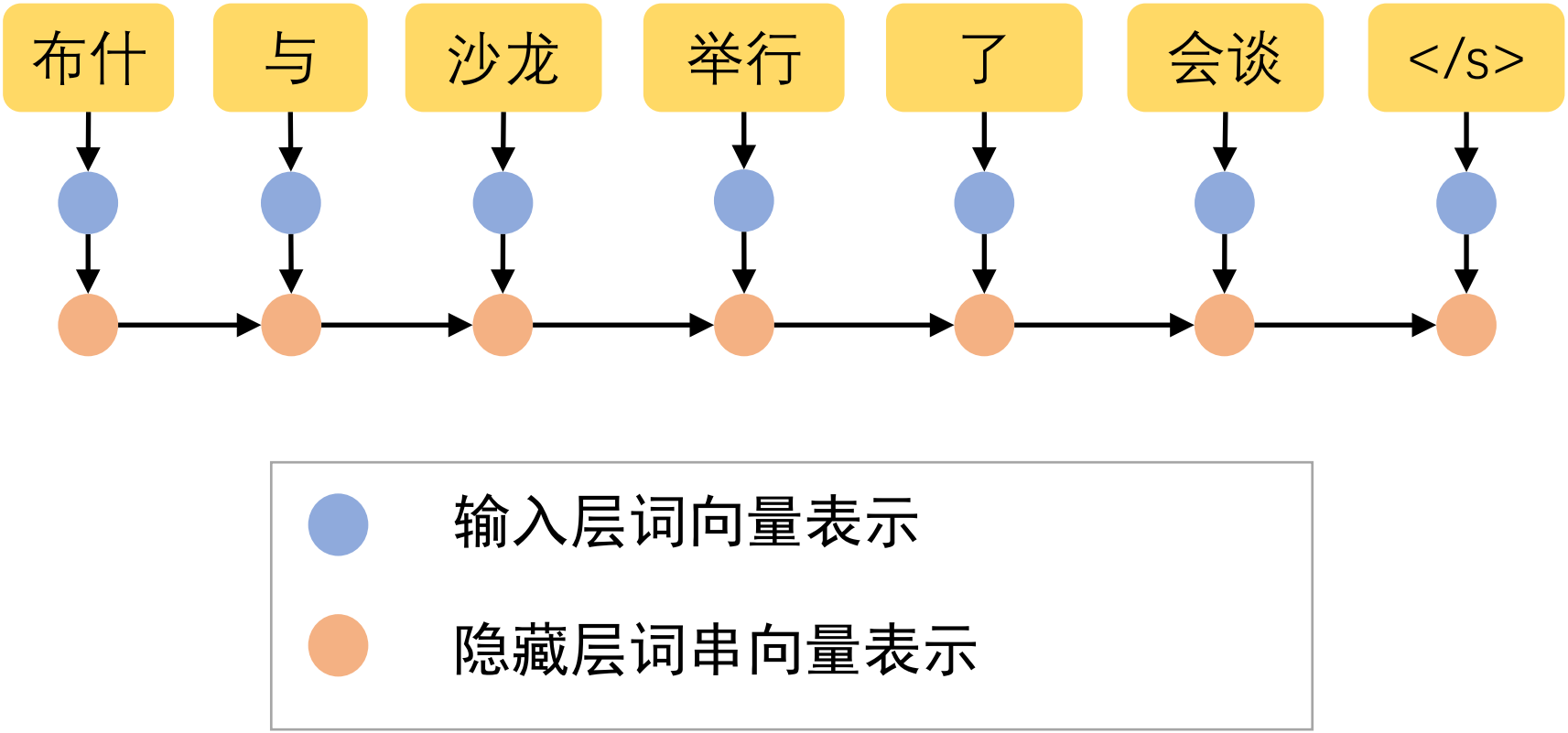
$$= \frac{\exp(\varphi(\mathbf{y}_n, \mathbf{x}, \mathbf{y}_{<n}, \boldsymbol{\theta}))}{\sum_{y \in \mathcal{Y}} \exp(\varphi(y, \mathbf{x}, \mathbf{y}_{<n}, \boldsymbol{\theta}))}$$

$$= \frac{\exp(\varphi(\mathbf{v}_{\mathbf{y}_n}, \mathbf{c}_s, \mathbf{c}_t, \boldsymbol{\theta}))}{\sum_{y \in \mathcal{Y}} \exp(\varphi(\mathbf{v}_y, \mathbf{c}_s, \mathbf{c}_t, \boldsymbol{\theta}))}$$

$\mathbf{v}_y$	目标语言词向量	$\mathcal{Y}$	目标语言词汇
$\mathbf{c}_s$	源语言上下文向量	$\mathbf{c}_t$	目标语言上下文向量

句子向量表示

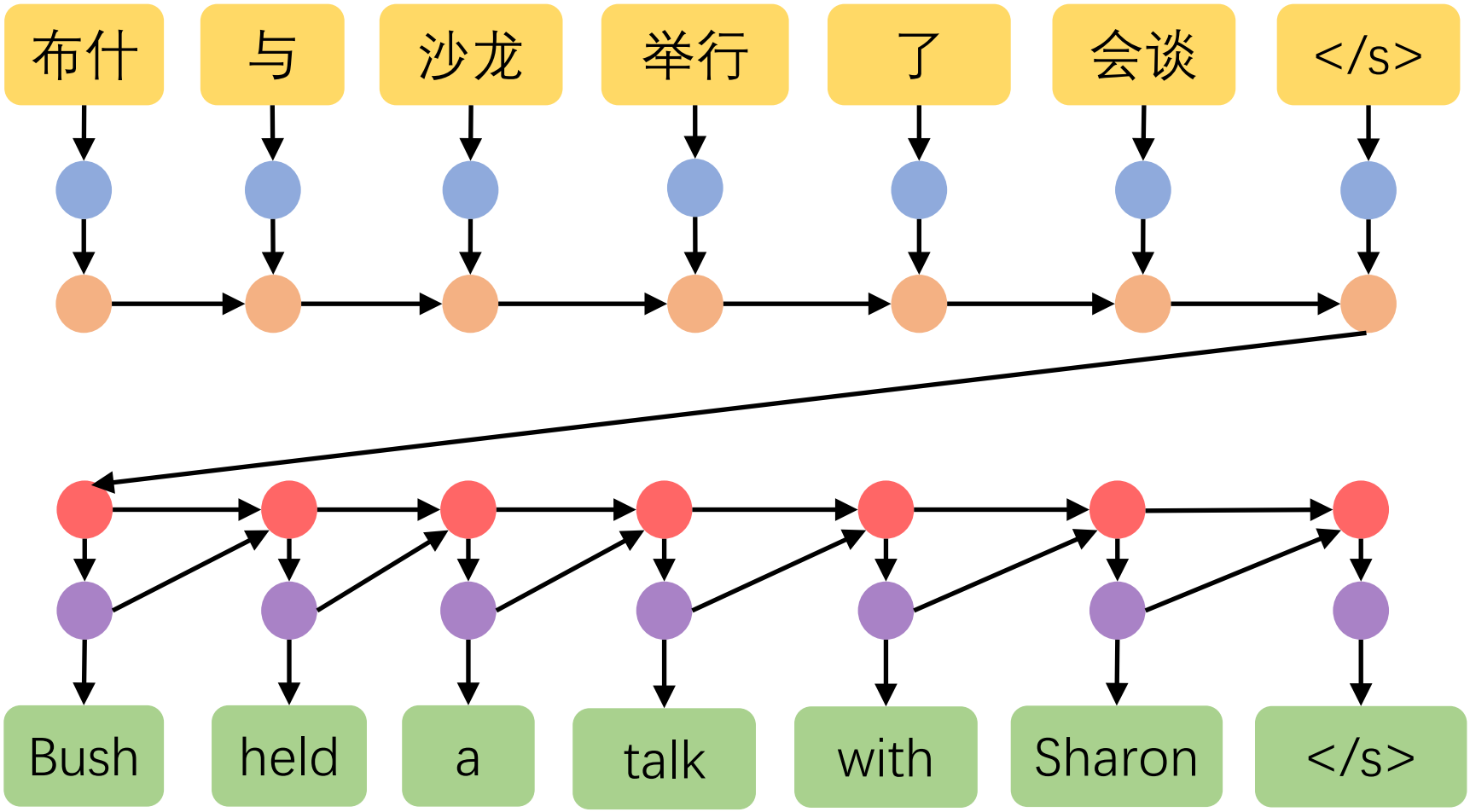
- 利用递归神经网络计算句子的向量表示



递归神经网络将句子“编码”为向量表示

编码器–解码器框架

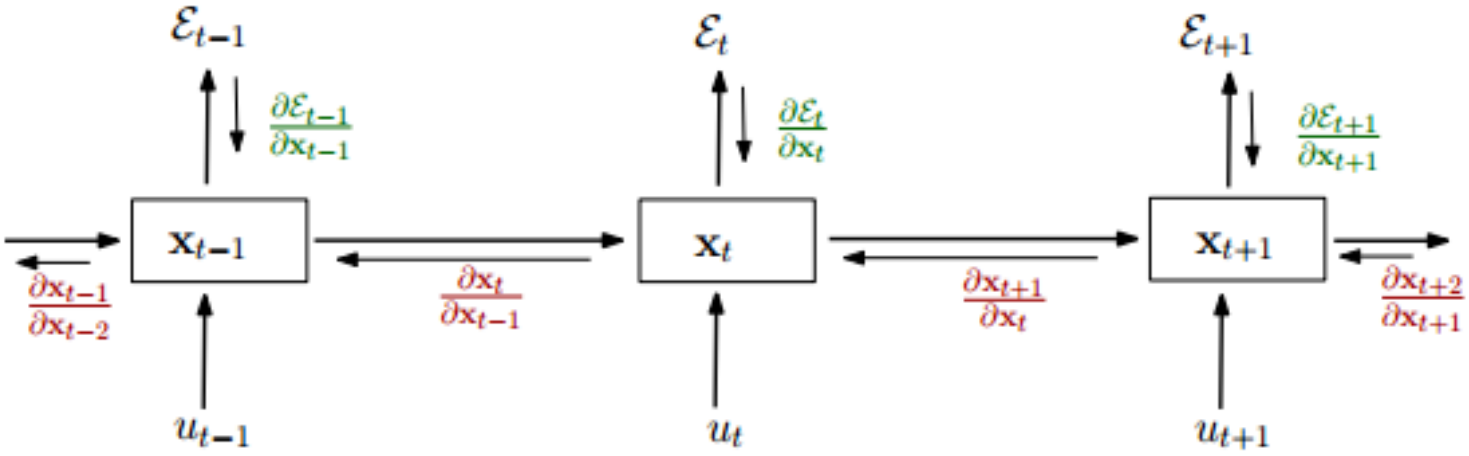
- 利用递归神经网络实现源语言的编码和目标语言的解码



(Sutskever et al., 2014)

递归神经网络的优缺点

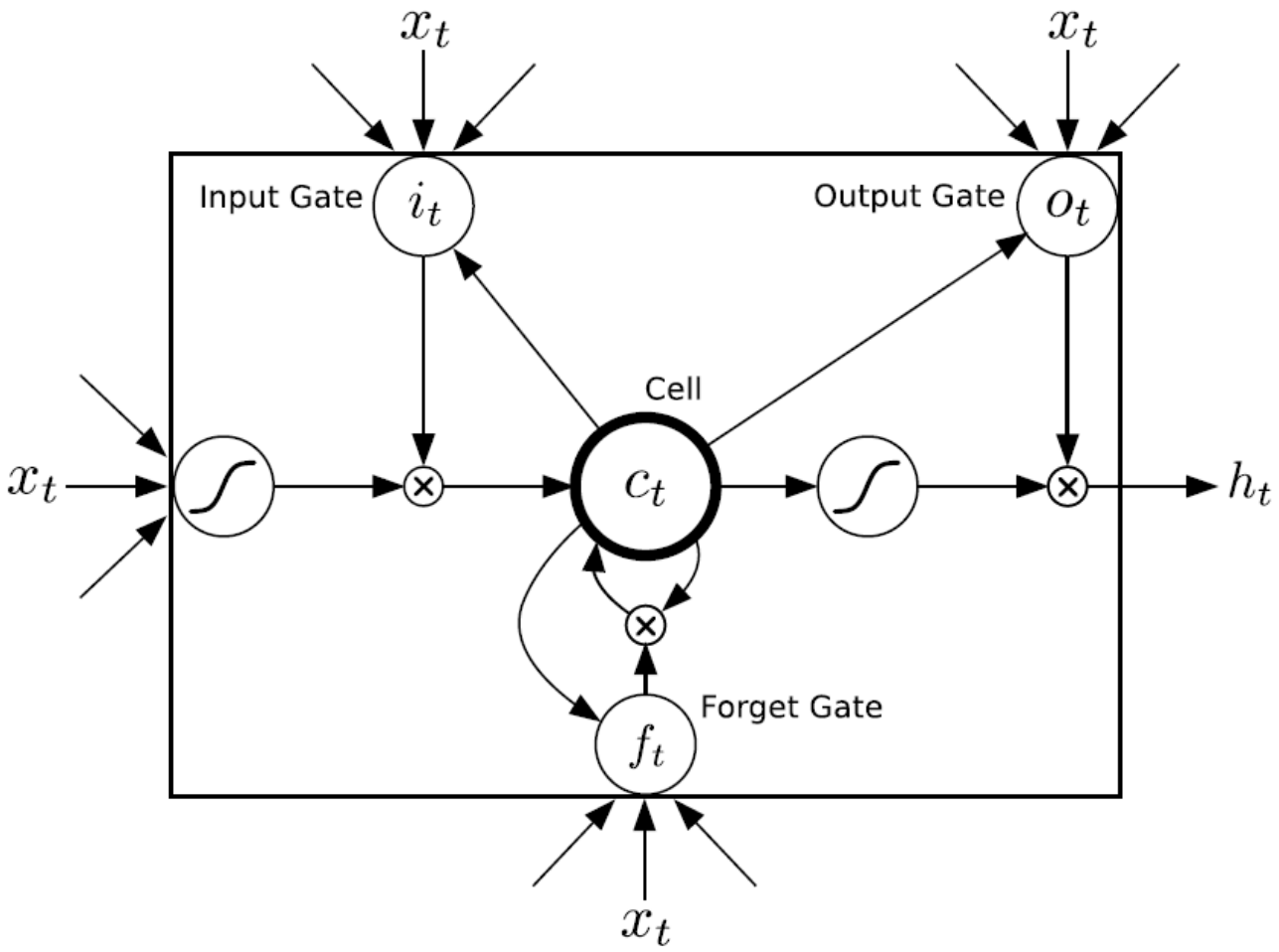
- 优点
  - 适合处理变长线性序列
  - 理论上能够利用无限长的历史信息
- 缺点
  - “梯度消失”或“梯度爆炸”



