

第 15 章 机器翻译

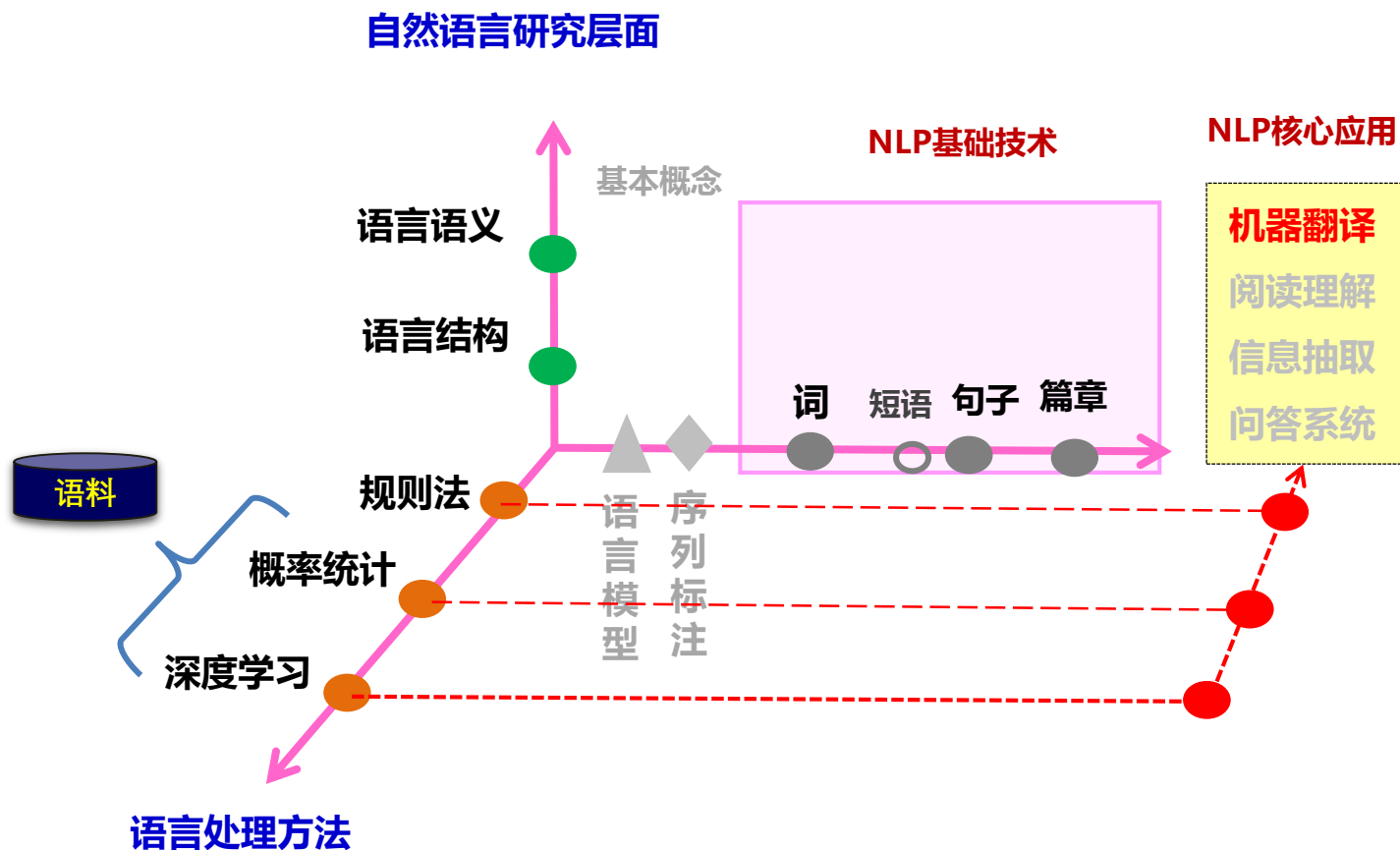
中科院信息工程研究所第二研究室

胡玥

huyue@iie.ac.cn

自然语言处理课程内容及安排

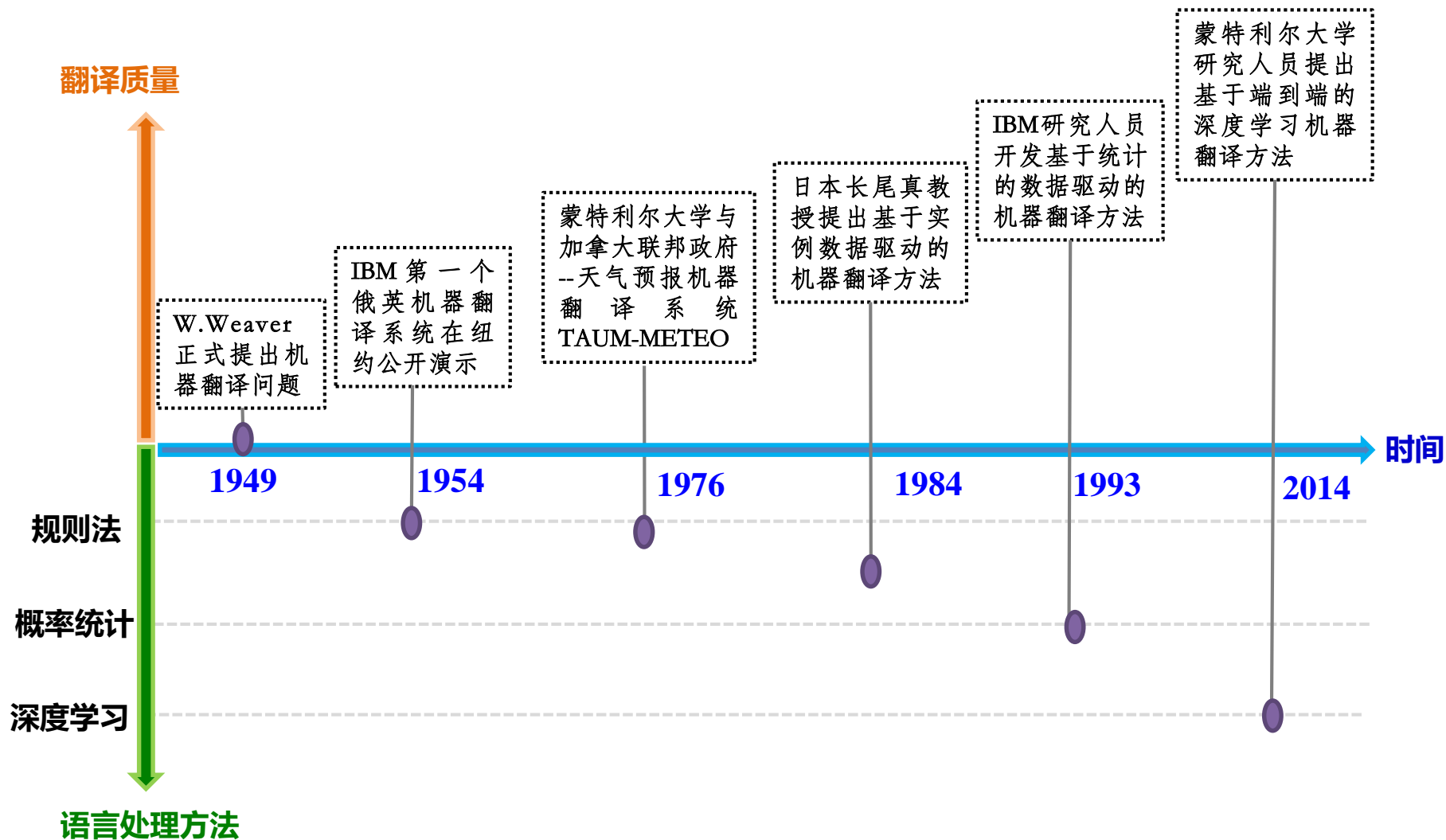
◇ 课程内容：



机器翻译 (machine translation, MT)：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)

机器翻译概述

机器翻译发展历史：



内 容 提 要

15.1 早期（1993前）机器翻译方法

15.2 统计机器翻译方法

15.3 神经网络机器翻译方法

15.1 早期机器翻译方法

早期机器翻译方法

- 基于规则机器翻译方法 (语言知识由语言专家人工编写)
- 基于实例的机器翻译方法 (语言知识来源语料库)
- 翻译记忆方法 (基于实例翻译法的一种特例)

15.1 早期机器翻译方法

基于规则机器翻译方法

规则机器翻译系统基本步骤：

- 需要有字典
- 词语切分
- 引入人名识别与翻译
- 引入词性标注
- 引入句法分析
- 引入翻译规则：插入译文词
- 引入翻译规则：调整语序
- 引入翻译规则：多层次调序与插入
- 引入目标语言属性

各步的所有规则均需要人工编写

15.1 早期机器翻译方法

基于规则的方法优点：

- 直观，能够直接表达语言学家的知识
- 规则的颗粒度具有很大的可伸缩性。大颗粒度的规则具有很强的概括能力，小颗粒度的规则具有精细的描述能力
- 便于处理复杂的结构和进行深层次的理解，如解决长距离依赖问题
- 大颗粒度的规则具有较强的系统适应性，不依赖于具体的训练语料

基于规则的方法缺点：

- 规则的覆盖性差，特别是细颗粒度的规则很难总结得比较全面
- 规则之间的冲突没有好的解决办法（翘翘板现象）
- 规则一般只局限于某一个具体的系统，规则库开发成本太高

15.1 早期机器翻译方法

基于实例的机器翻译方法

长尾真(Makoto Nagao)在1984年发表了《采用类比原则进行日-英机器翻译的一个框架》一文指出：

- 人类并不通过做深层的语言学分析来进行翻译，人类的翻译过程是：首先把输入的句子正确地分解为一些短语碎片，接着把这些短语碎片翻译成其它语言的短语碎片，最后再把这些短语碎片构成完整的句子，每个短语碎片的翻译是通过类比的原则来实现的。
- 因此，应该在计算机中存储一些实例，并建立由给定的句子找寻类似例句的机制，这是一种由实例引导推理的机器翻译方法，也就是基于实例的机器翻译。

实例翻译方法是基于语料库的机器翻译方法

15.1 早期机器翻译方法

在基于实例的机器翻译系统中，翻译知识以实例和义类词典的形式来表示，系统的主要是双语对照的翻译实例库，实例库主要有两个字段，一个字段保存源语言句子，另一个字段保存与之对应的译文，每输入一个源语言的句子时，系统把这个句子同实例库中的源语言句子字段进行比较，找出与这个句子最为相似的句子，并模拟与这个句子相对应的译文，最后输出译文。如果利用了较大的翻译实例库并进行精确的对比，有可能产生高质量译文，而且避免了基于规则的那些传统的机器翻译方法必须进行深层语言学分析的难点。在翻译策略上是很有吸引力的。