(Pascanu et al., 2013)

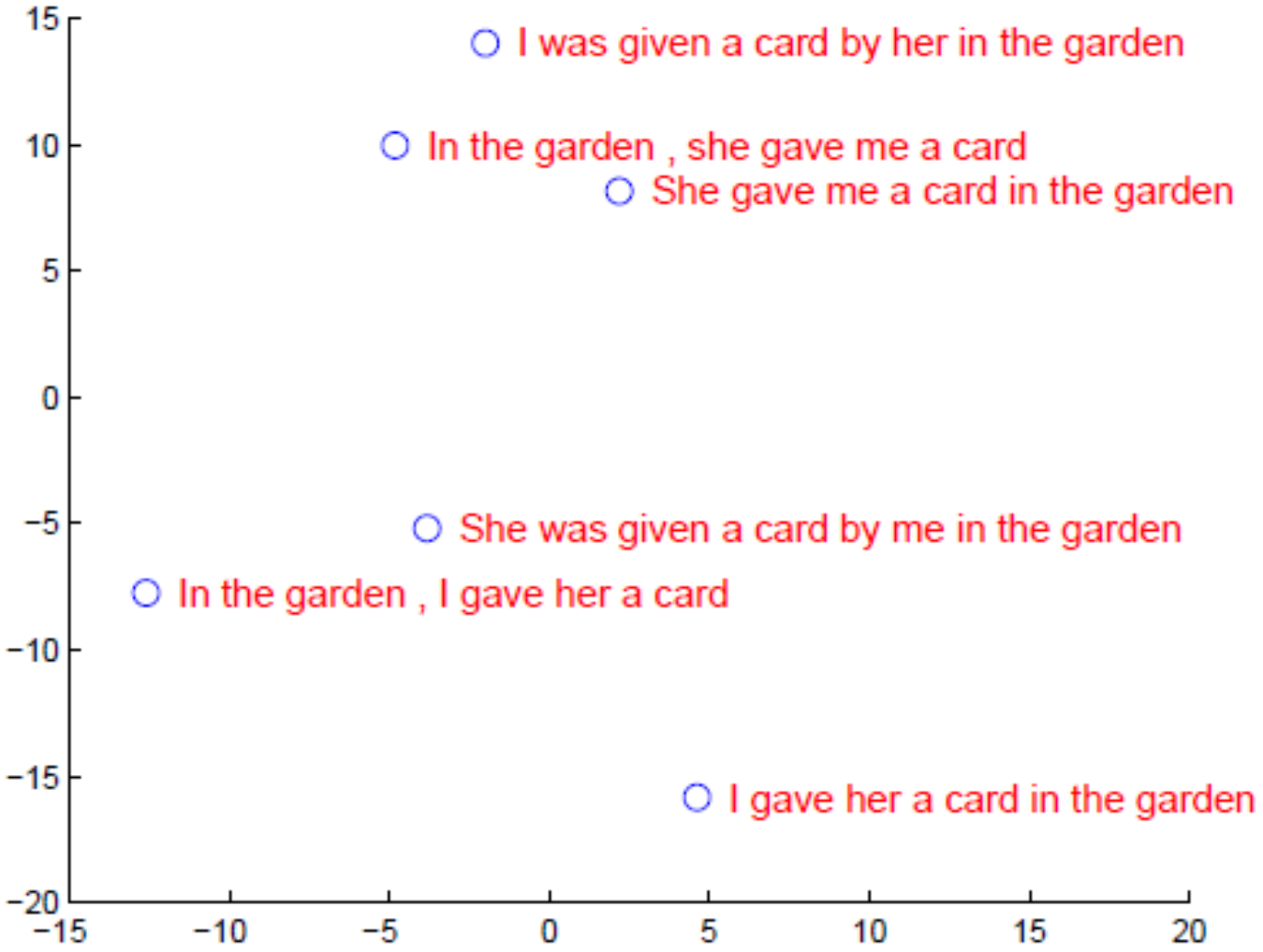
长短时记忆

- 通过门阀技术缓解“梯度消失”和“梯度爆炸”



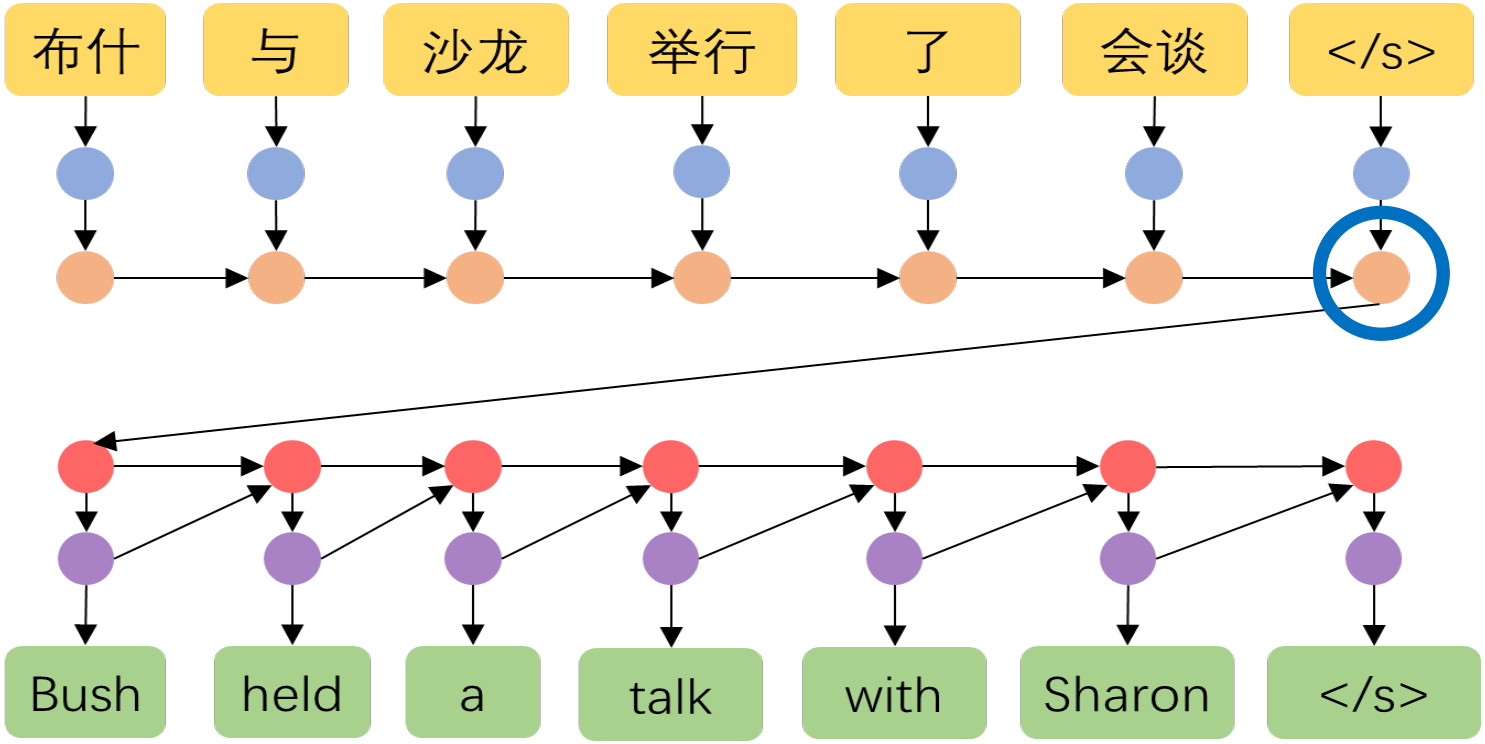
(Hochreiter and Schmidhuber, 1997)

神经网络学到了什么？



(Sutskever et al., 2014)

编码器–解码器架构的优缺点

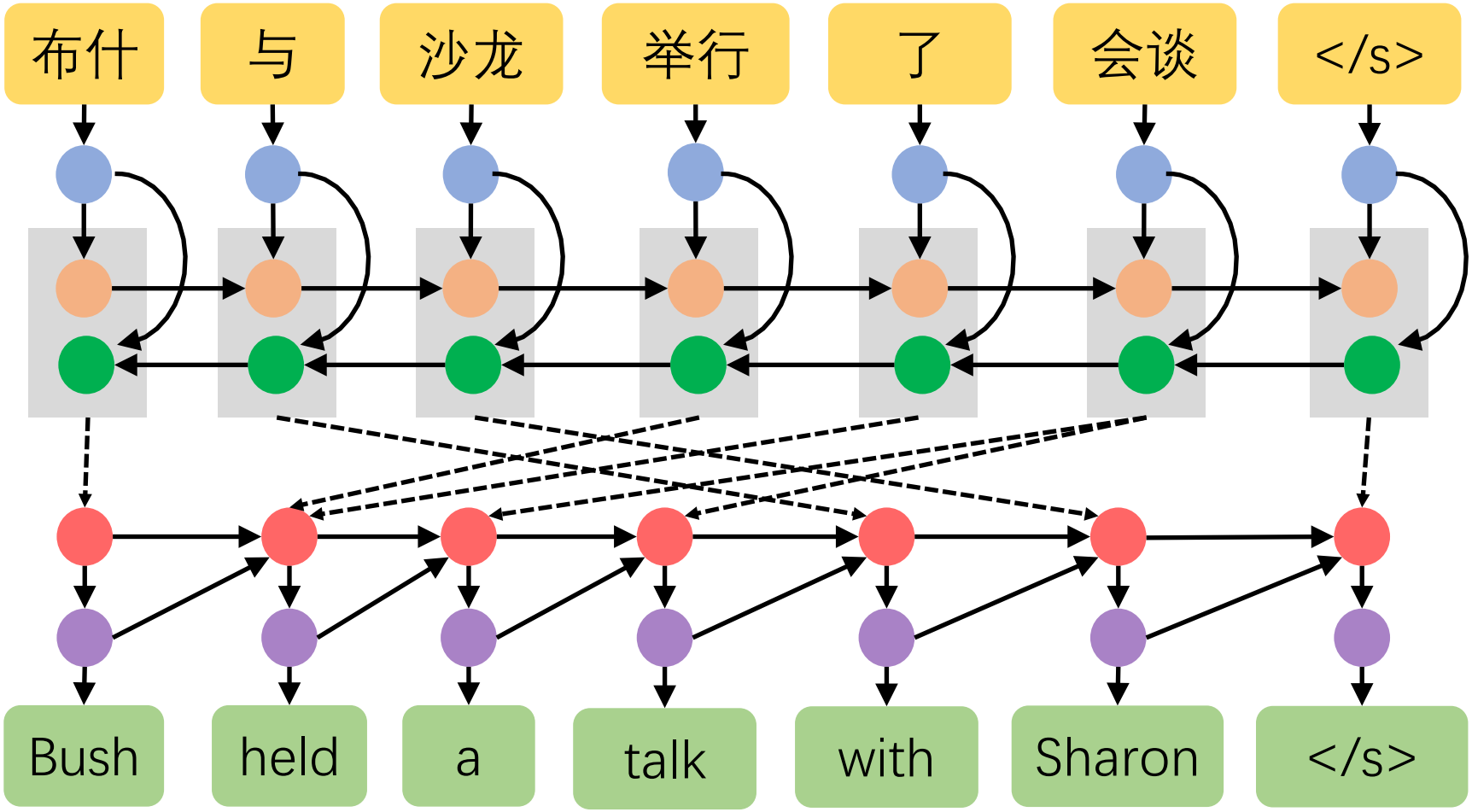


- 优点：利用长短时记忆处理长距离依赖
- 缺点：任意长度的句子都编码为固定维度的向量

(Sutskever et al., 2014)

基于注意力的神经机器翻译

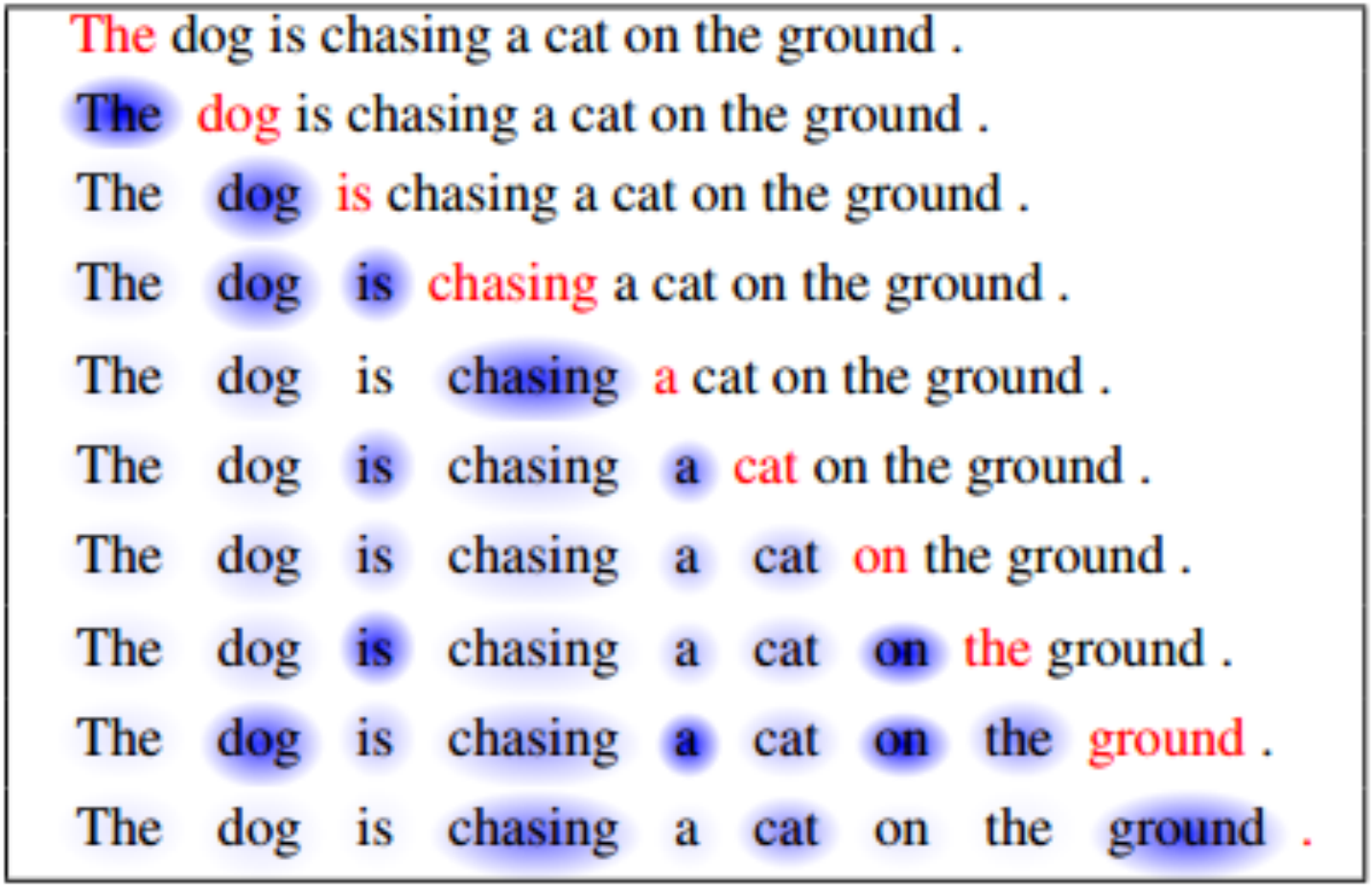
- 利用注意力机制动态计算源语言端相关上下文



(Bahdanau et al., 2015)

注意力

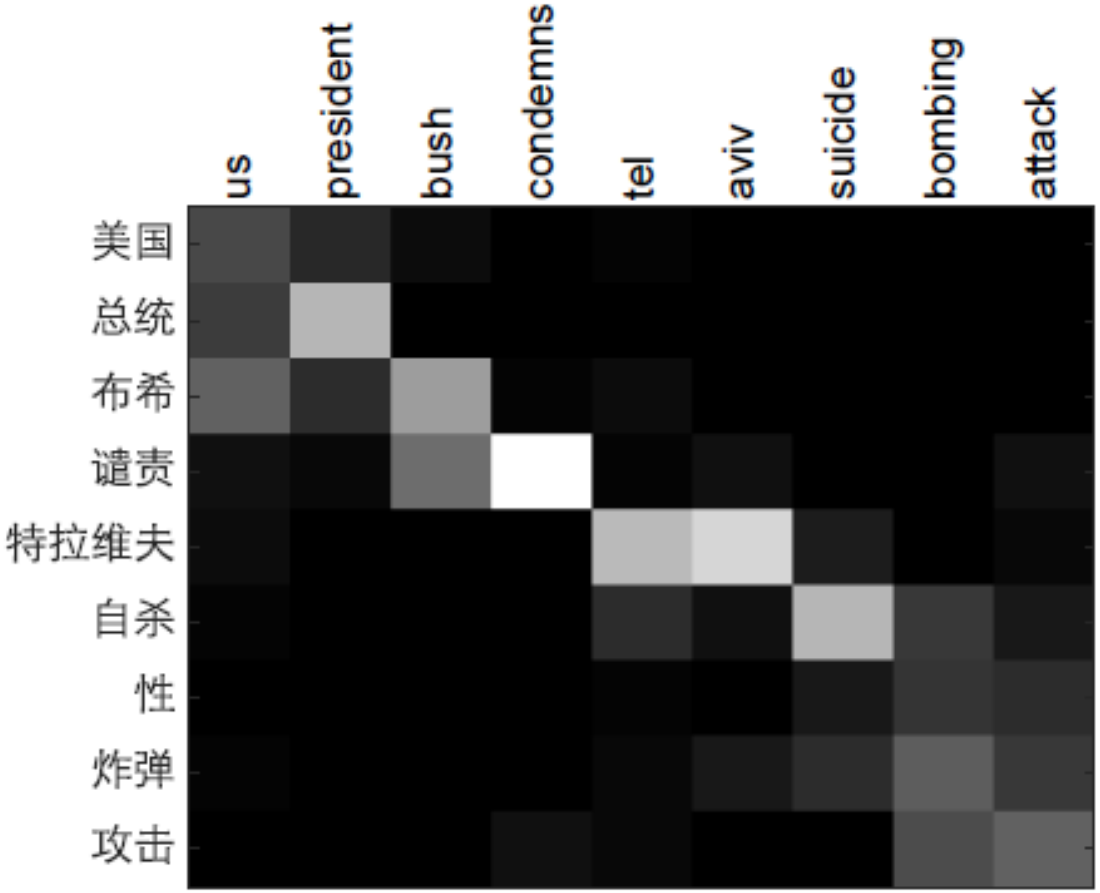
- 思想：集中关注影响当前词的上下文



(Cheng et al., 2016a)

神经机器翻译中的注意力

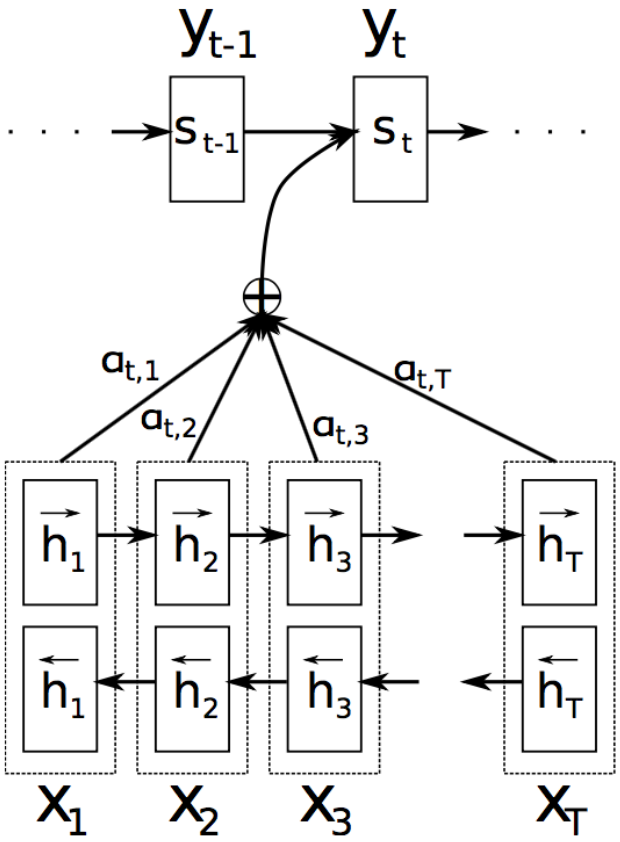
- 源语言词语目标语言词的关联强度



(Bahdanau et al., 2015; Cheng et al., 2016b)

神经机器翻译中的注意力

- 源语言词语目标语言词的关联强度



$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

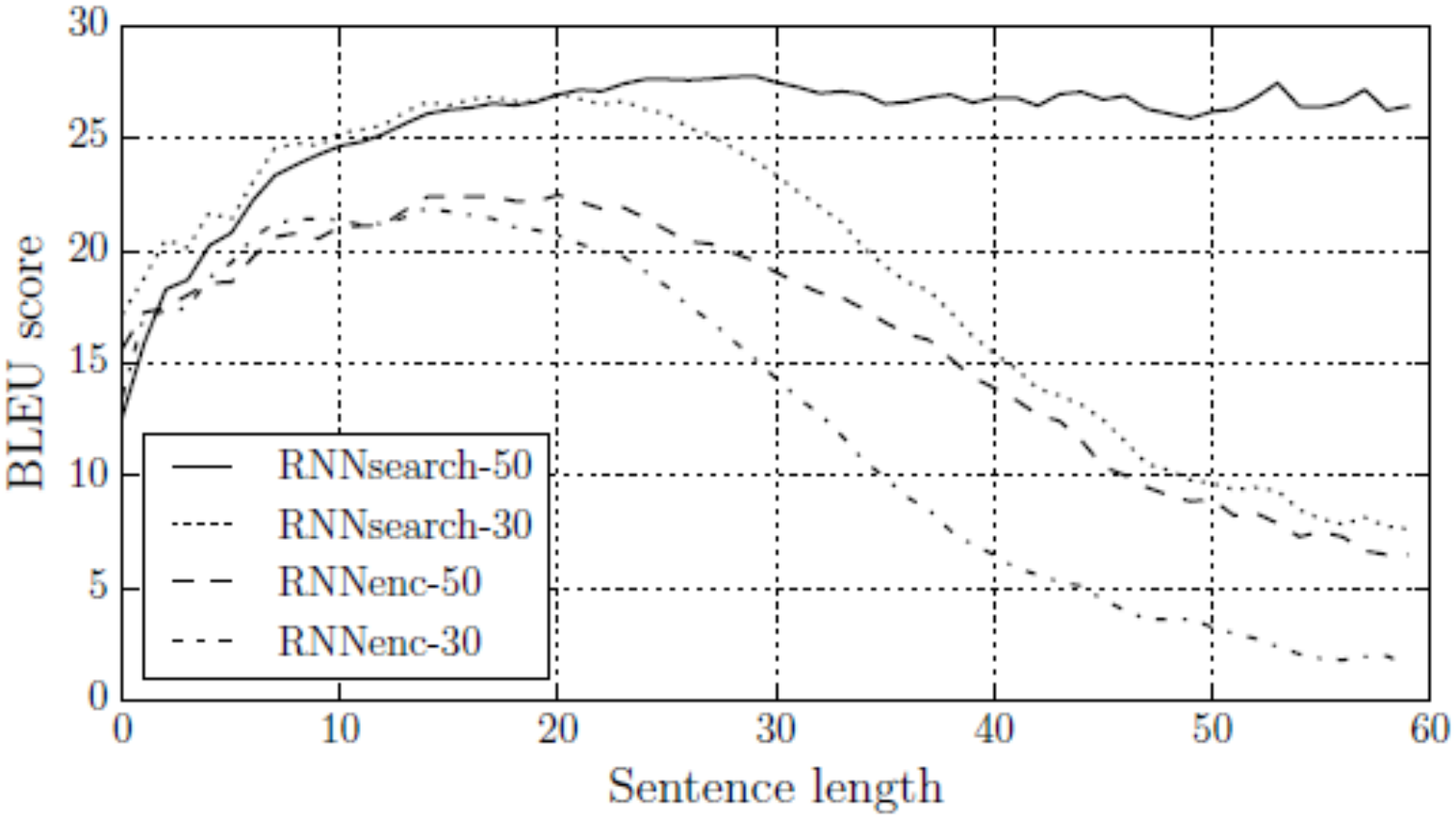
$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

$$e_{ij} = a(s_{i-1}, h_j)$$

(Bahdanau et al., 2015)

注意力机制提升长句翻译效果



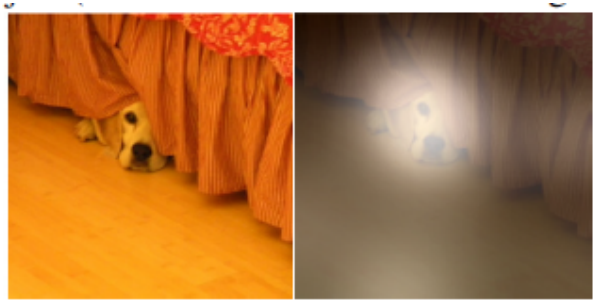
RNNenc: 固定源语言上下文，**RNNsearch**: 动态源语言上下文

(Bahdanau et al., 2015)

注意力机制的其他应用

- 注意力机制已成为深度学习的主流技术

A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

“看图说话”：为图片自动生成文本描述

(Xu et al., 2015)

近期研究进展

- 神经机器翻译在近两年取得飞速发展
  - 受限词汇量 (Luong et al., 2015a ; Jean et al., 2015)
  - 先验约束 (Tu et al., 2016; Cohn et al., 2016)
  - 记忆和隐状态 (Wang et al., 2016; Zhang et al., 2016)
  - 训练准则 (Shen et al., 2016; Ranzato et al., 2016)
  - 单语数据利用 (Cheng et al., 2016c; Sennrich et al., 2016b)
  - 多语言 (Dong et al., 2015; Zoph and Knight, 2016)
  - 多模态 (Duong et al., 2016; Hitschler et al., 2016)

进展1

受限词汇量问题

进展1：受限词汇量

- 受计算复杂度限制，仅能使用有限的词汇量

$$P(\mathbf{y}_n|\mathbf{x},\mathbf{y}_{<n};\boldsymbol{\theta}) = \frac{\exp(\varphi(\mathbf{v}_{\mathbf{y}_n},\mathbf{c}_s,\mathbf{c}_t,\boldsymbol{\theta}))}{\sum_{y\in\mathcal{Y}}\exp(\varphi(\mathbf{v}_y,\mathbf{c}_s,\mathbf{c}_t,\boldsymbol{\theta}))}$$

$\mathbf{v}_y$	目标语言词向量	$\mathcal{Y}$	目标语言词汇
$\mathbf{c}_s$	源语言上下文向量	$\mathbf{c}_t$	目标语言上下文向量

(Luong et al., 2015a; Jean et al., 2015; Li et al., 2016)

进展1：受限词汇量

- 未登录词替换：在后处理阶段单独翻译未登录词

	Sentences
src	An additional *2600* operations including *orthopedic* and *cataract* surgery will help clear a backlog .
trans	En outre , $unkpos_1$ opérations supplémentaires , dont la chirurgie $unkpos_5$ et la $unkpos_6$, permettront de résorber l' arriéré .
+unk	En outre , *2600* opérations supplémentaires , dont la chirurgie *orthopédiques* et la *cataracte* , permettront de résorber l' arriéré .
tgt	2600 opérations supplémentaires , notamment dans le domaine de la chirurgie orthopédique et de la cataracte , aideront à rattraper le retard .