15.1 早期机器翻译方法

基于实例方法优点:

- 使用语料库作为翻译知识来源, 无需人工编写规则系统开发成本低, 速度快
- 从语料库中学习到的知识比较客观
- 从语料库中学习到的知识覆盖性比较好

基于实例方法缺点:

- 系统性能依赖于语料库
- 数据稀疏问题严重
- 语料库中不容易获得大颗粒度的高概括性知识

15.1 早期机器翻译方法

翻译记忆方法

**翻译记忆方法 (Translation Memory) 是基于实例方法的特例；
也可以把基于实例的方法理解为广义的翻译记忆方法；**

翻译记忆的基本思想：

- 把已经翻译过的句子保存起来
- 翻译新句子时，直接到语料库中去查找
 - 如果发现相同的句子，直接输出译文
 - 否则交给人去翻译，但可以提供相似的句子的参考译文

内 容 提 要

15.1 早期（1993前）机器翻译方法

15.2 统计机器翻译方法

15.2.1 统计翻译方法简介

15.2.2 翻译系统评价

15.3 神经网络机器翻译方法

15.2 统计机器翻译方法

统计机器翻译起源



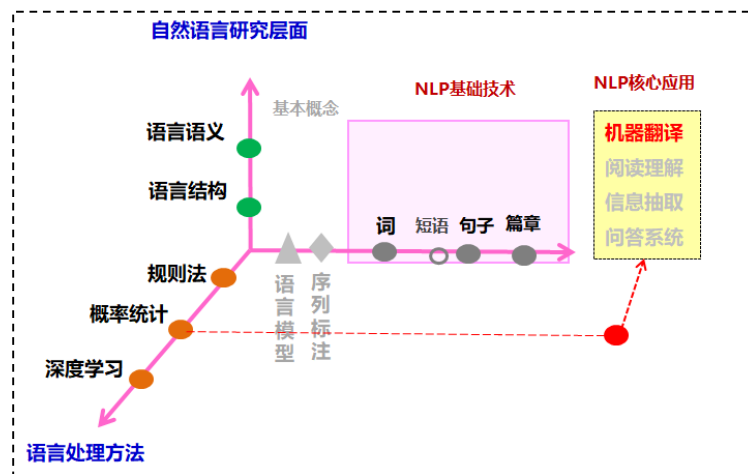
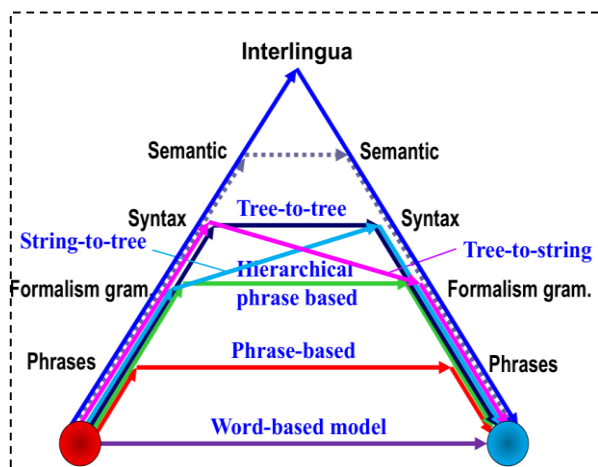
W. Weaver

1946年 美国洛克菲勒基金会 (Rockefeller Foundation) 副总裁 W. Weaver提出机器翻译的想法； 1949年，韦弗发表了一份以《翻译》为题的备忘录，提出：“当我阅读一篇用汉语写的文章的时候，我可以说，这篇文章实际上是用英语写的，只不过它是用另外一种奇怪的符号编了码而已，在阅读时，我是在进行解码”。韦弗的卓越思想成为了而后统计机器翻译 (Statistic Machine Translation, 简称SMT) 的理论基础。

1990 年IBM 的Peter F. Brown 等人在Computational Linguistics 上发表论文 “统计机器翻译方法”； 1993 年他们发表在该杂志发表论文 “统计机器翻译的数学：参数估计” 两篇文章奠定了统计机器翻译的理论基础。

15.2 统计机器翻译方法

统计机器翻译方法



1. 基于词的翻译方法

2. 基于短语的翻译方法

3. 基于句法的翻译方法

基于层次化短语方法（形式句法）

基于树的方法（语义句法）

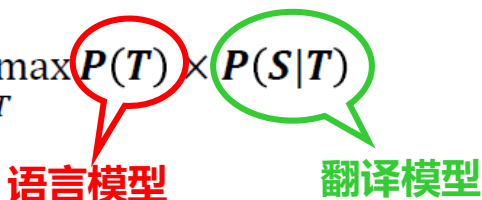
1. 基于词的翻译方法

问题： 源语言句子： $S=s_1^m=s_1s_2\cdots s_m$ **目标：** 求 $t^* = \operatorname{argmax}_t \Pr(t|s)$

目标语言句子： $T=t_1^l= t_1t_2\cdots t_l$

解决方法： 贝叶斯公式： $P(T|S) = \frac{P(T) \times P(S|T)}{P(S)}$

$$T' = \operatorname{argmax}_T P(T) \times P(S|T)$$



语言模型 翻译模型

统计翻译中的三个关键问题：

1. 估计语言模型概率 $p(T)$;
2. 估计翻译概率 $p(S|T)$;
3. 快速有效地搜索 T 使得 $p(T) \times p(S | T)$ 最大。

1. 基于词的翻译方法

1. 估计语言模型概率 $p(T)$

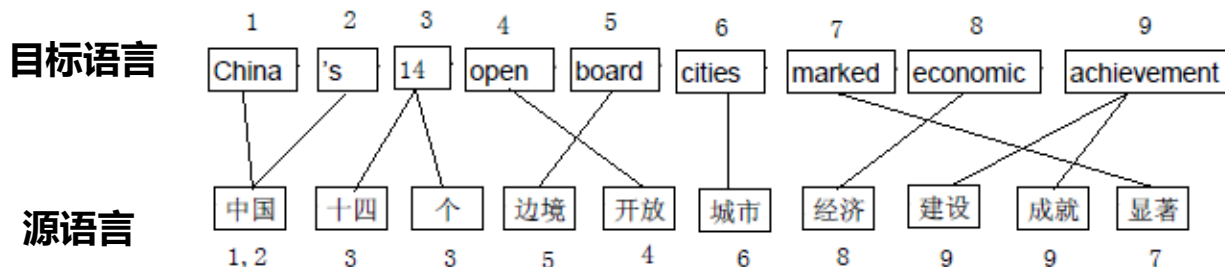
目标语言模型学习问题（详见“语言模型”）

2. 估计翻译概率 $p(S|T)$

$$T' = \operatorname{argmax}_T P(T) \times P(S|T)$$

翻译模型

翻译时词之间的对应情况



根据如何定义词对齐关系有：IBM 翻译模型 1-5

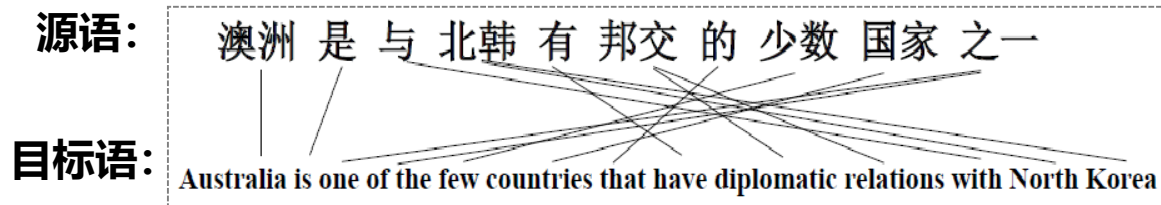
1. 基于词的翻译方法

3. 快速有效地搜索T 使得 $p(T) \times p(S | T)$ 最大

给定S, 求T, 使得 $P(T) * P(S|T)$ 最大

经典的算法: • 贪婪算法, • 堆栈搜索, •

基于词翻译例:



最早模型, 性能已经被其他方法所超越, 但其中建立在各种词对齐思想上的词语对齐工具例如(GIZA++), 是统计机器翻译方法的基础工作。

1. 基于词的翻译方法

基于词的翻译模型存在的问题

基于词的翻译模型只刻画了词到词的翻译概率，词翻译的时候没有考虑上下文，在词语调序方面能力很差。

- 难以刻画一些固定搭配、习惯用法的翻译；
- 很难处理词义消歧问题
- 很难处理一对多、多对一和多对多的翻译问题

很多研究者想到了在短语层面进行建模，以改进局部词语调序的效果。最成功的工作是 Och、Zens、Koehn 等人的工作。

[Koehn, 2003] 提出基于短语的对数线性模型翻译模型

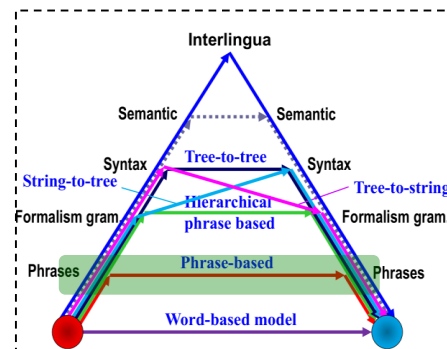
2. 基于短语的翻译方法

2. 基于短语的翻译方法

代表模型：

基于短语的对数线性翻译模型[Koehn, 2003]

要素：{ 翻译单位： 短语
翻译模型： 对数线性模型



基本思想：

以短语为基本翻译单元 把训练语料库中所有对齐的短语及其翻译概率 存储起来，
作为一部带概率的短语词典，翻译的时候将输入的句子与短语词典进行匹配，
选择最好的短语划分，按短语进行翻译，然后将得到的短语译文重新排序，得
到最优的译文

2. 基于短语的翻译方法

如： 对齐的短语：

(澳洲 是, Australia is)	概率
(与 北韩, with North Korea)	概率
(有 邦交, have the diplomatic relations)	概率
(的 少数 国家 之一, one of the few countries that)	概率

短语：

指一个连续的词串(n -gram), 不一定是语言学中定义的短语(phrase)

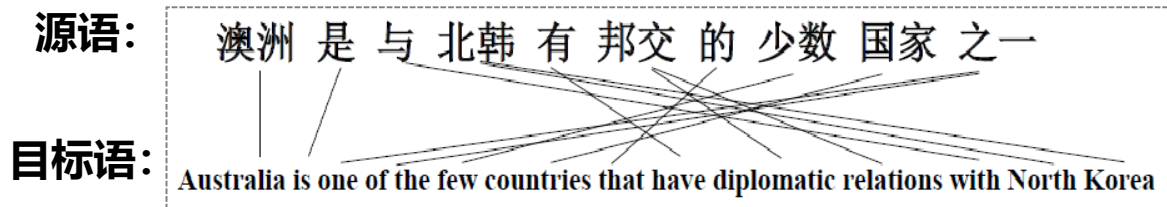
如： 我想预订一个单人间。

I would like to reserve a single room.