(Luong et al., 2015a)

进展1：受限词汇量

- 采样近似：利用采样近似计算期望

$$\nabla \log p(y_t \mid y_{<t}, x)$$

$$= \nabla \mathcal{E}(y_t) - \sum_{k:y_k \in V} p(y_k \mid y_{<t}, x) \nabla \mathcal{E}(y_k)$$

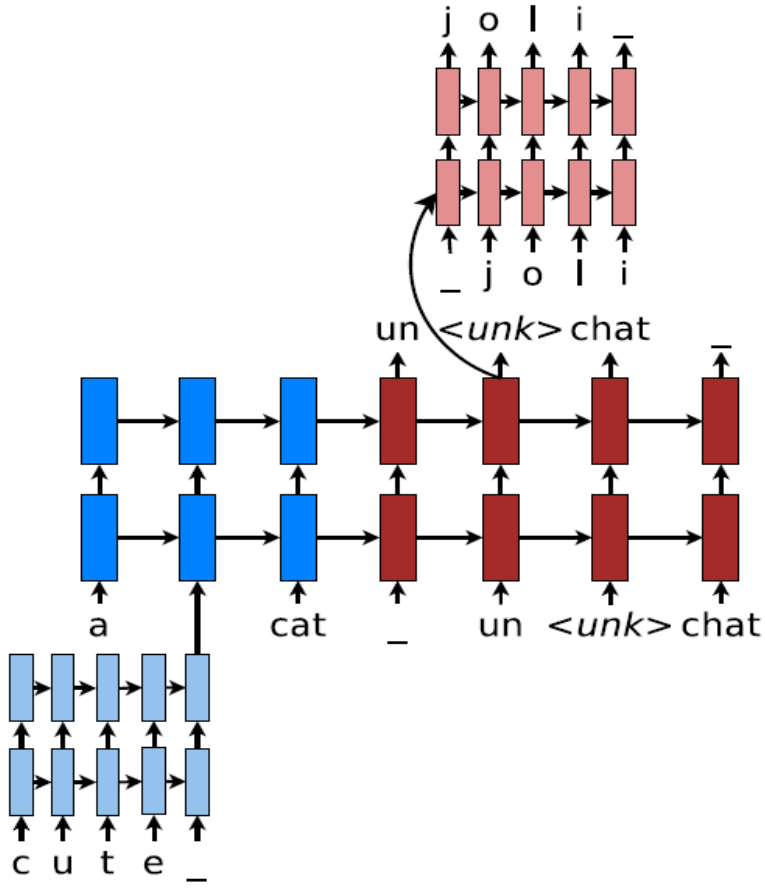
$$\approx \nabla \mathcal{E}(y_t) - \sum_{k:y_k \in V'} \frac{\omega_k}{\sum_{k':y_{k'} \in V'} \omega_{k'}} \nabla \mathcal{E}(y_k)$$

计算采样空间的上的概率分布，近似计算期望

(Luong et al., 2015a)

进展1：受限词汇量

- 基于字母的模型：用细粒度意义单元降低词汇量

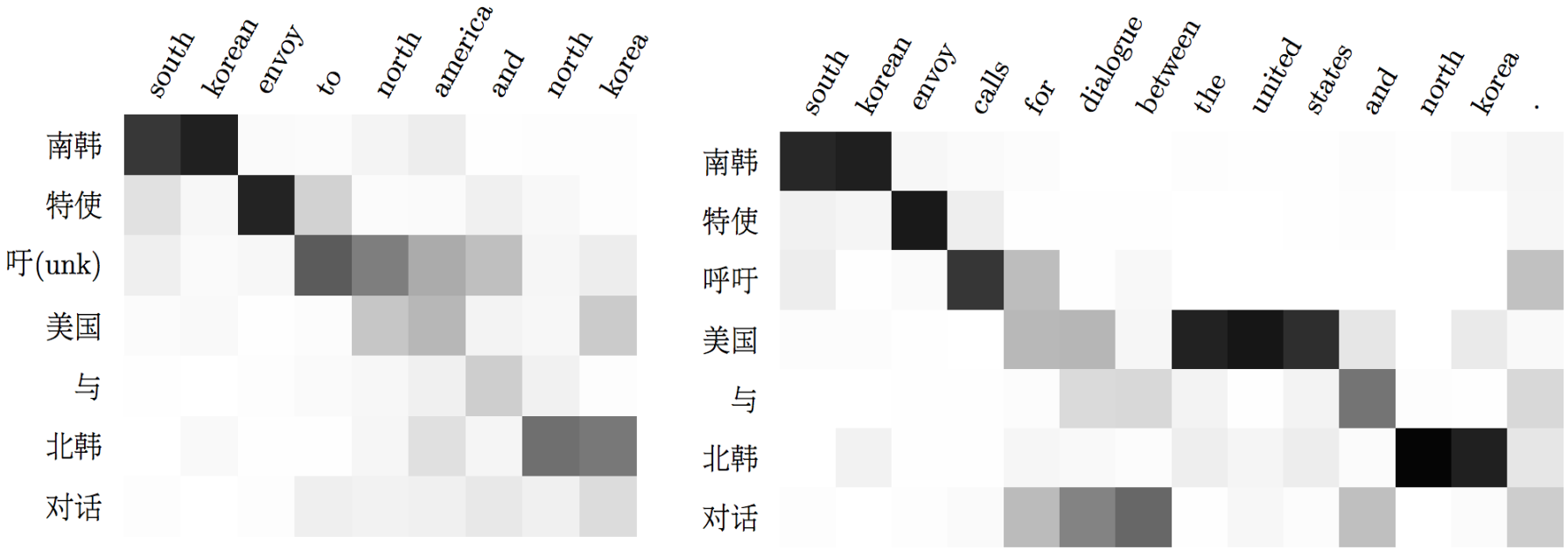


字母与词语的混合模型，
词语模型解决常用词翻译，
字母模型解决生僻词翻译

(Luong and Manning, 2016)

进展1：受限词汇量

- 相似词替换：用相似词代替未登录词进行训练



翻译结束后再局部还原为未登录词的译文

(Li et al., 2016)

进展2

增加先验约束

进展2：先验约束

- 如何利用先验知识约束神经机器翻译？

类型	示例
双语词典	“中国” 一般被翻译成 “China”
繁殖率	“白宫” 一般被翻译成两个英文词
覆盖	源文中每个词大多被只被翻译一次
结构差异	中文 “VP+PP” ， 英文则 “PP+VP”

(Tu et al., 2016; Cohn et al., 2016; Cheng et al., 2016b; Shi et al., 2016)

进展2：先验约束

- 基于覆盖率的神经机器翻译

输入	很多 机场 都 被迫 关闭 了
不考虑覆盖度	Many airports were closed to close
考虑覆盖度	Many airports were forced to close down

先验：不应重复翻译，也不应漏翻

(Tu et al., 2016)

进展2：先验约束

- 在注意力机制中嵌入结构化约束

约束：源语言和目标语言句子之间大致是对角线对齐

$$f_{ji} = \boldsymbol{v}^\top \tanh\left(\mathbf{W}^{(\text{ae})}\boldsymbol{e}_i + \mathbf{W}^{(\text{ah})}\boldsymbol{g}_{j-1} + \mathbf{W}^{(\text{ap})}\psi(j,i,I)\right)$$

$$\boldsymbol{\alpha}_j = \text{softmax}\left(\boldsymbol{f}_j\right)$$

$$\boldsymbol{c}_j = \sum_i \alpha_{ji}\boldsymbol{e}_i$$

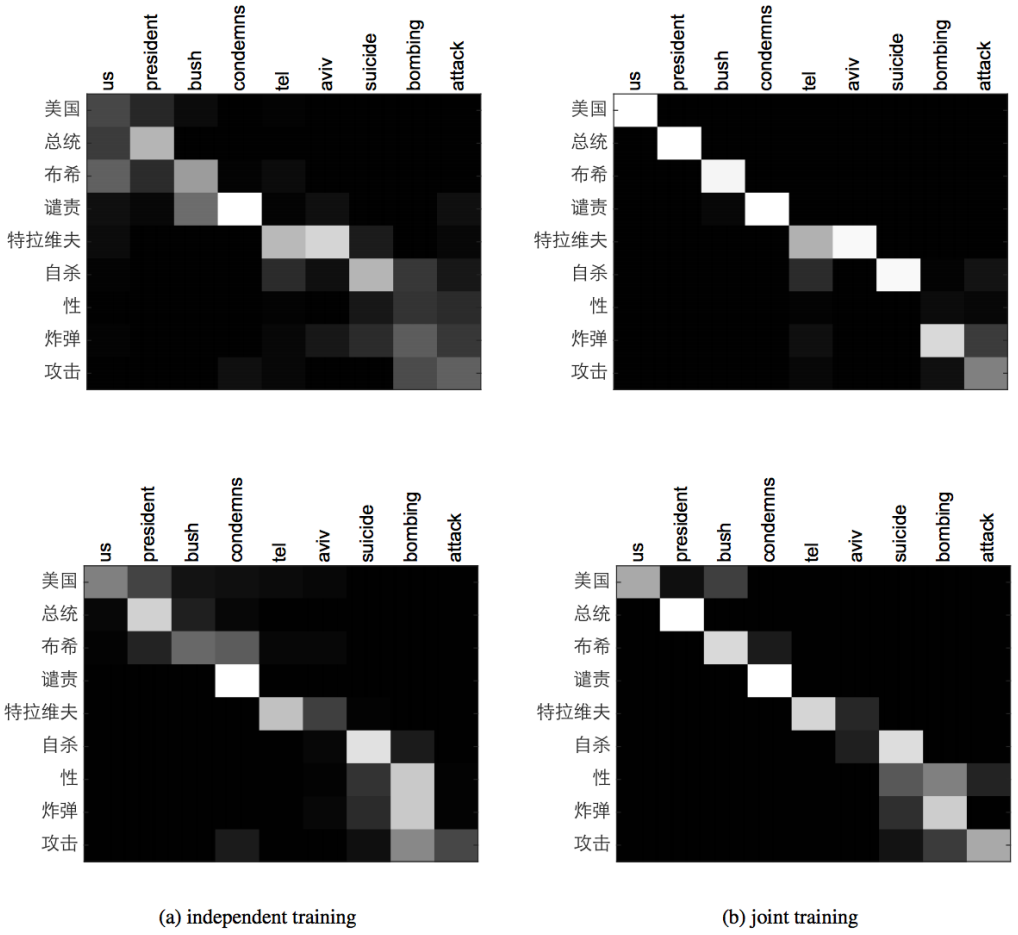
其中

$$\psi(j,i,I) = \left[\log(1+j), \log(1+i), \log(1+I)\right]^\top$$

(Cohn et al., 2016)

进展2：先验约束

- 一致性训练：正向和反向翻译模型具有互补性



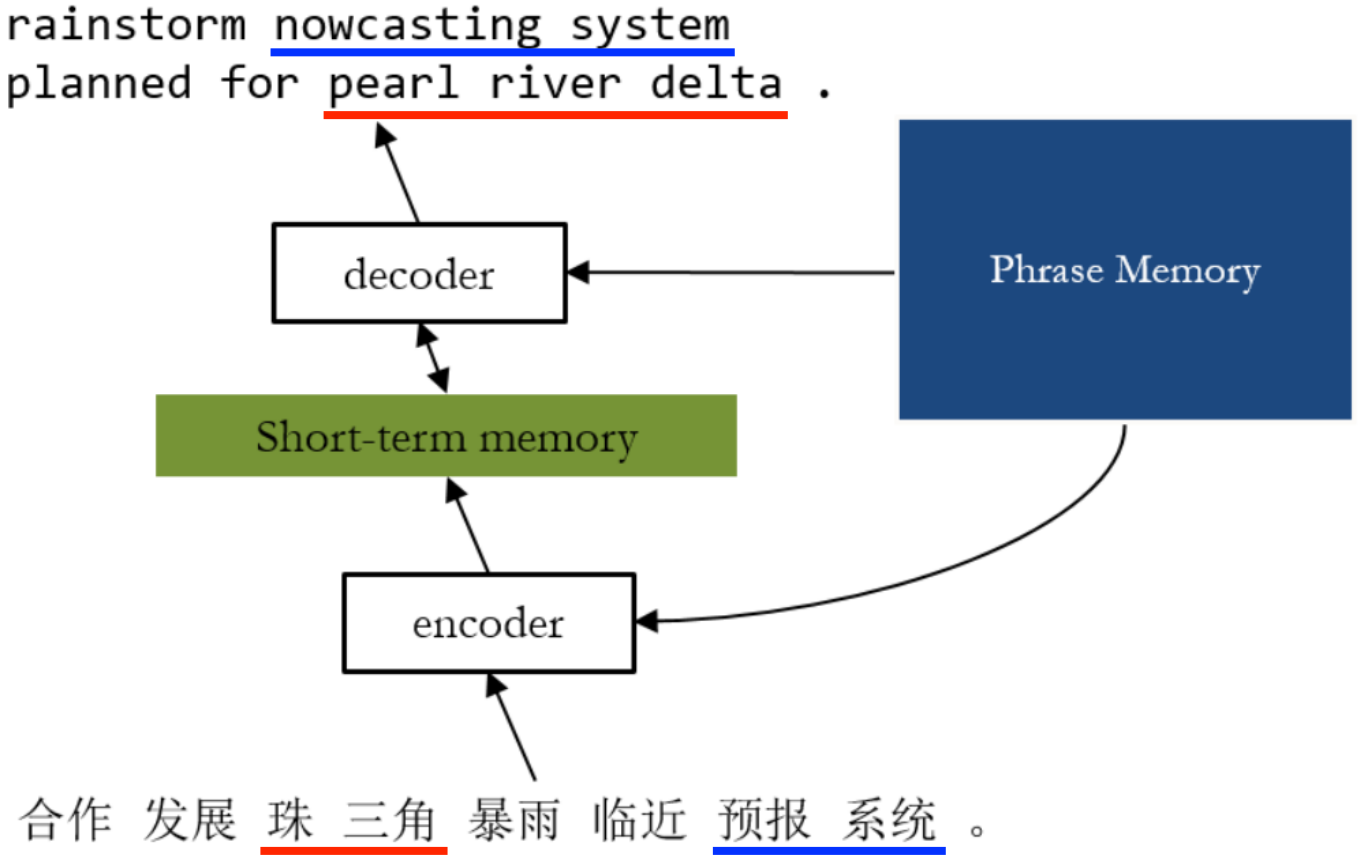
(a) independent training

(b) joint training

(Cheng et al., 2016)