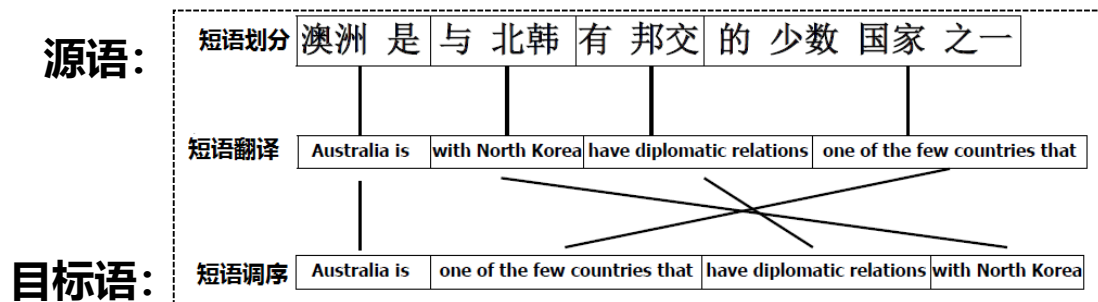
2. 基于短语的翻译方法

基于短语翻译优势：

基于词翻译：

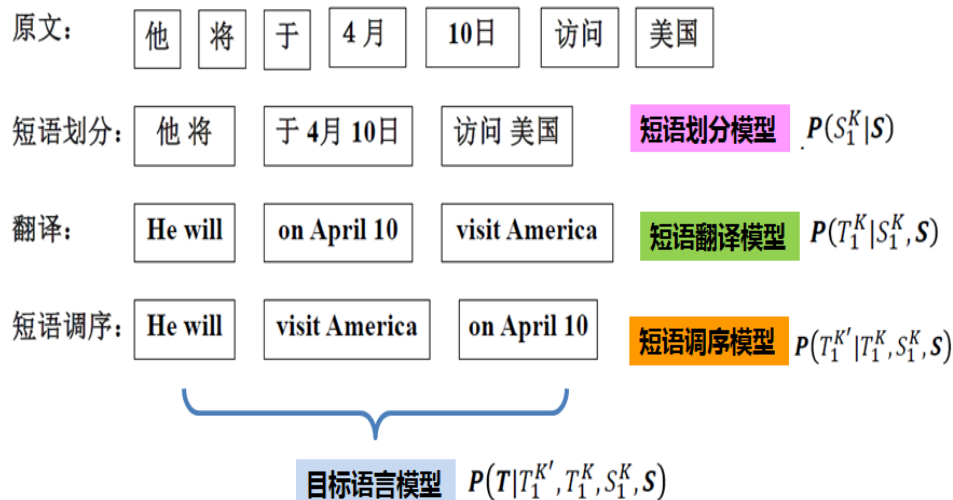


基于短语翻译：

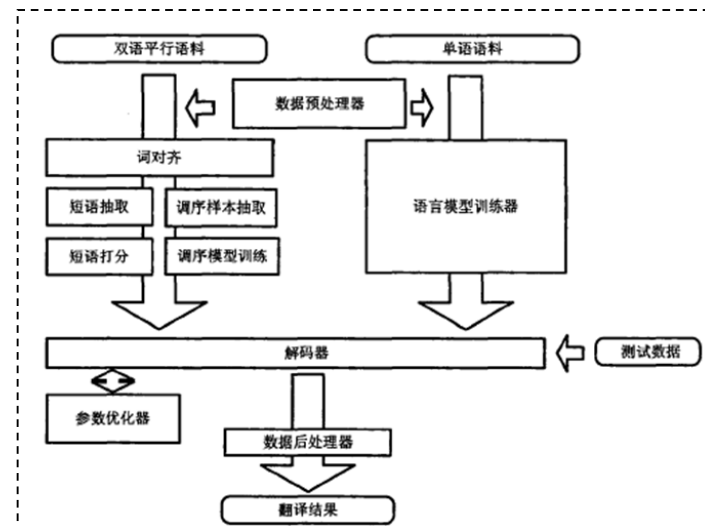


2. 基于短语的翻译方法

基于短语的统计翻译过程



基于短语的对数线性模型翻译系统



线性对数模型

$$T' = \operatorname{argmax}_T P(T|S) = \operatorname{argmax}_T \frac{\exp\{\sum_1^M \lambda_m h_m(T, S)\}}{\sum_{T^*} \exp\{\sum_1^M \lambda_m h_m(T^*, S)\}}$$

$$= \operatorname{argmax}_T \left\{ \sum_1^M \lambda_m h_m(T, S) \right\}$$

3. 基于句法的翻译方法

3. 基于句法的翻译方法

- ◆ 基于层次化短语方法（形式句法） 括号转录语法
- ◆ 基于树的方法（语义句法） 同步树替换文法（STSG）

特点： 用句法分析方法进行机器翻译，需要双语**同步上下文无关文法**

同步上下文无关文法 (SCFG)

SCFG是CFG的扩展，是上下文无关文法针对两个输出符号串的泛化。CFG (Σ, N, P, S) 包括终结符集合 Σ ，非终结符集合 N ，和产生式集合 $\{P \rightarrow \{N^* \times N^*\}\}$ ，而在**同步上下文无关文法**中，文法指定每个产生式包含两个输出。这些产生式通过共标的非终结符在两个输出字符串间建立联系。

如

$NP \rightarrow DT_1 NPB_2 / DT_1 NPB_2$

$NPB \rightarrow NPB_1 AJ_2 / AJ_2 NPB_1$

$NPB \rightarrow JJ_1 NN_2 / JJ_1 NN_2$

$DT \rightarrow \text{the} / \epsilon$

$AJ \rightarrow \text{strong} / \text{呼啸}$

$JJ \rightarrow \text{north} / \text{北}$

$NN \rightarrow \text{wind} / \text{风}$

同步上下文无关文法生成同构的源语和目标语一对树，树上对应的非终结符对齐。其中一个树可以通过旋转非终结符节点转换为另一个树。

同步上下文无关文法

$NP \rightarrow DT_1NPB_2 / \underline{DT_1NPB_2}$

$NPB \rightarrow NPB_1AJ_2 / \underline{AJ_2NPB_1}$

$NPB \rightarrow \underline{JJ_1NN_2} / \underline{JJ_1NN_2}$

$DT \rightarrow the / \epsilon$

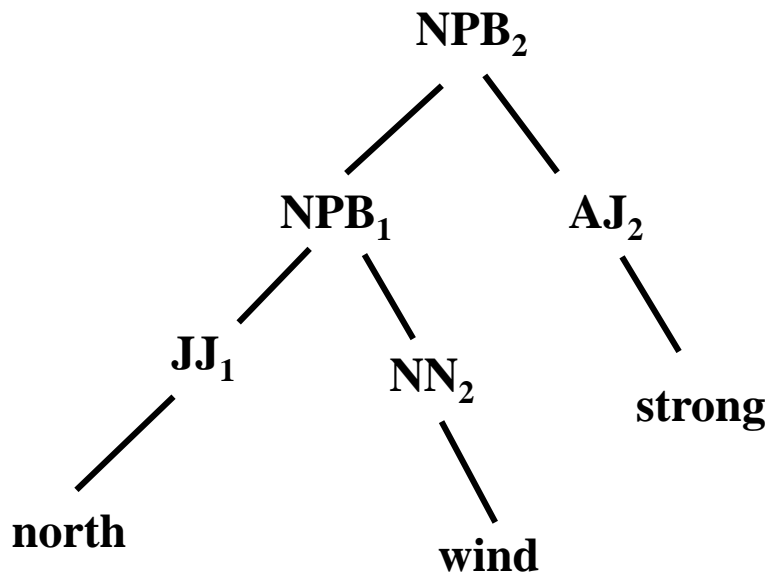
$AJ \rightarrow \underline{strong} / \underline{呼啸}$

$JJ \rightarrow \underline{north} / \underline{北}$

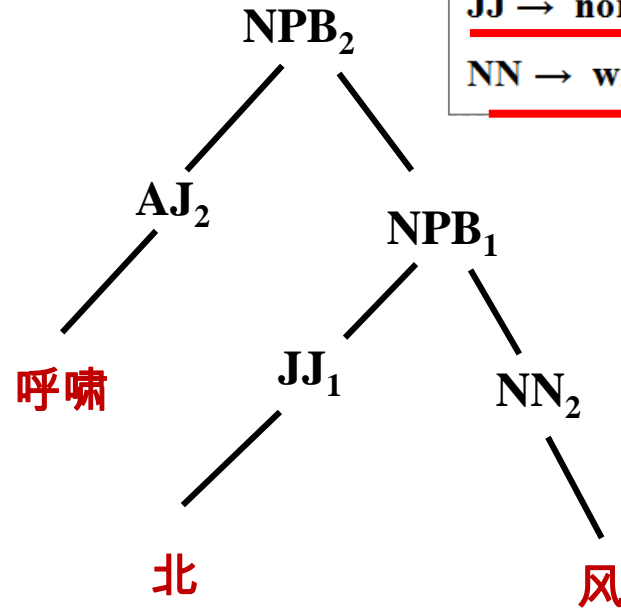
$NN \rightarrow \underline{wind} / \underline{风}$

如： 源语言： north wind strong

用同步上下文无关文法对源语言进行分析



源语言： north wind strong



目标语： 呼啸北风

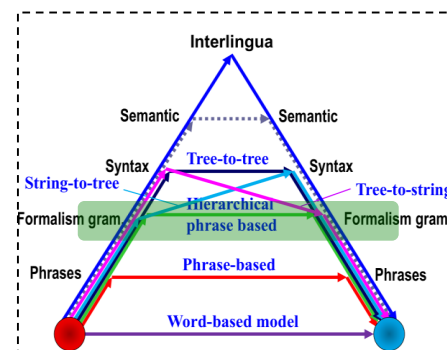
3. 基于句法的翻译方法

◆ 基于层次化短语方法：

代表模型：

基于层次短语的翻译模型[David Chiang,2005]