进展2：先验约束

- 短语记忆：利用传统的短语表提高NMT



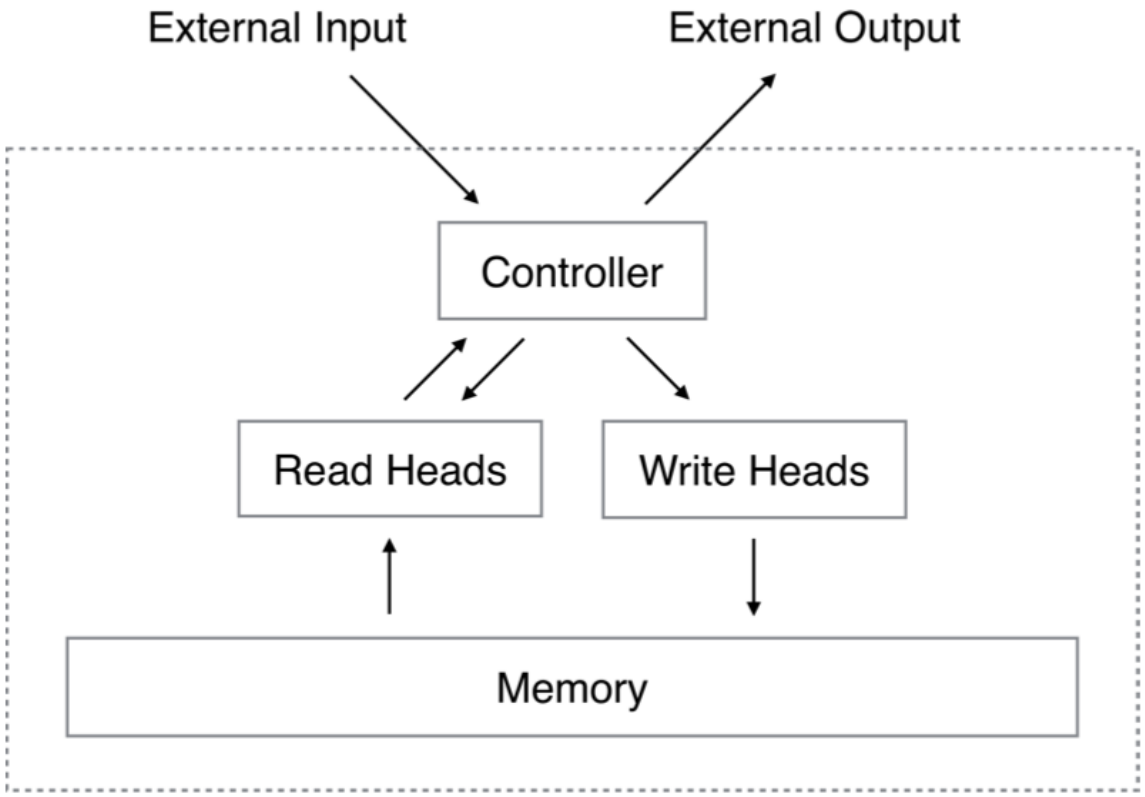
(Tang et al., 2016)

进展3

新机制：记忆和隐状态

进展3：记忆和隐状态

- 神经网络图灵机



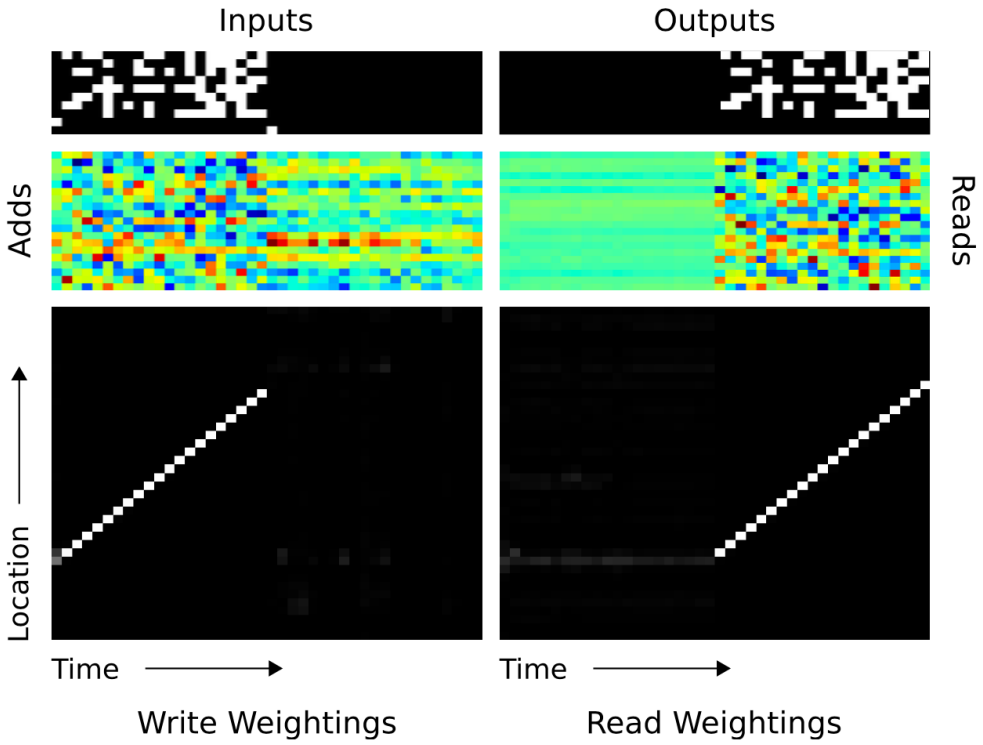
门阀 => 注意力 => 记忆？

(Graves et al., 2014)

进展3：记忆和隐状态

- 神经网络图灵机

序列拷贝：　　3 1 5 4 2　　⇨　　3 1 5 4 2

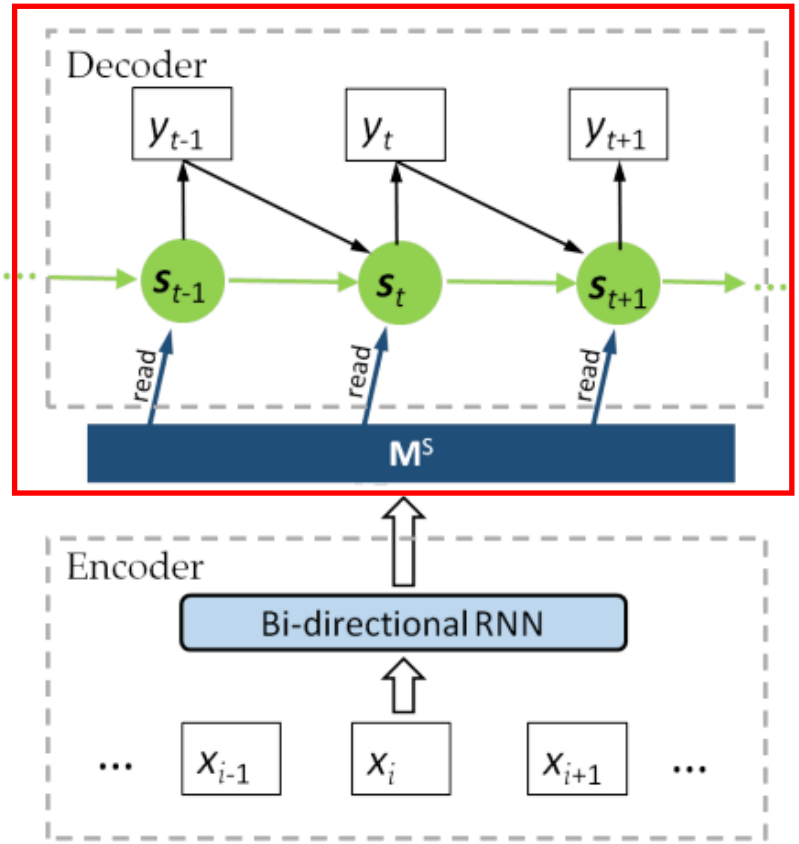


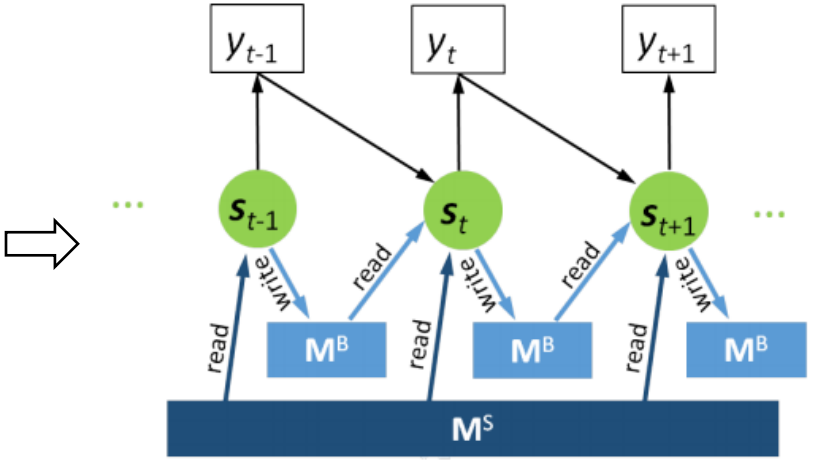
initialise: move head to start location
while input delimiter not seen **do**
    receive input vector
    write input to head location
    increment head location by 1
end while
return head to start location
while true **do**
    read output vector from head location
    emit output
    increment head location by 1
end while

(Graves et al., 2014)

进展3：记忆和隐状态

- 利用记忆机制提高神经机器翻译





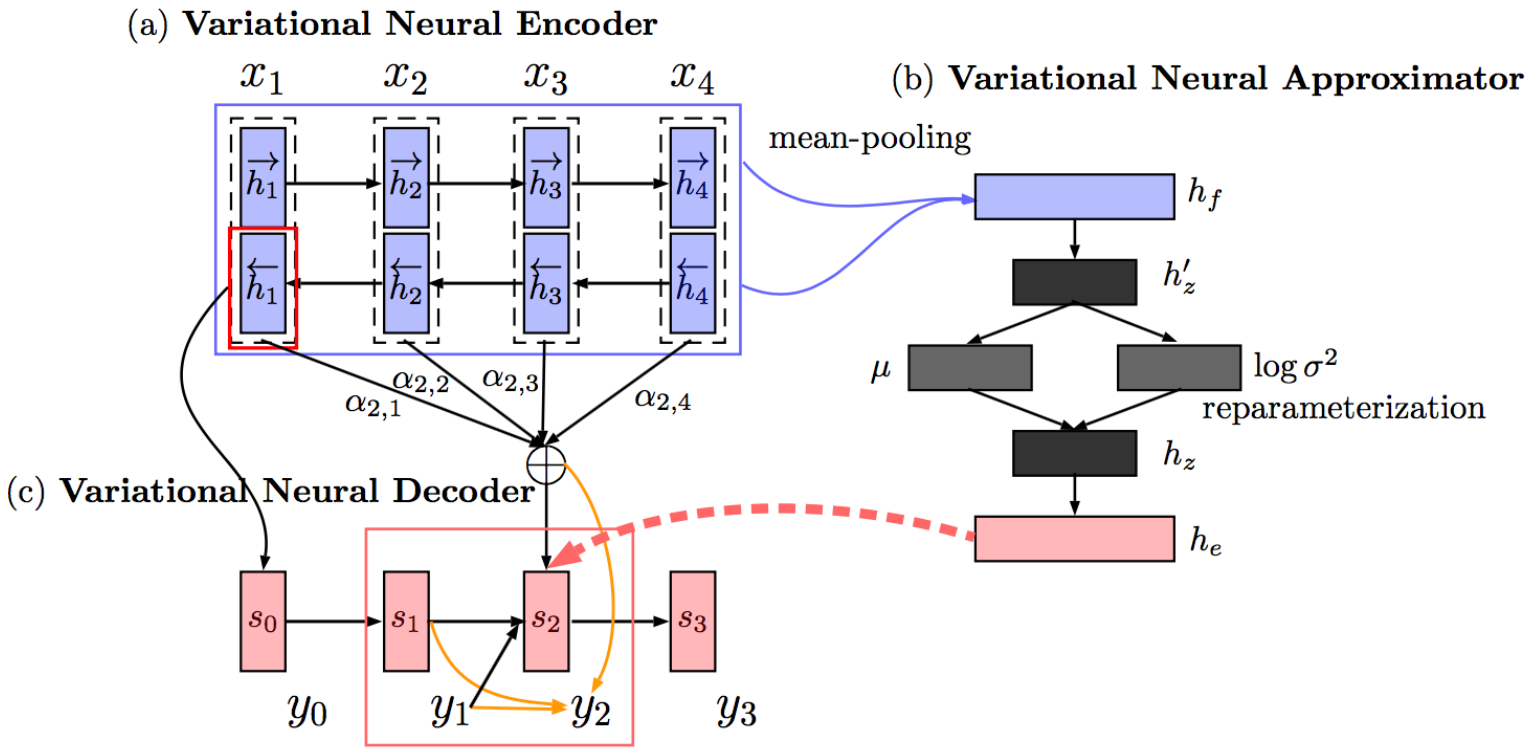
将“外存”引入神经机器翻译

(Wang et al., 2016)

进展3：记忆和隐状态

- 变分神经机器翻译

$$p(\mathbf{y}|\mathbf{x}) = \sum_z p(\mathbf{y}, z|\mathbf{x}) = \sum_z p(\mathbf{y}|z, \mathbf{x})p(z)$$



(Zhang et al., 2016)

进展4

针对评价指标的训练准则

进展4：训练准则

- 标准的极大似然估计面临挑战

训练数据 $\{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$

训练目标 $$\mathcal{L}(\boldsymbol{\theta}) = \sum_{s=1}^{S} \log P(\mathbf{y}^{(s)} | \mathbf{x}^{(s)}; \boldsymbol{\theta})$$

$$= \sum_{s=1}^{S} \sum_{n=1}^{N^{(s)}} \log P(\mathbf{y}_n^{(s)} | \mathbf{x}^{(s)}, \mathbf{y}_{<n}^{(s)}; \boldsymbol{\theta})$$

优化 $$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \left\{ \mathcal{L}(\boldsymbol{\theta}) \right\}$$

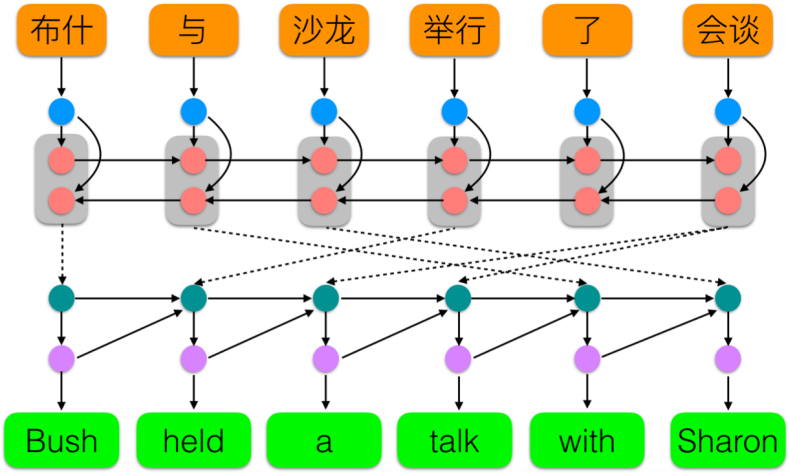
挑战：exposure bias问题和词级损失函数

(Shen et al., 2016; Rezanto et al., 2016)

进展4：训练准则

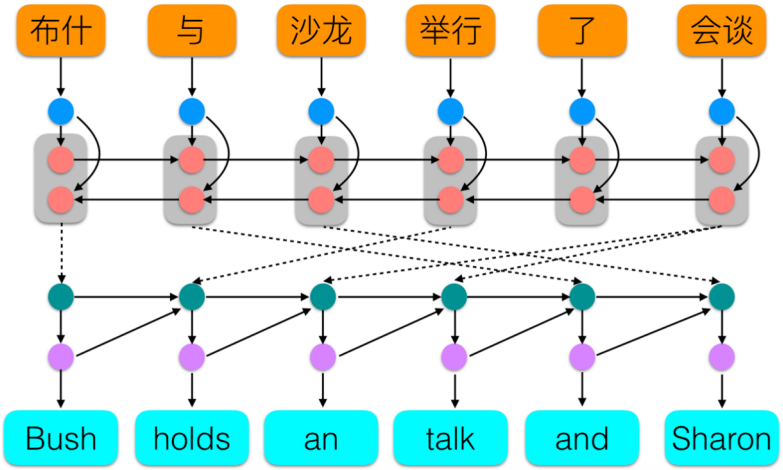
- 极大似然估计面临“exposure bias”问题

训练



生成目标词基于来自
观测数据的上下文

测试

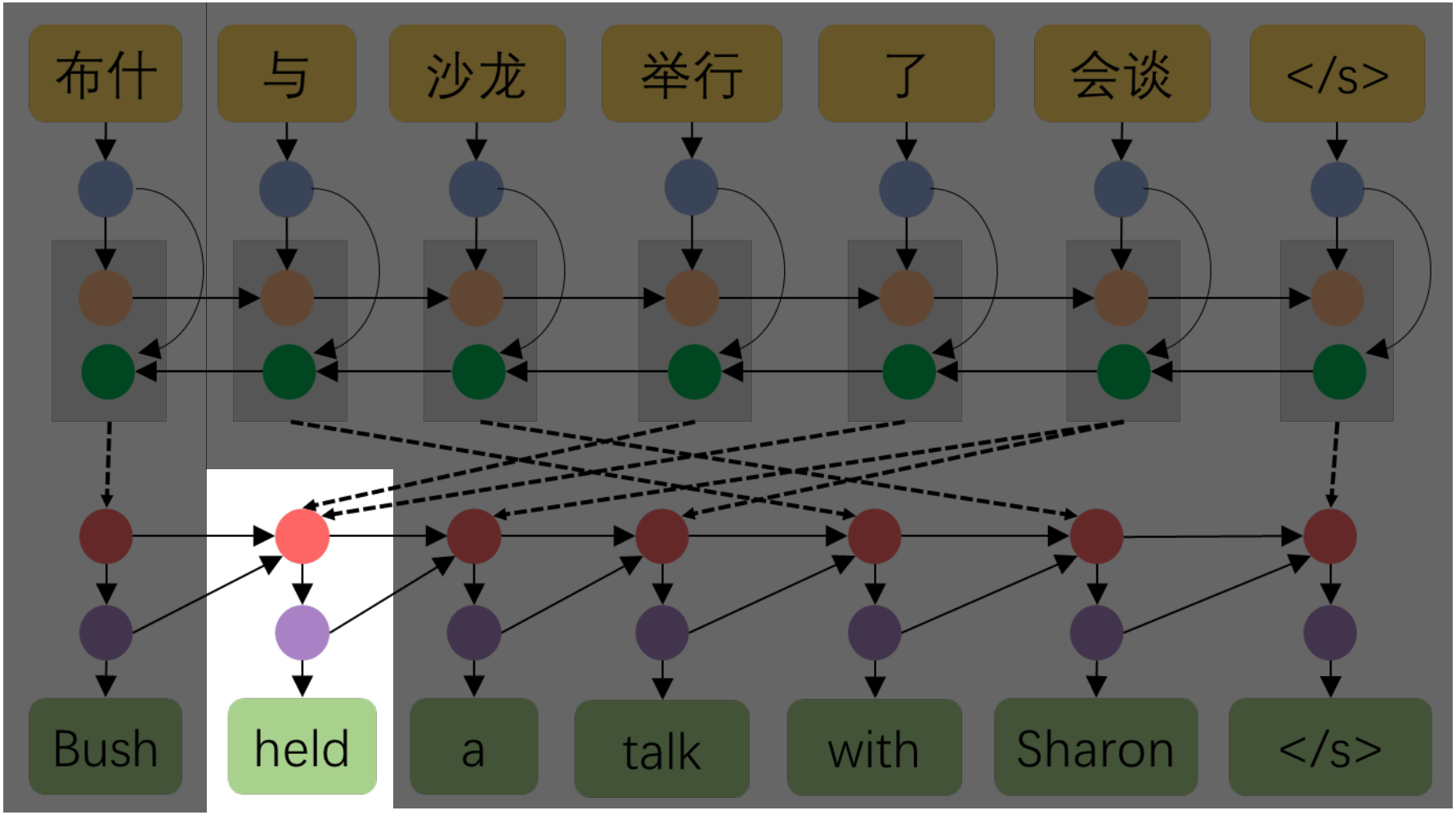


生成目标词基于来自
模型预测的上下文

(Rezanto et al., 2016)

进展4：训练准则

- 极大似然估计仅使用词级损失函数



(Rezanto et al., 2016)

进展4：训练准则

- MIXER：利用增强学习针对评价指标优化模型

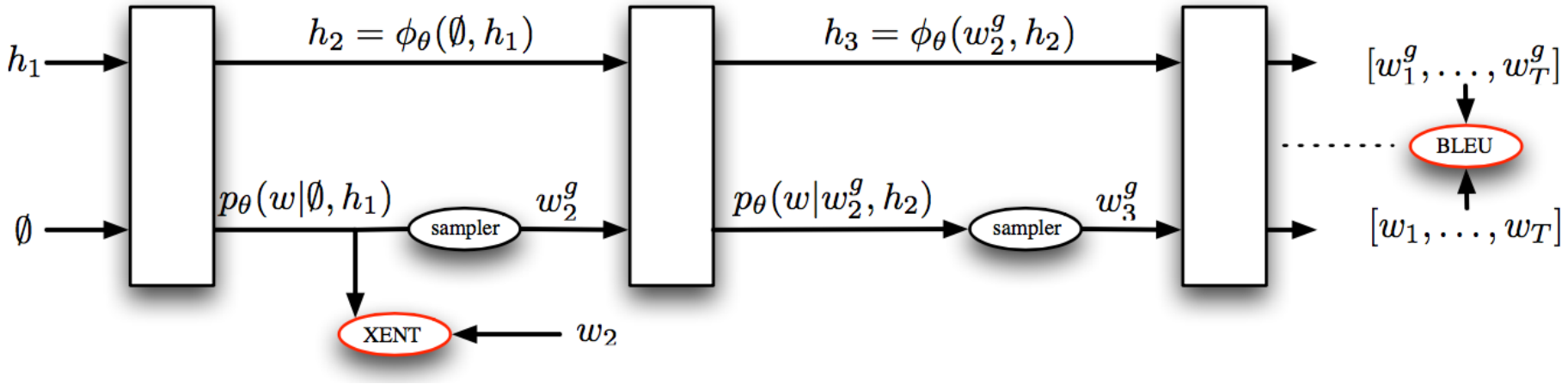


Figure 4: Illustration of MIXER. In the first s unrolling steps (here $s = 1$), the network resembles a standard RNN trained by XENT. In the remaining steps, the input to each module is a sample from the distribution over words produced at the previous time step. Once the end of sentence has been reached (or the maximum sequence length), a BLEU reward is computed. REINFORCE is then used to backpropagate the gradients, even through the sequence of samplers. We employ an annealing schedule on s, starting with s equal to the maximum sequence length T and finishing with $s = 1$.

(Rezanto et al., 2016)

进展4：训练准则

- 最小风险训练：针对评价指标训练神经网络

训练数据 $\{\langle \mathbf{x}^{(s)}, \mathbf{y}^{(s)} \rangle\}_{s=1}^{S}$

训练目标 $$\mathcal{R}(\boldsymbol{\theta}) = \sum_{s=1}^{S} \mathbb{E}_{\mathbf{y}|\mathbf{x}^{(s)};\boldsymbol{\theta}} \left[\Delta(\mathbf{y}, \mathbf{y}^{(s)}) \right]$$

$$= \sum_{s=1}^{S} \sum_{\mathbf{y} \in \mathcal{Y}(\mathbf{x}^{(s)})} P(\mathbf{y}|\mathbf{x}^{(s)};\boldsymbol{\theta}) \Delta(\mathbf{y}, \mathbf{y}^{(s)})$$

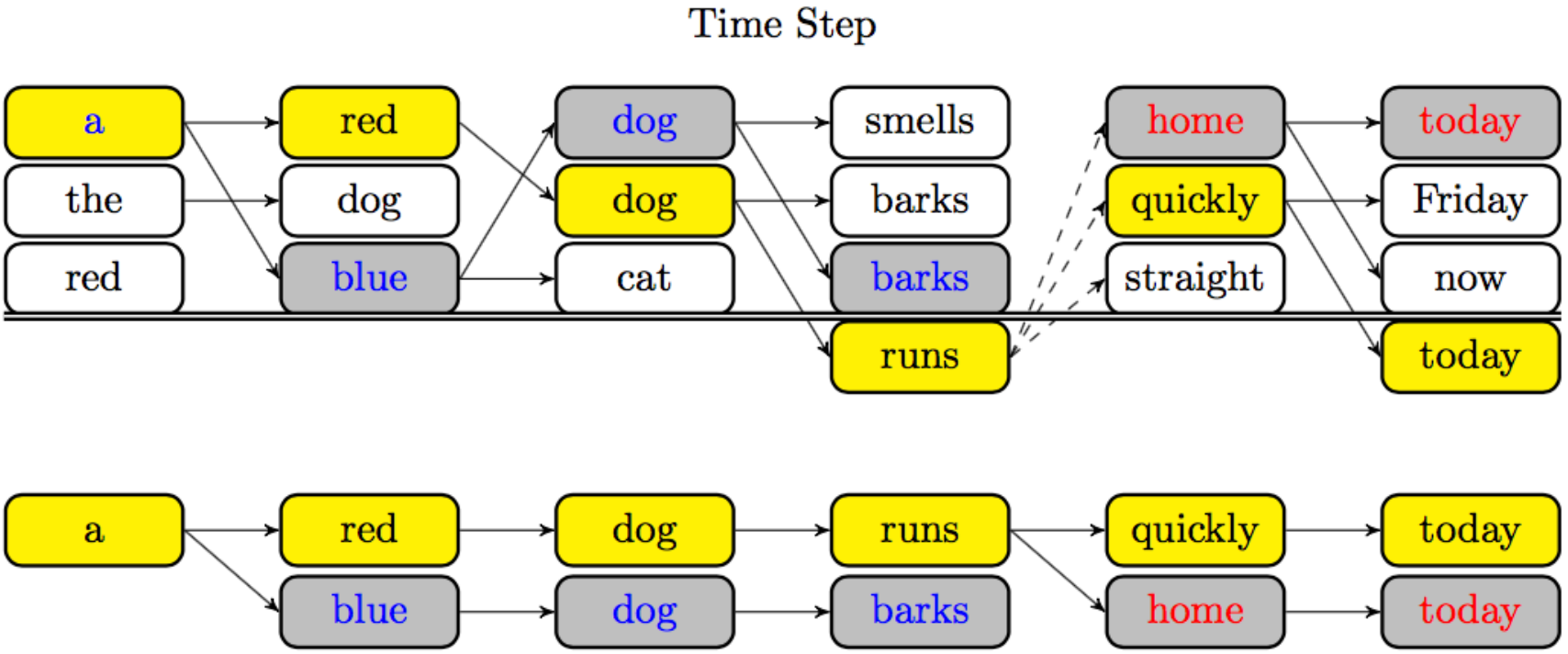
优化 $$\hat{\boldsymbol{\theta}}_{\text{MRT}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \left\{ \mathcal{R}(\boldsymbol{\theta}) \right\}$$