要素：{ 翻译单位： 层级短语模板
翻译模型： 句法分析方法进行机器翻译



核心是引入了**嵌套层次短语**的思想，并采用**括号转录语法**（同步上下文无关语法的特例）作为形式化方法，在完成源语言句法分析的同时，目标语言就生成了，因此可利用各种成熟的句法分析算法进行机器翻译，而无需另外设计专门的翻译算法。

优点：易于实现，解码过程复杂度相对较低，效果比传统短语模型有很大提高

3. 基于句法的翻译方法

层次短语文法规则采用同步文法（括号转录语法），所有句法结构规则（短语模板）
不使用任何语言学知识直接从平行语料库中自动学习得到

语法规则：

- (1) $X \rightarrow \langle \text{与 } X_1 \text{ 有 } X_2, \text{ have } X_2 \text{ with } X_1 \rangle$
- (2) $X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_2 \text{ that } X_1 \rangle$
- (3) $X \rightarrow \langle X_1 \text{ 之一, one of } X_1 \rangle$
- (4) $X \rightarrow \langle \text{澳洲, Australia} \rangle$
- (5) $X \rightarrow \langle \text{邦交, diplomatic relations} \rangle$
- (6) $X \rightarrow \langle \text{少数国家, few countries} \rangle$
- (7) $X \rightarrow \langle \text{北韩, North Korea} \rangle$
- (8) $(S \rightarrow S_1 X_2, S_1 X_2)$
- (9) $(S \rightarrow X_i, X_i)$

翻译过程：在对源语端进行形式文法分析时可以自动生成译文。

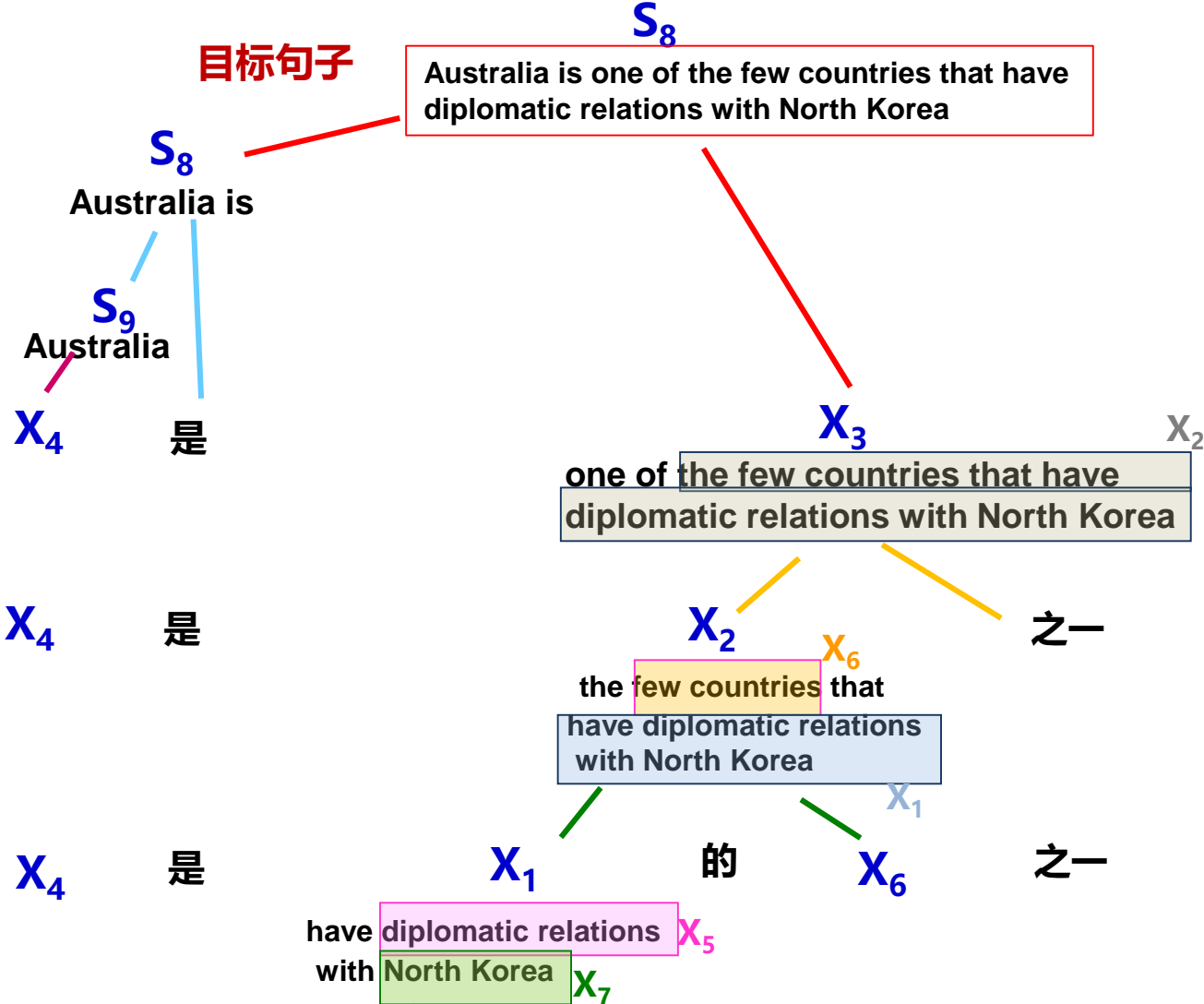
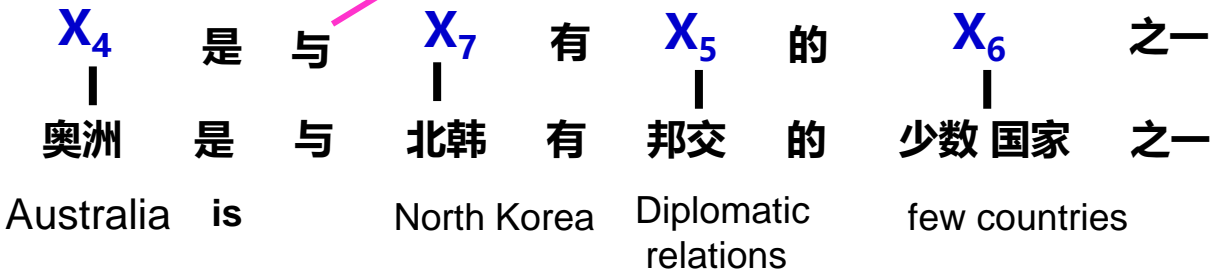
如，源语：澳洲 是 与北韩 有邦交 的少数国家 之一

目标语：Australia is one of the few countries that have diplomatic relations with North Korea

如，翻译过程

- 语法规则：
- (1) $X \rightarrow \langle \text{与 } X_1 \text{ 有 } X_2, \text{ have } X_2 \text{ with } X_1 \rangle$
 - (2) $X \rightarrow \langle X_1 \text{ 的 } X_2, \text{ the } X_2 \text{ that } X_1 \rangle$
 - (3) $X \rightarrow \langle X_1 \text{ 之一, one of } X_1 \rangle$
 - (4) $X \rightarrow \langle \text{澳洲, Australia} \rangle$
 - (5) $X \rightarrow \langle \text{邦交, diplomatic relations} \rangle$
 - (6) $X \rightarrow \langle \text{少数国家, few countries} \rangle$
 - (7) $X \rightarrow \langle \text{北韩, North Korea} \rangle$
 - (8) $(S \rightarrow S_1 X_2, S_1 X_2)$
 - (9) $(S \rightarrow X_1, X_1)$

源语句子



3. 基于句法的翻译方法

◆ 基于树的翻译方法：

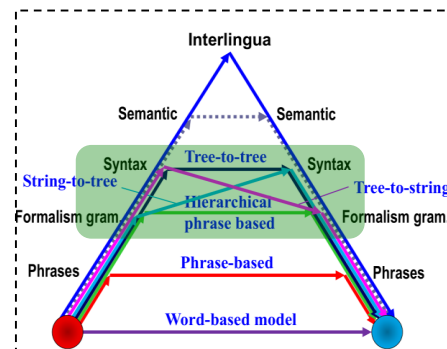
特点：

树翻译模型属于语言学上基于句法的模型

使用语言学知识

语言知识规则通常从句法树库训练得到

翻译模型：句法分析方法进行机器翻译



树模型分为：

- **树到树模型**：在源语言端和目标语言端都使用语言知识
- **树到串模型**：只在源语言端使用语言知识
- **串到树模型**：只在目标语言端使用语言知识

3. 基于句法的翻译方法

- 树到树的代表模型：

Zhang et al.(2007, 2008) 提出了树到树的翻译模型

- 特点：**
- 句法分析：将源语言句子分析为一棵句法结构树（短语结构树）
 - 树到树的转换：递归地将源语言句子的句法结构树转换为目标语言句子的句法结构树，拼接叶结点得到译文。

- 树到串的代表模型：

Yang Liu (ACL2006) 提出了树到串的翻译模型

- 特点：**
- 在源语言端进行句法分析
 - 在目标语言端不进行句法分析
 - 从源语言端句法分析和词语对齐的语料库中抽取翻译规则
 - 递归地将源语言句子的句法结构树转换为目标语言句子
(树到串转换)

3. 基于句法的翻译方法

- 串到树的代表模型:

Galley et al.(2004, 2006) , 提出了串到树的翻译模型

- 特点:**
- 在源语言端进行不句法分析
 - 在目标语言端进行句法分析
 - 从目标语言端句法分析和词语对齐的语料库中抽取翻译规则并构造翻译模型
 - 利用串到树转换规则, 将源语言句子分析为一棵目标语言句法结构树, 拼接叶结点得到译文

内 容 提 要

15.1 早期（1993前）机器翻译方法

15.2 统计机器翻译方法

15.2.1 统计翻译方法简介

15.2.2 翻译系统评价

15.3 神经网络机器翻译方法

15.2.2 翻译系统评价

常用的评测指标

主观评测:

(1) 流畅度 (2) 充分性 (3) 语义保持性

自动评测:

由评测系统依据一定的数学模型对译文句子自动计算得分。

常用的自动打分方法有:

- BLEU 评价方法
- NIST评测方法
- mWER 方法
- GTM方法
- METEOR 评测

15.2.2 翻译系统评价

BLEU 评价方法 [Papineni, 2002] - BiLingual Evaluation Understudy, IBM

基本思想：

将机器翻译产生的候选译文与人翻译的多个参考译文相比较，越接近，候选译文的正确率越高。

实现方法：

统计同时出现在系统译文和参考译文中的 n 元词的个数，最后把匹配到的 n 元词的数目除以系统译文的单词数目，得到评测结果。

例如：

系统译文：	I am a teacher .	1-gram : 分数 = $3/4$
参考译文1：	I am a student.	2-gram : 分数 = $2/3$
参考译文2：	I am a worker.	