通用性：适用于任意架构和任意损失函数

(Shen et al., 2016)

进展4：训练准则

- 柱搜索优化：引入基于搜索的损失函数



$$\mathcal{L}(f) = \sum_{t=1}^{T} \Delta(\hat{y}_{1:t}^{(K)}) \left[1 - f(y_t, \boldsymbol{h}_{t-1}) + f(\hat{y}_t^{(K)}, \hat{\boldsymbol{h}}_{t-1}^{(K)})\right]$$

(Wiseman and Rush, 2016)

进展5

单语数据利用

进展5：单语数据利用

- 神经机器翻译只使用平行语料库进行训练

双语语料库的规模、质量和覆盖面都非常受限

(Cheng et al., 2016c; Sennrich et al., 2016b)

进展5：单语数据利用

- 集成语言模型：修改网络架构，嵌入语言模型

浅层集成

$$\log p(\mathbf{y}_t = k) = \log p_{\text{TM}}(\mathbf{y}_t = k) + \beta \log p_{\text{LM}}(\mathbf{y}_t = k)$$

深层集成

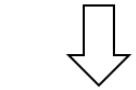
$$p(\mathbf{y}_t|\mathbf{y}_{<t}, \mathbf{x}) \propto \exp(\mathbf{y}_t^\top(\mathbf{W}_o \mathbf{f}_o(\mathbf{s}_t^{\text{LM}}, \mathbf{s}_t^{\text{TM}}, \mathbf{y}_{t-1}, \mathbf{c}_t) + \mathbf{b}_o))$$

(Gulcehre et al., 2015)

进展5：单语数据利用

- 伪数据：翻译单语语料库，构造伪平行语料库

单语 布什 与 沙龙 举行 了 会谈



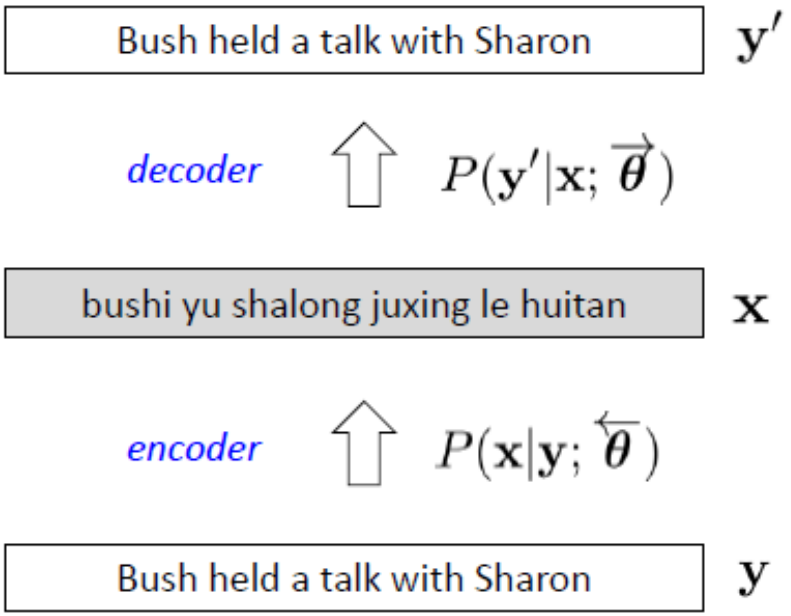
机器翻译系统

译文 Bush held a talk with Sharon

(Sennrich et al., 2016b)

进展5：单语数据利用

- 半监督学习：同时利用平行语料库和单语语料库



$$P(\mathbf{y}'|\mathbf{y};\overrightarrow{\theta},\overleftarrow{\theta}) = \sum_{\mathbf{x}} P(\mathbf{y}',\mathbf{x}|\mathbf{y};\overrightarrow{\theta},\overleftarrow{\theta}) = \sum_{\mathbf{x}} \underbrace{P(\mathbf{x}|\mathbf{y};\overleftarrow{\theta})}_{encoder}\underbrace{P(\mathbf{y}'|\mathbf{x};\overrightarrow{\theta})}_{decoder}$$

自动编码器：正向模型编码，反向模型解码

(Cheng et al., 2016c)

进展6

多语言翻译

进展6：多语言

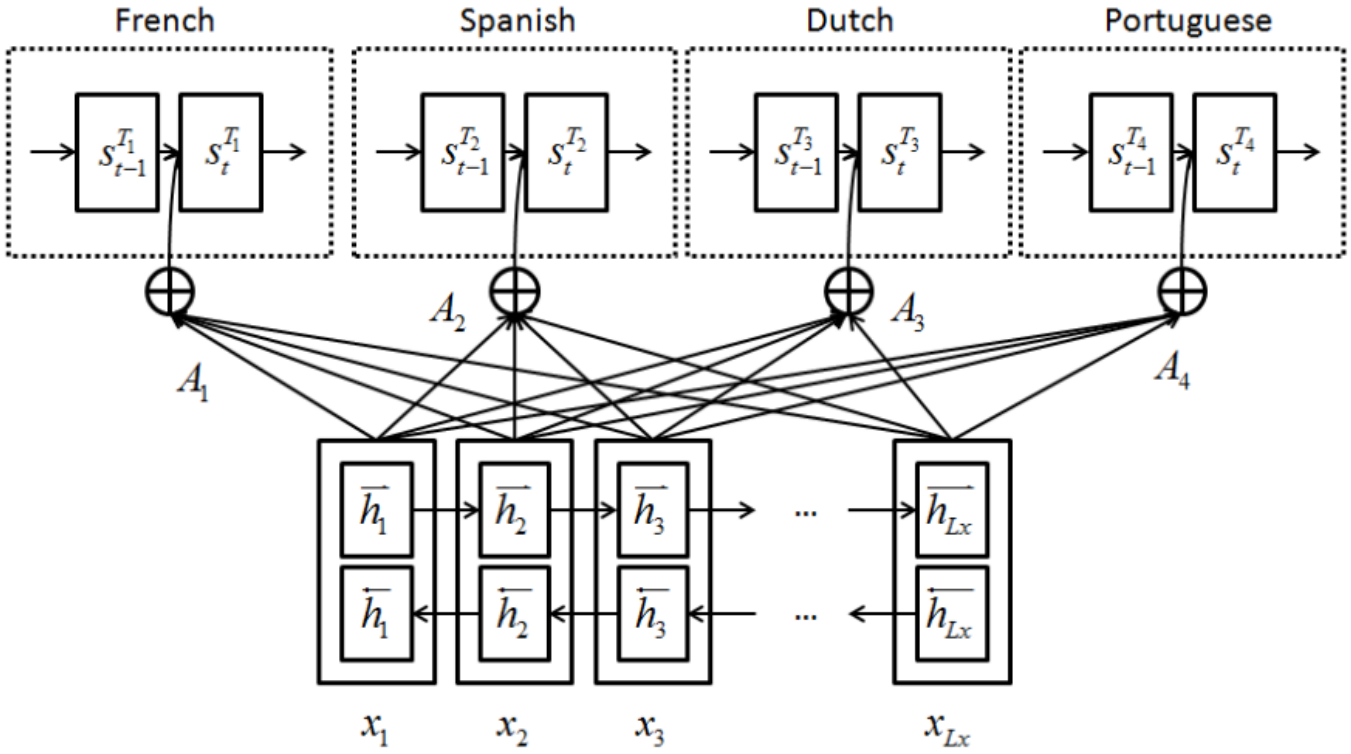
- 各种语言对之间的资源丰富程度迥异，如何有效利用？

Parallel Corpus (L1-L2)	Sentences	L1 Words	English Words
Bulgarian-English	406,934	-	9,886,291
Czech-English	646,605	12,999,455	15,625,264
Danish-English	1,968,800	44,654,417	48,574,988
German-English	1,920,209	44,548,491	47,818,827
Greek-English	1,235,976	-	31,929,703
Spanish-English	1,965,734	51,575,748	49,093,806
Estonian-English	651,746	11,214,221	15,685,733
Finnish-English	1,924,942	32,266,343	47,460,063
French-English	2,007,723	51,388,643	50,196,035
Hungarian-English	624,934	12,420,276	15,096,358
Italian-English	1,909,115	47,402,927	49,666,692
Lithuanian-English	635,146	11,294,690	15,341,983
Latvian-English	637,599	11,928,716	15,411,980
Dutch-English	1,997,775	50,602,994	49,469,373
Polish-English	632,565	12,815,544	15,268,824
Portuguese-English	1,960,407	49,147,826	49,216,896
Romanian-English	399,375	9,628,010	9,710,331
Slovak-English	640,715	12,942,434	15,442,233
Slovene-English	623,490	12,525,644	15,021,497
Swedish-English	1,862,234	41,508,712	45,703,795

Europarl多语言平行语料库 (Koehn, 2005)

进展6：多语言

- 多任务学习：同时训练多种语言对翻译模型

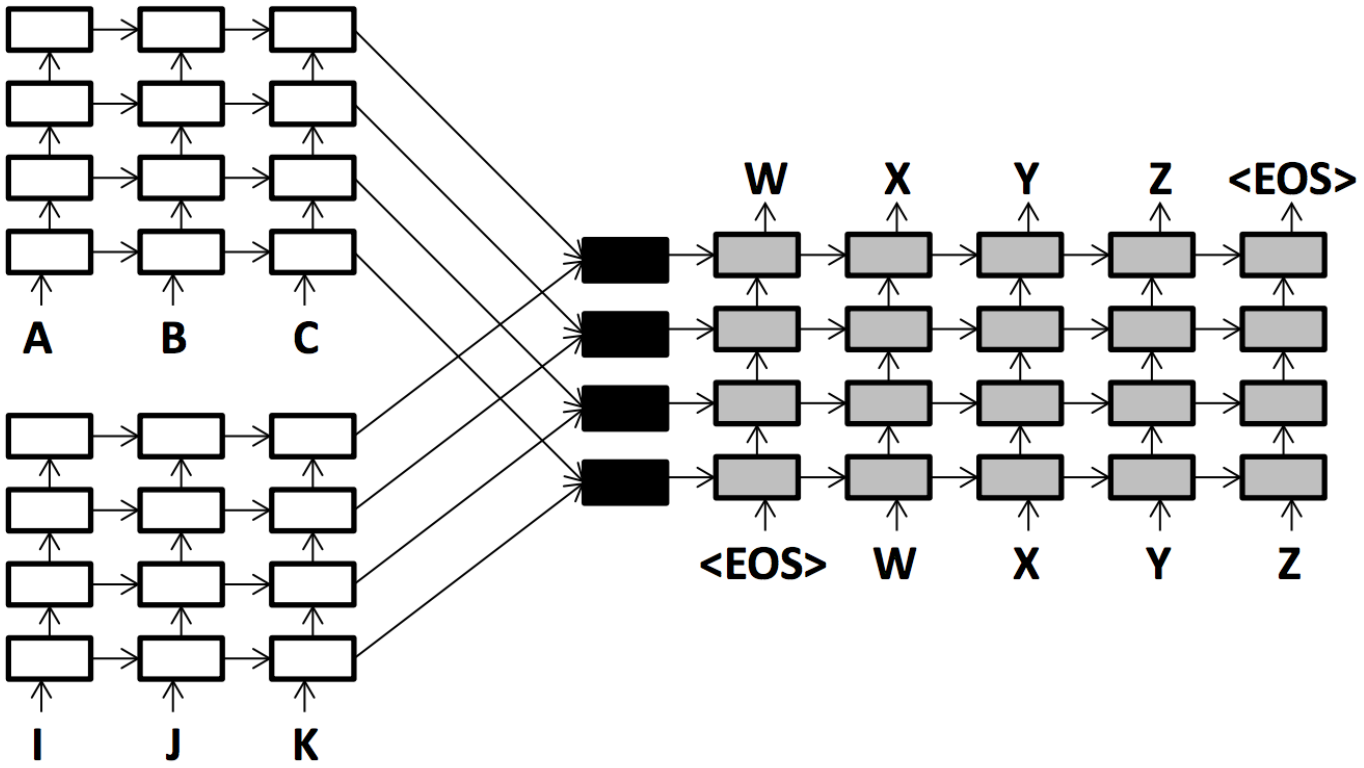


多种语言共享源语言编码器

(Dong et al., 2015)

进展6：多语言

- 多源神经机器翻译：多个输入，单个输出



基于多个编码器的注意力机制

(Zoph and Knight, 2016)