15.2.2 翻译系统评价

但会出现如下问题:

系统译文: the the the the the the the.

参考译文1: The cat is on the mat.

参考译文2: There is a cat on the mat.

按照上述计算方法, 如果 n 取1 的话, 打分 = $7/7$

但显然这种翻译结果几乎没有任何意义。

修正方法:

Max_Ref_Count: 每个单词在所有参考译文中出现次数的最大值

Count: 该单词在系统译文中出现的总次数

计数 $\text{Count clip} = \min(\text{Count}, \text{Max_Ref_Count})$

15.2.2 翻译系统评价

Total_Count clip=所有单词的Count clip 值累加起来，最后，用 Total_Count clip 除以**系统译文**中全部单词的个数。

$$p_n = \frac{\sum_{C \in \{Candidate\}} \sum_{n\text{-gram} \in C} Count_{clip}(n\text{-gram})}{\sum_{C' \in \{Candidate\}} \sum_{n\text{-gram}' \in C'} Count(n\text{-gram}')}$$

如：系统译文中的单词the 在参考译文1 中出现的次数最多，Max_Ref_Count = 2，而 the 在系统译文中出现的次数为7，即 Count = 7，因此，Count clip = min(7, 2) = 2。候选译文中全部单词的个数等于7，因此，该例中修正后的一元语法精确度为2/7。

15.2.2 翻译系统评价

在修正的 n 元语法精度计算中，随着 n 值的增大精度值几乎成指数级下降，因此，BLEU 方法中采用了修正的 n 元语法精度的对数加权平均值，相当于对修正的精度值进行几何平均， n 值最大为4。另外，考虑到句子的长度对上述BLEU 评分也有一定的影响，例如，如果一个机器翻译系统只翻译最可靠的词汇，译文句子就可能比较短，按上述方法计算出的精度值就会较高。因此，需要进一步考虑候选译文的句子长度对计算评分的影响。

15.2.2 翻译系统评价

BLEU 值定义:

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

最大语法的阶数，实际取4

长度过短句子的惩罚因子

$w_n = 1/N$

出现在答案译文中的n元词语接续组占候选译文中n元词语接续组总数的比例。

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

c为候选译文中单词的个数，r为答案译文中与c最接近的译文单词个数。

BLEU 分值范围: 0 ~ 1, 分值越高表示译文质量越好, 分值越小, 译文质量越差。

内 容 提 要

15.1 早期（1993前）机器翻译方法

15.2 统计机器翻译方法

15.3 神经网络机器翻译方法

15.3 神经网络机器翻译方法

神经网络机器翻译方法

特点：端到端的神经网络机器翻译

- 有监督翻译模型 
 1. 基本RNN架构翻译模型
 2. Attention+ RNN架构翻译模型
 3. Transformer 翻译模型
- 无监督翻译模型

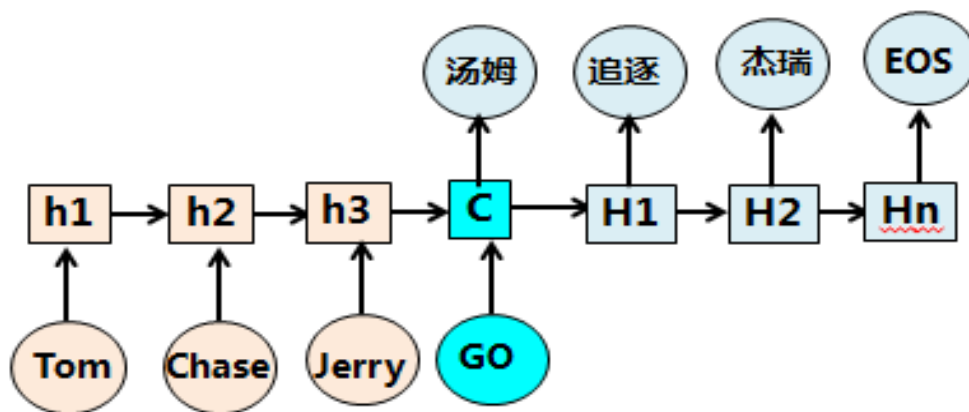
1. 基本RNN架构翻译模型

1. 基本RNN架构翻译模型

基本架构： Encoder-Decoder RNN框架

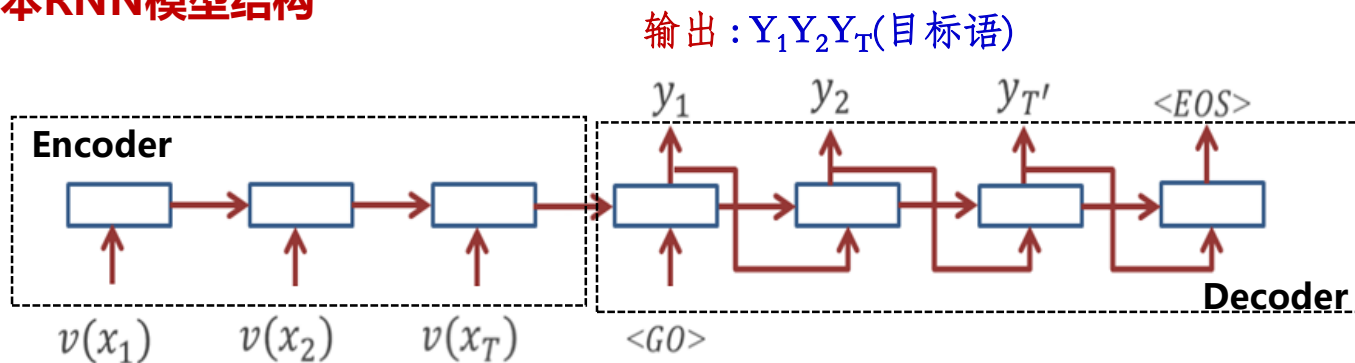
源语： Tom chase Jerry

目标语： 汤姆追逐杰瑞

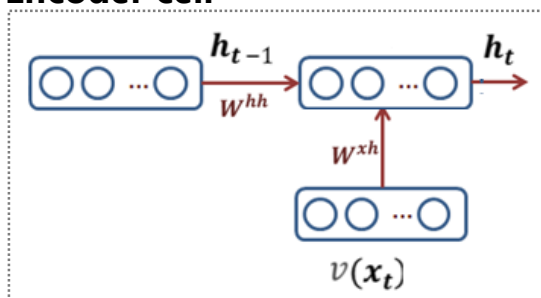


1. 基本RNN架构翻译模型

■ 基本RNN模型结构



Encoder cell



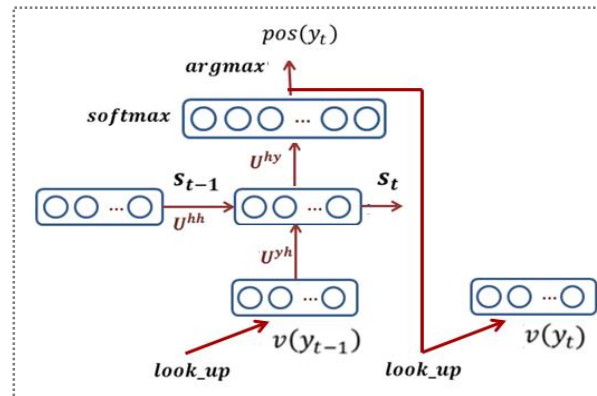
参数: W^{xh} , W^{hh}

$$h_t = \text{sigm}(v(x_t)W^{xh} + h_{t-1}W^{hh})$$

$$h_0 = 0; \quad C = h_T$$

其中, $V(w_i)$ 表示 w_i 的词向量

Decoder cell



参数: U^{yh} , U^{hh} , U^{hy}

$$S_t = \text{sigm}(v(y_{t-1})U^{yh} + S_{t-1}U^{hh})$$

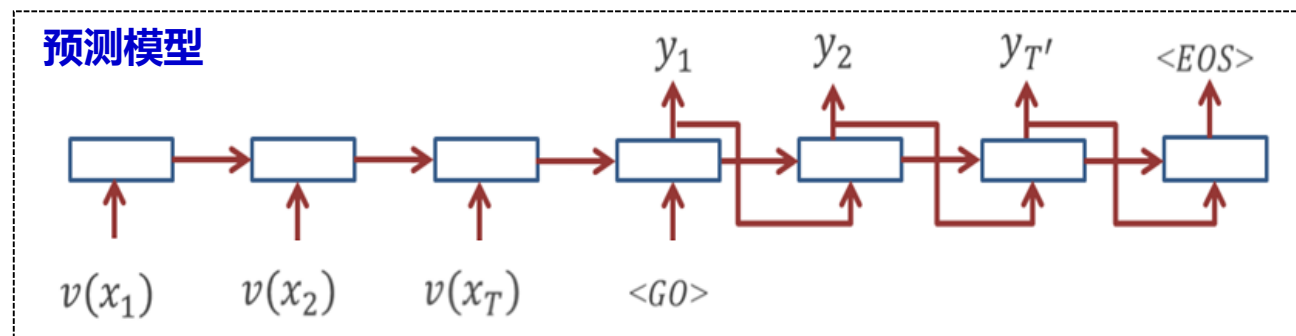
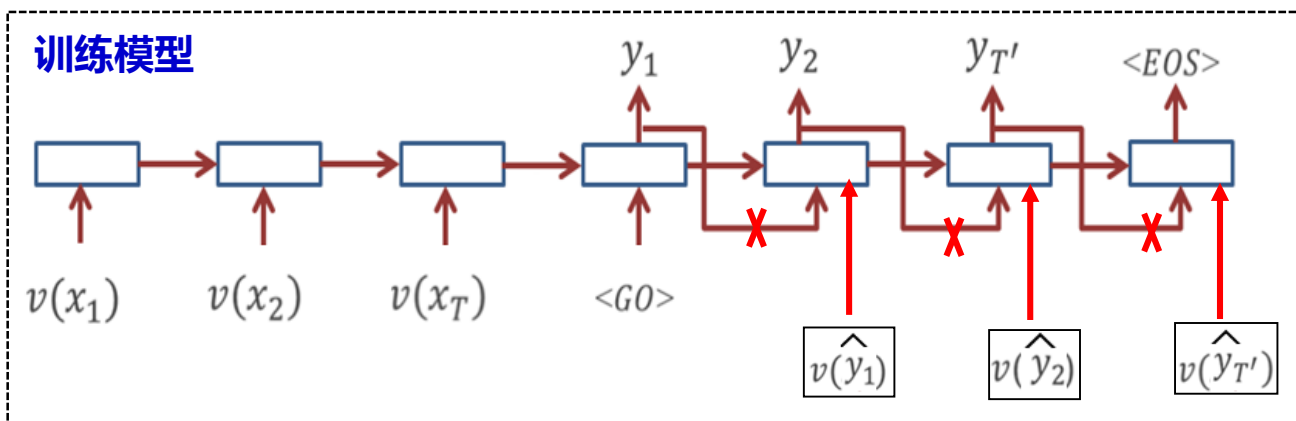
$$\text{Pos}(y_t) = \text{arg}(\text{softmax}(S_t U^{hy}))$$

$$S_0 = C; \quad \text{Pos}(y_0) = \text{Pos}(<\text{Go}>)$$

1. 基本RNN架构翻译模型

■ 基本RNN模型学习

模型训练过程 实例 (x^i, \hat{y}^i) 源语: $x_1 x_2 x_T$ 目标语: $\hat{y}_1 \hat{y}_2 \hat{y}_{T'}$



1. 基本RNN架构翻译模型

■ 基本RNN模型学习

- 损失函数: $\theta = [W^{xh}, W^{hh}, U^{yh}, U^{hh}, U^{hy}]$

交叉熵损失 $J(\theta; x, y) = -\frac{1}{T'} \sum_{t=1}^{T'} \sum_{i=1}^{tar_vocab} \hat{y}_t^{(i)} \log(pred_t^{(i)})$

其中: $Pred_t = g(y_{t-1}, S_{t-1}, U^{yh}, U^{hh}, U^{hy}, C)$

$$C = f(x_1, x_2, \dots, x_T; W^{xh}, W^{hh})$$

- 参数学习: 梯度下降, BPTT

$$\theta := \theta - \alpha [\nabla_{\theta} J((\theta; x, y))]$$

学得参数: $\theta = [W^{xh}, W^{hh}, U^{yh}, U^{hh}, U^{hy}]$

1. 基本RNN架构翻译模型

翻译模型开发:

1. 准备语料

① 训练词向量语料:

用于源语词向量训练的源语语料

```
1 在国内，车臣非法武装的分裂活动威胁到地区安全、联邦主权和领土完整。
2 这些都有力地证明了中国人民实现祖国统一的愿望是无法阻挡的历史潮流。
3 然而，他们却颇有大将风度，化险为夷，转危为安，直到发射成功。
4 这一系列政策的颁布实施，有力地调动了该院科研人员的创新热情。
5 新华社耶路撒冷2月11日电 新闻分析：以色列空袭黎巴嫩不利于中东和平进程 新华社
6 根据历史唯物主义的基本观点，人民，只有人民，才是创造世界历史的动力。
7 “更寄希望于台湾人民”不仅是一个想法，而且有具体的内容。
8 朱邦造指出，中国一贯主张世界事务的磋商、协调和决策，应当具有广泛代表性。
9 在讨论全球性问题时，也有必要多听取其他国家，特别是发展中国家的意见。
```

需要用于目标语词向量训练的目标语语料

```
1 internally , the separatist activities of the illegal chechen militia threaten regional s
2 all this eloquently proves that the desire of the chinese people to realize the reunifica
3 however , they have the posture of generals , and they can disguise their feelings and pu
4 the publication and implementation of these numerous policies has thus powerfully mobiliz
5 eyc news analysis by xna reporter : " israel 's air raids over lebanon are unfavorable to
6 according to the basic viewpoint of historical materialism , the people , and only the pe
7 " all the more pinning hopes on the taiwan people " is not only an idea ; it also has a s
8 zhu bangzao pointed out : china has always advocated that consultation , coordination , a
9 when discussing global issues , it is necessary to listen more to opinions of other count
```

1. 基本RNN架构翻译模型

② 句对齐双语平行语料（用于翻译模型训练）：

目前,粮食出现阶段性过剩,恰好可以以粮食换森林、换草地,再造西部秀美山川。工程的立项、设计都要充分论证,反复比较,使工程经得起时间检验。建设现代
论的焦点。与此同时,发展中国家代表对电子时代贫富差距的拉大表现出极大的担忧。南非总统曼德拉说,因特网带来的信息革命使得发展中国家与发达国家之间
6.2%。对国外评价系统的过分强调,既不符合客观实际,也不利于中国科技期刊的发展。在干预过程中,它甚至为美国创造出一种新的战争形态:“不接触战”。
中小学校必须坚决取消招生考试,公布学生名次,错误的做法。一个更重要的方面,是充分体现了我们对于台湾人民意愿、作用的重视与尊重。结果两市十四日平均升
有家较少一些网站还缺乏安全意识。目前我国的重要机构对网络的依赖性越来越强,如让黑客袭击,后果将非常严重。另一方面,要切实加强对进出口产品
将参加梅西奇总统的就职典礼。据悉,在上述背景下,才有昨日传出长和泰及盈动洽购香港电讯的消息。当这个训练法刚刚提出时,曾遭到专家们的强烈反对。
在的四百五十亿美元增加到二〇〇五年的一千亿美元。通知要求,各级人事部门要充分发挥所雇人才市场的作用,为毕业生就业提供优质服务。去年的九届人大一次
行员个个过得硬!全体飞行员掌握了灵活多变的战术和克敌制胜的规律,收到了很好的效果。联合国贸发会议第十届大会开幕前夕,笔者读到一篇报道,说一
中国的鸡肉产量只占肉类总产量的12%左右,而世界鸡肉占肉类总产量的23%。可奶作为补剂却渐渐为许多中老年人所接受,也构成稳定消费市场的一部分。
反映多年来成本的实际变化趋势。比较一下粮食的含税成本与平均出售价格,涨幅的百分比,后者高出大约1?5个百分点。在计划经济体制下,1965—1978年的13
感动,主动表示彻底脱离“法轮功”,回到党和人民的正确立场上来。罗干在全国加强基层政法队伍建设“电视电话会”上强调,加强基层政法队伍建设,维护社会
露端倪。此项政策使全国8400万人受益,居民收入每年可望增加1000亿元以上。有的代表评价,这是“没有重复建设”。在加强
the present food surplus can specifically serve the purpose of helping western china restore its woodlands, grasslands, and the beauty of its landscapes .the establishment and design of pr
ated here in the late years of song dynasty, which was 700 years ago .acting russian prime minister putin specifically pointed out at the recent russian federation security council meeting
angya reiterated that the chinese government will " resolutely defend china 's sovereignty, territorial integrity, national dignity, and national security . "in this year 's new year fire
the gap between the developing and the developed countries .the hard-working, brave, clever, and wise chinese people can surely solve their own affairs and will finally realize the mother
umph, i can finally talk about it now .pointing at wang yongzhi, qian xuesen said to the chief designer : " this young man 's opinion is correct, and do as he says ! "it turned out that
ectations . "china 's science and technology publishers have borne the dual pressures of a surplus number of science and technology magazines and an excessive emphasis on foreign appraisal s
ing to the relevant wto provisions, developing countries will in some respects enjoy differentials and favored treatment that are different from the developed countries .the ecologically fr
ao 's return to the motherland, the relationship between the msar and the region of taiwan has become a special component part of the cross-strait relationship .last year the port handled l
ies with and give support to education undertakings . "starting immediately after this spring term begins, all middle schools and primary schools must resolutely abolish the practice of publ
i to a six-year imprisonment on the charge of engaging in illegal business and confiscated his property of 10,000 yuan .china also has developed " 53h2-type " guided missile escort ships , w

需要对语料中目标语句加入指示开始和结束的符号 <GO>和<EOS>

如: *encoder* = 在国内,车臣非法武装的分裂活动威胁到地区安全、联邦主权和领土完整。

decoder = <GO> internally, the separatist activities of the ... integrity of the Russian federation. <EOS>

1. 基本RNN架构翻译模型

2. 分别训练源语和目标语 词向量

- 源语言词向量:

embedding_lookup

```
0.050833 0.128765 -0.146094 0.069694 0.007024 -0.096424 -0.157390 0.011784 -0.038783
0.143288 -0.204101 0.099999 0.069656 0.109964 0.084649 0.235120 0.051784 -0.111841 0.
-0.105407 0.024066 0.140674 -0.024447 -0.014990 0.122730 -0.167001 -0.084349 0.073150
0.019436 -0.190295 0.022214 0.053573 -0.099395 0.040390 0.050647 -0.160216 0.186505 0.
0.055955 0.143302 -0.076757 -0.083509 -0.134587 -0.182467 -0.167716 0.012705 0.120715
0.056205 -0.255456 0.058806 0.005022 -0.052727 -0.029080 0.201284 0.035767 -0.149297
0.032334 -0.135239 0.022567 0.021690 -0.043513 -0.117100 0.022756 -0.109634 0.006124
0.064289 0.174939 -0.199188 0.067879 -0.037430 -0.106827 -0.060625 0.000091 0.068930
0.001241 -0.196275 -0.000043 -0.246263 -0.157610 -0.005805 -0.313385 -0.104575 0.0932
```

- 目标语言词向量:

embedding_lookup

```
0.120087 -0.135869 -0.123477 -0.042004 0.044839 -0.081288 0.088841 0.128275 -0.011523
0.079033 -0.028754 0.158046 -0.040011 0.025214 0.215386 -0.059622 0.099403 0.159544 0.
0.011927 -0.201049 0.117777 -0.008809 0.101911 -0.160056 -0.291013 -0.106556 0.087172
0.058882 -0.130262 -0.109831 0.002600 0.110658 -0.031069 0.111842 -0.007345 0.090138
0.160734 0.053305 -0.012226 0.041163 0.110672 -0.019010 -0.006176 -0.153083 -0.011480
0.089265 -0.094931 -0.375949 0.099409 0.036721 -0.102113 0.007733 0.127761 -0.095580
0.062840 -0.188951 0.169604 0.109607 -0.225956 -0.091960 0.122545 0.051216 0.231229 0.
-0.049814 0.033591 0.074439 0.183768 0.226920 0.015473 0.144618 0.140309 -0.107489 -0.
0.092753 0.011977 0.120027 0.180937 0.276690 0.204090 -0.038654 0.049030 -0.035187 0.
-0.074427 -0.063666 -0.008831 0.140220 -0.135664 0.121244 -0.011726 -0.267773 -0.0529
```


1. 基本RNN架构翻译模型

3. 训练模型参数

交叉熵损失
$$J(\theta; x, y) = -\frac{1}{T'} \sum_{t=1}^{T'} \sum_{i=1}^{tar \ vocab} \hat{y}_t^{(i)} \log(pred_t^{(i)})$$