进展7

多模态翻译

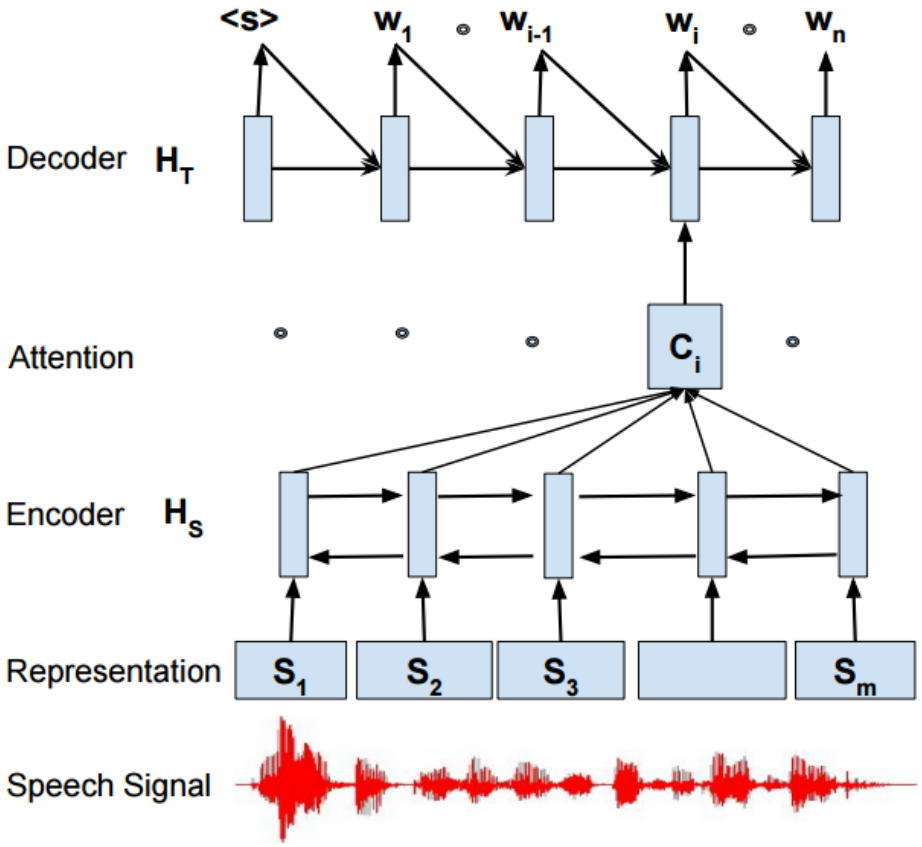
进展7：多模态

- 利用向量空间贯通文本、语音和图像

(Duong et al., 2016; Hitschler et al., 2016)

进展7：多模态

- 直接语音翻译：建立语音信号与文本的注意力机制



不经过语音识别，直接将源语言语音翻译成目标语言文本

(Duong et al., 2016)

进展7：多模态

- 基于多模态桥接的图像描述翻译

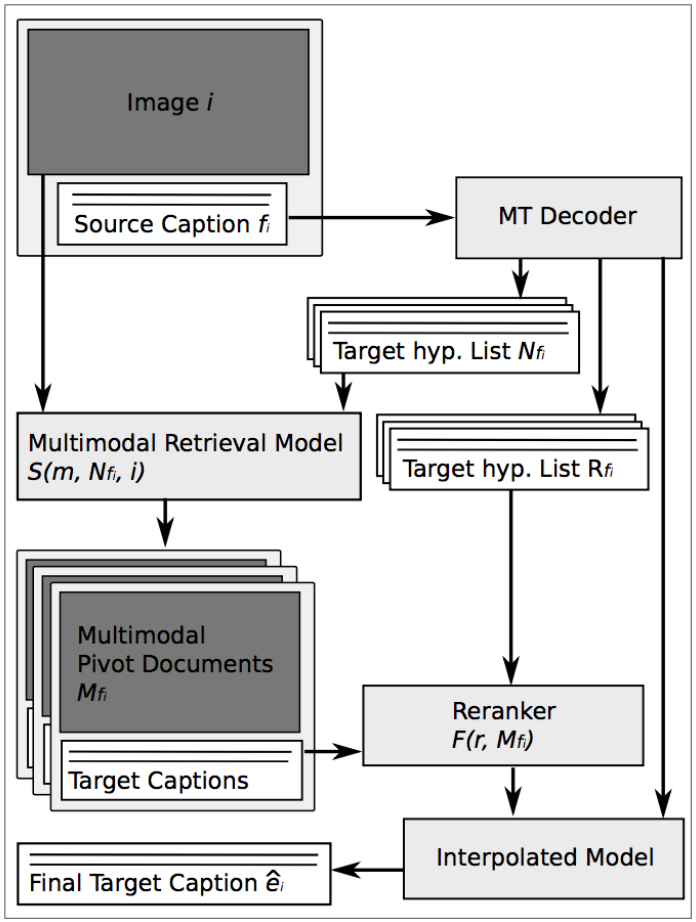
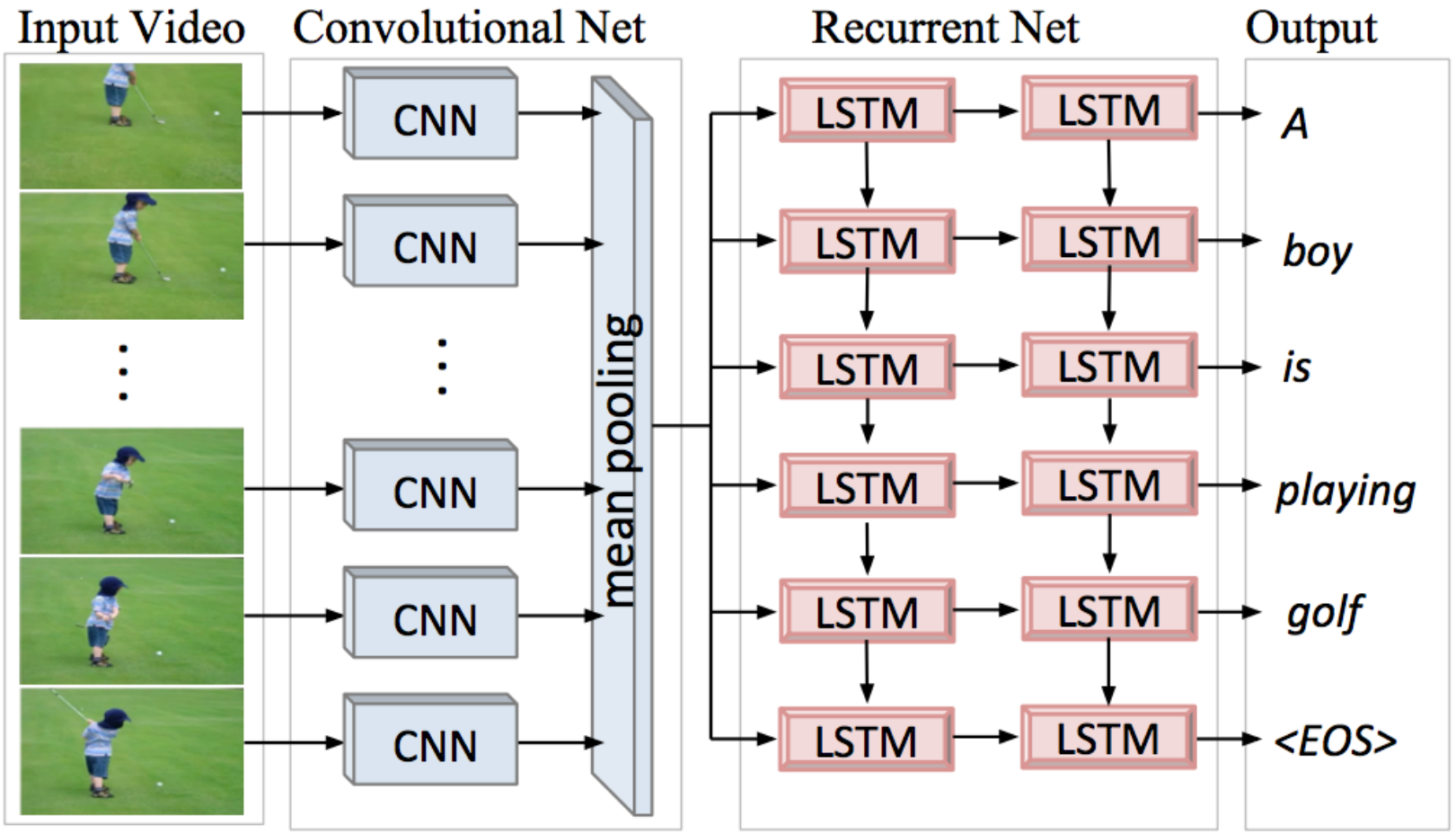


Image:	
Source:	Ein großer Stadtbus in dichtem Verkehr.
cdec:	a great city in heavy traffic .
TMR-TXT:	a great city in dense traffic .
TMR-CNN:	a large city bus in heavy traffic .
TMR-HCA:	a large city bus in heavy traffic .
Reference:	the large city bus is pulling into the traffic .

(Hitschler et al., 2016)

进展7：多模态

- 为视频生成描述文字

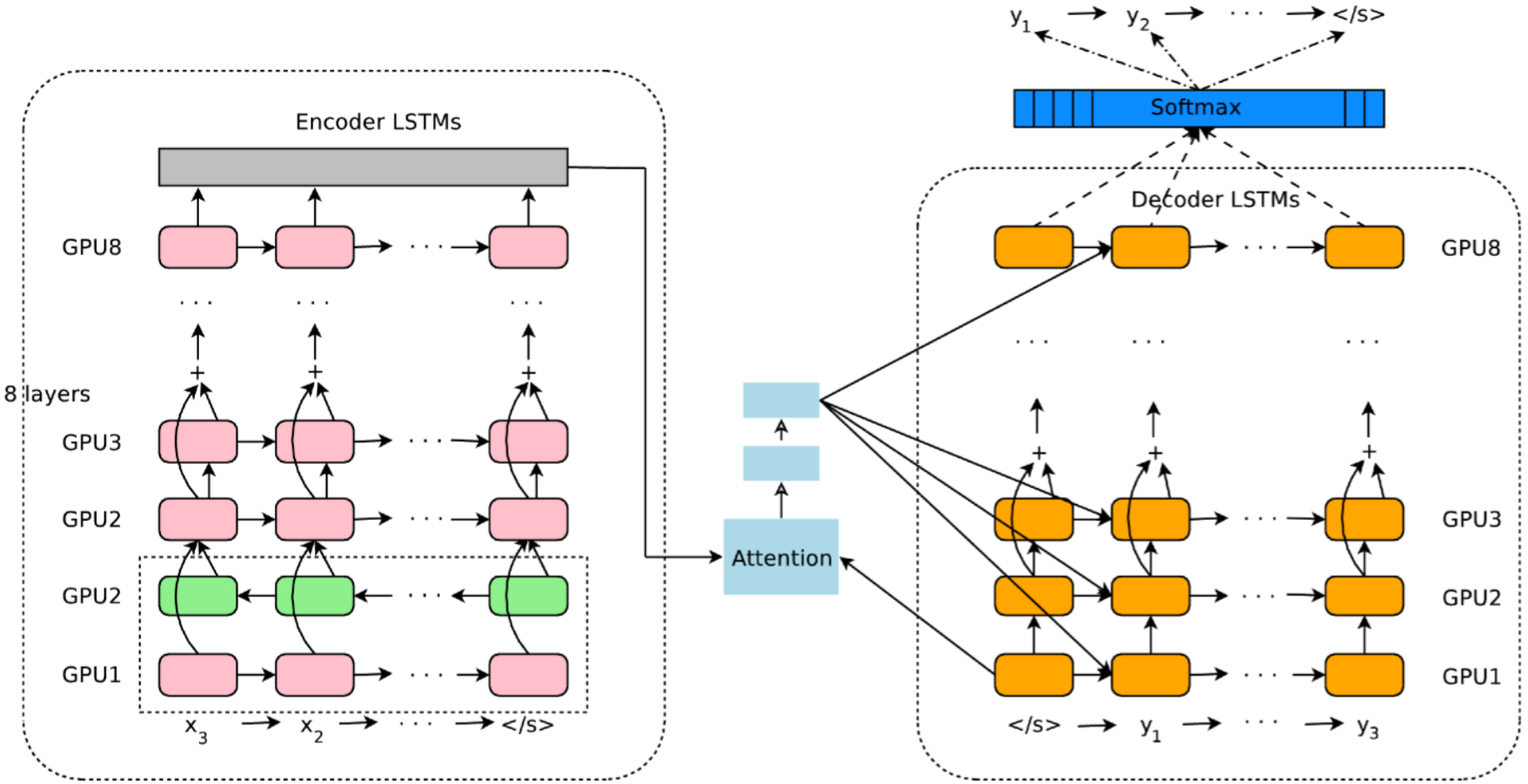


(Venugopalan et al., 2015)

神经机器翻译教程和开源工具

- 教程
  - Neural Machine Translation, ACL 2016 Tutorials
  - Introduction to NMT with GPUs, Kyunghyun Cho
- 开源工具
  - GroundHog：加拿大蒙特利尔大学
  - Blocks：加拿大蒙特利尔大学
  - TensorFlow：Google
  - EUREKA-MangoNMT：中国科学院自动化研究所

工业界NMT系统



Google Neural Machine Translation (GNMT)

工业界NMT系统

美国总统奥巴马昨天在白宫与来访的中国国家主席习近平就朝鲜半岛局势问题举行了一个小时的会谈。

U.S. President Barack Obama held talks with visiting Chinese President Xi Jinping at the White House yesterday for an hour on the situation on the Korean Peninsula.

Baidu 翻译

统计机器翻译 Vs 神经机器翻译

	统计机器翻译	神经机器翻译
表示	离散	连续
模型	线性	非线性
训练	MERT	MLE / MRT
可解释性	高	低
训练复杂度	低	高
处理全局调序	句法	门阀、注意力
内存需求	高	低

神经机器翻译面临的挑战

- 如何设计表达能力更强的模型？
- 如何提高语言学方面的可解释性？
- 如何降低训练复杂度？
- 如何与先验知识相结合？
- 如何实现多模态翻译？

总结

- 神经机器翻译：通过神经网络直接实现自然语言的相互映射。
- 神经机器翻译近年来取得迅速发展，有望取代统计机器翻译成为新的主流技术。
- 神经机器翻译在架构、可解释性、训练算法等方面仍面临挑战，需要进一步深入探索。

谢谢！

http://nlp.csai.tsinghua.edu.cn/~ly/