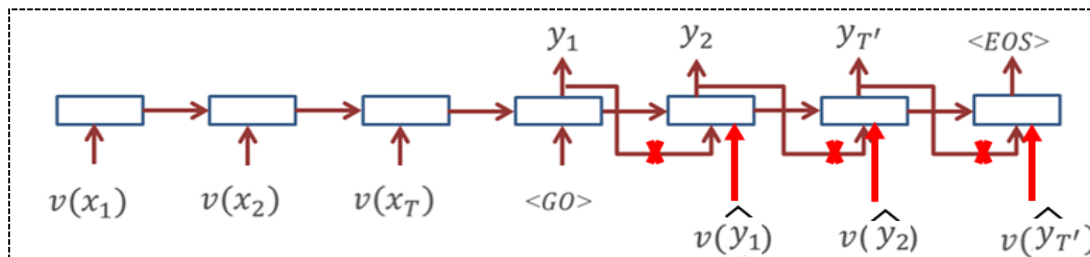
其中: $Pred_t = g(y_{t-1}, S_{t-1}, U^{yh}, U^{hh}, U^{hy}, C)$

$C = f(x_1, x_2, \dots, x_T; W^{xh}, W^{hh})$

- 参数学习: 梯度下降, BPTT 训练实例: 双语平行句对

学得参数: $\theta = [W^{xh}, W^{hh}, U^{yh}, U^{hh}, U^{hy}]$

训练模型

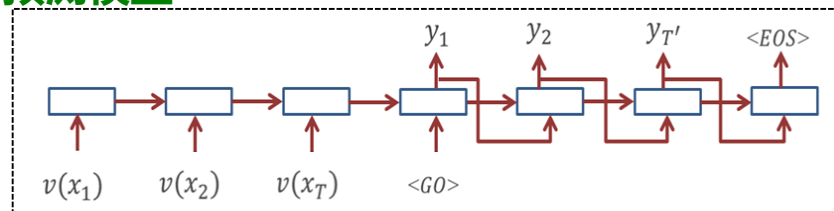


1. 基本RNN架构翻译模型

实验结果：

```
>>>香港 必须 看到 全 国 、 看到 世界 ， 把 自己 的 发展 与 国家 的 发展 结合  
起来 。  
>>>_GO hong kong must be able to see the whole country and the world and inte  
grate its own development with china 's national development . _PAD  
  
>>>hong kong must be taken to launch the process country and the united 's re  
alizing its development development . relations 's sovereignty defense .  
  
>>>我们 积极 开展 国际 交往 ， 但 一贯 坚持 原则 ， 坚决 捍卫 国家 的 主权 和  
民族 尊严 。  
>>>_GO he said : while actively unfolding international exchanges , we always  
uphold the principles and resolutely defend our national sovereignty and nat  
ional dignity . _PAD  
  
>>>he said : while actively unfolding international exchanges , we must uphol  
d the principles and resolutely defend our national sovereignty and national  
dignity .  
  
>>>即使 在 科学 技术 快速 发展 的 现代 社会 ， 人 仍然 是 最终 起 决定 作用 的  
因素 。  
>>>_GO even in modern society in which science and technology develop rapidly  
, people still consist of a decisive factor in the final analysis . _PAD  
  
>>>even in modern society in this science and technology . and in and in cons  
ist of a decisive factor in mistakes final analysis .
```

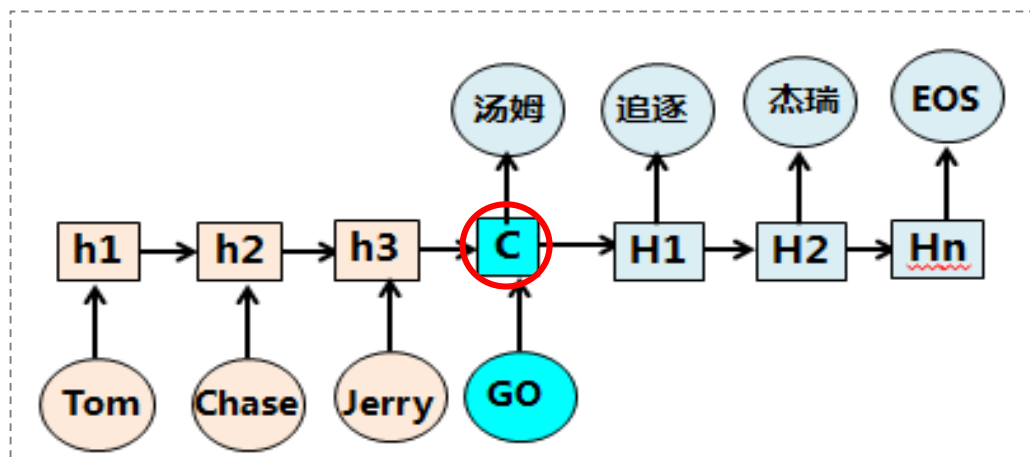
预测模型



1. 基本RNN架构翻译模型

基本RNN架构翻译模型存在问题

Encoder-Decoder RNN



问题: 对不同的输出 Y_i 中间语义表示C 相同

$$X = \langle x_1, x_2 \dots x_m \rangle$$

$$Y = \langle y_1, y_2 \dots y_n \rangle$$

$$C = \mathcal{F}(x_1, x_2 \dots x_m)$$

$$y_1 = f(C)$$

$$y_2 = f(C, y_1)$$

$$y_3 = f(C, y_1, y_2)$$

$$y_i = g(C, y_1, y_2 \dots y_{i-1})$$

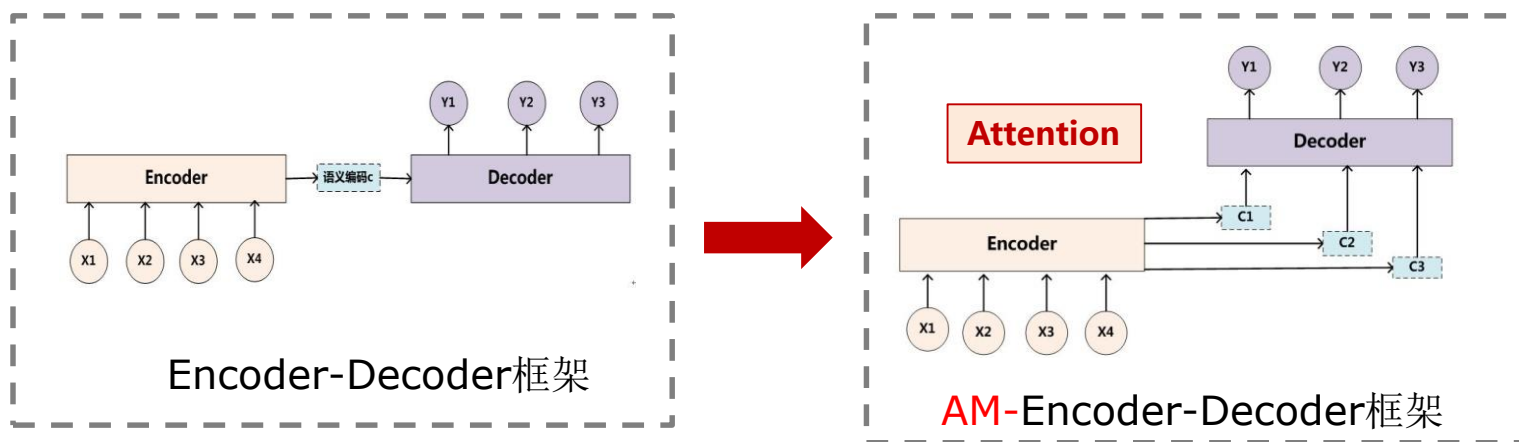
实际应该：在翻译“杰瑞”的时候，体现出英文单词对于翻译当前中文单词不同的影响程度，比如 (Tom,0.3) (Chase,0.2) (Jerry,0.5)

加入 Attention 模型

1. 基本RNN架构翻译模型

Attention 模型

一般用在Encoder-Decoder框架中的 Encoder 端和 Decoder 端之间关系计算中

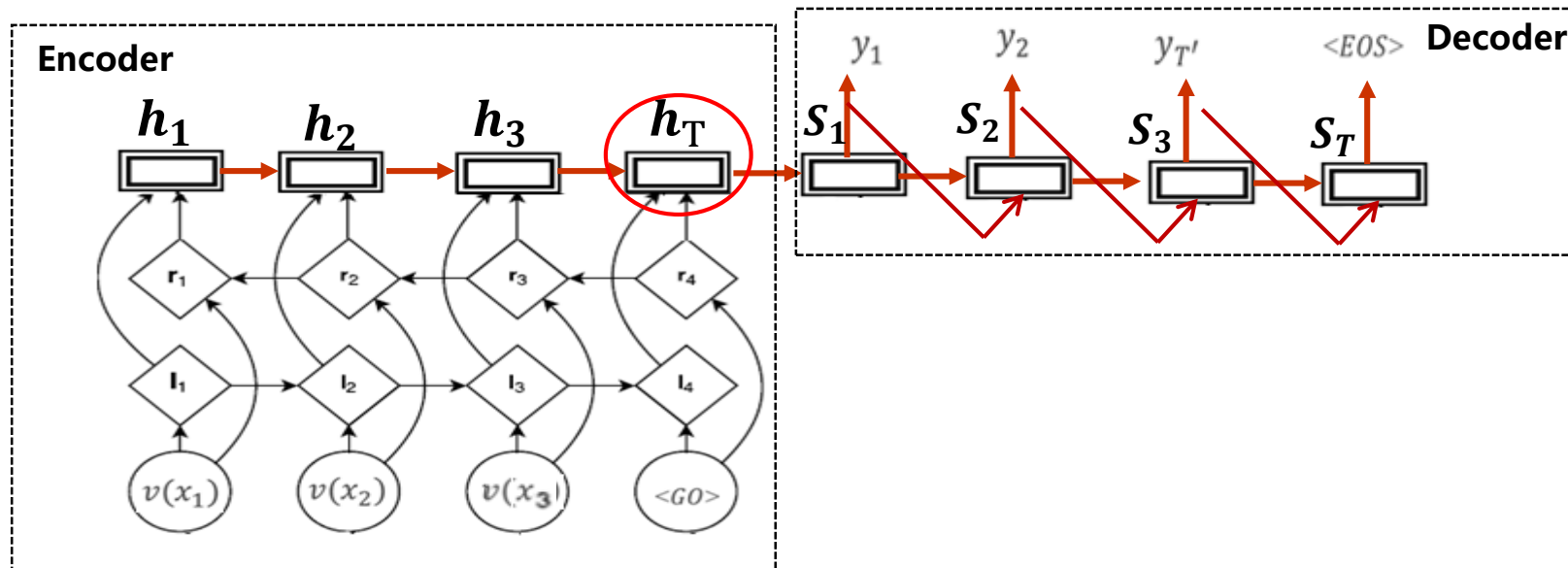


作用：让任务处理系统更专注于找到输入数据中显著的与当前输出相关的有用信息，从而提高输出的质量。

优势：不需要监督信号，可推理多种不同模态数据之间的难以解释、隐蔽性强、复杂映射关系，对于先验认知少的问题，极为有效。

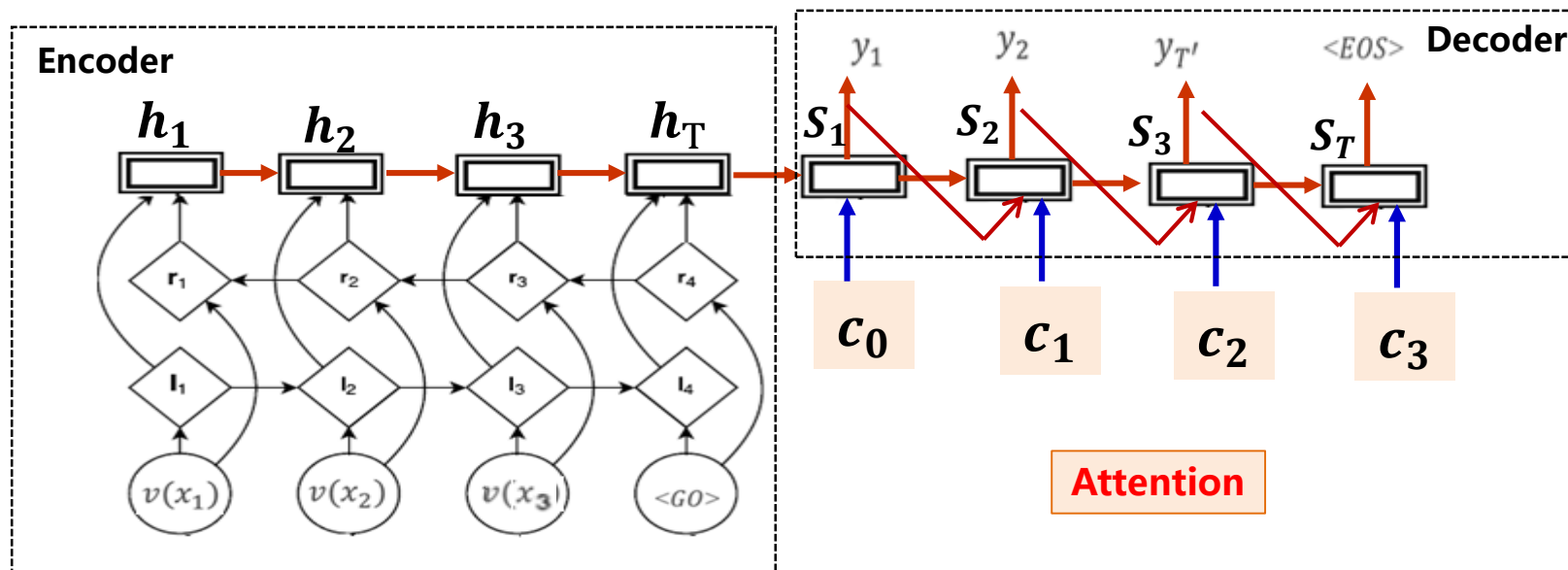
2. Attention+ RNN架构翻译模型

2. Attention+ RNN架构翻译模型



2. Attention+ RNN架构翻译模型

2. Attention+ RNN架构翻译模型



输入: $x_1 x_2 x_T$ (源语)

输出: $p(y_i | y_1, \dots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$

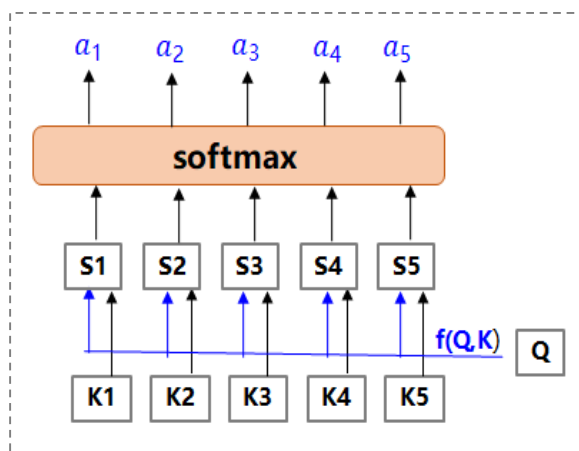
$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

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

$$e_{ij} = \underbrace{a(s_{i-1}, h_j)}_{f(Q, K)}$$

2. Attention+ RNN架构翻译模型

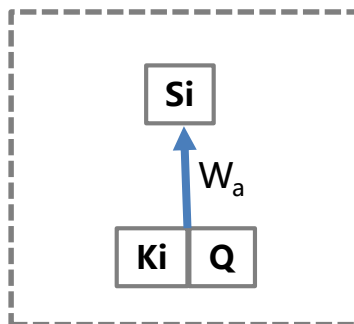
回顾注意力打分函数 $f(Q,K)$



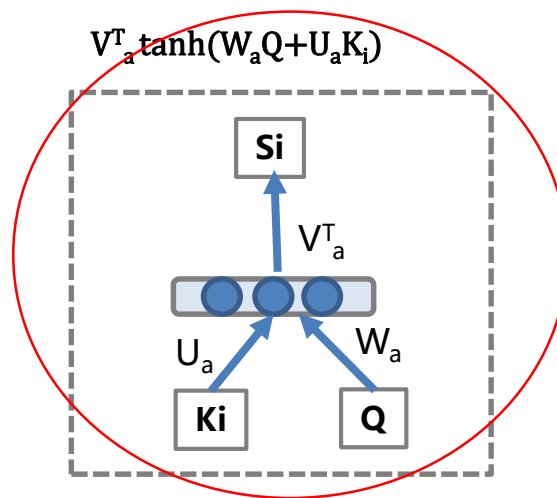
$$f(Q,K) = \begin{cases} Q^T K_i & \text{dot} \\ Q^T W_a K_i & \text{general} \\ W_a [Q, K_i] & \text{concat} \\ V_a^T \tanh(W_a Q + U_a K_i) & \text{perceptron} \end{cases}$$

乘法模型 (Multiplicative Model)
加法模型 (Additive Model)

$W_a [Q, K_i]$

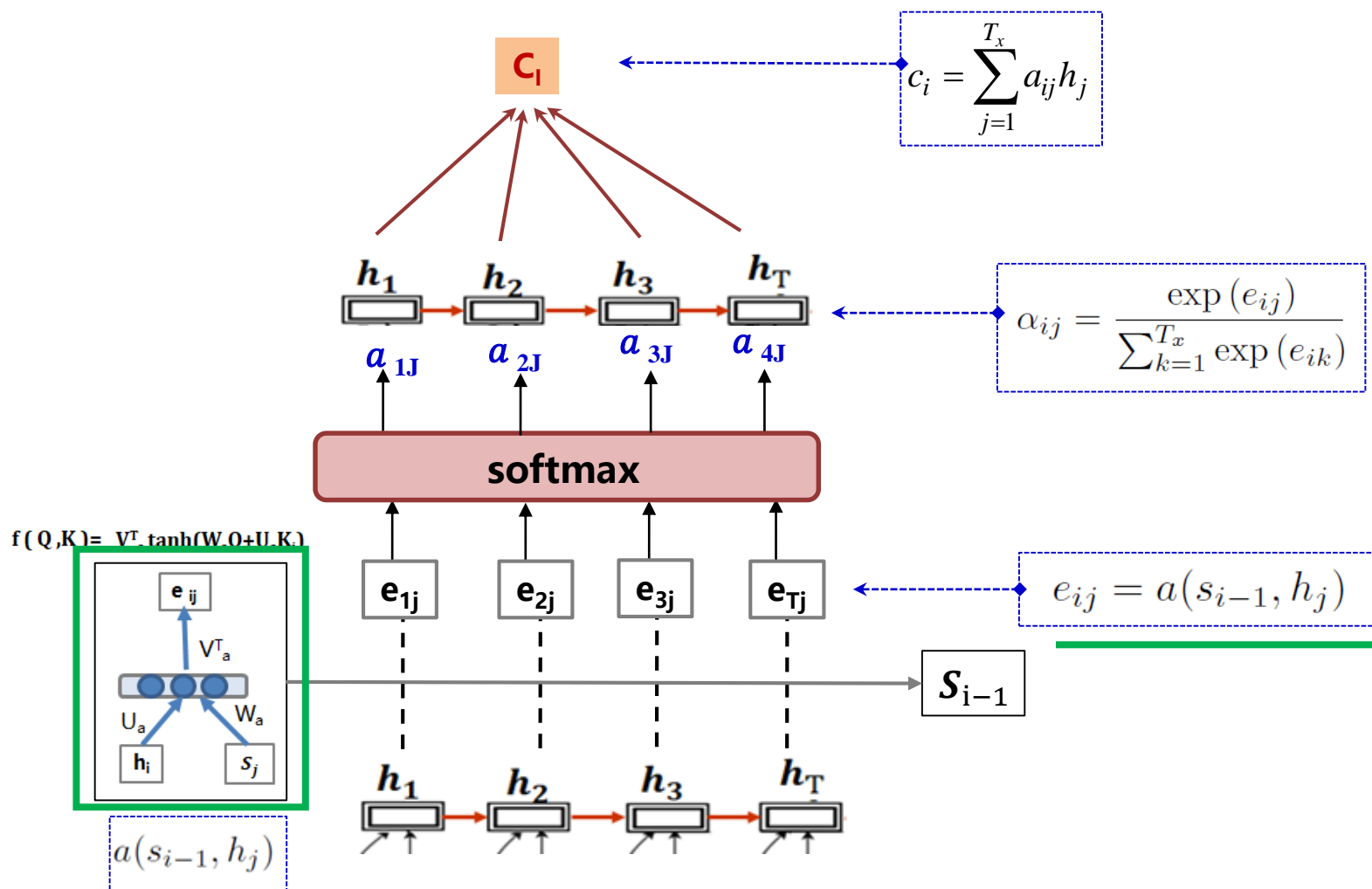


$V_a^T \tanh(W_a Q + U_a K_i)$



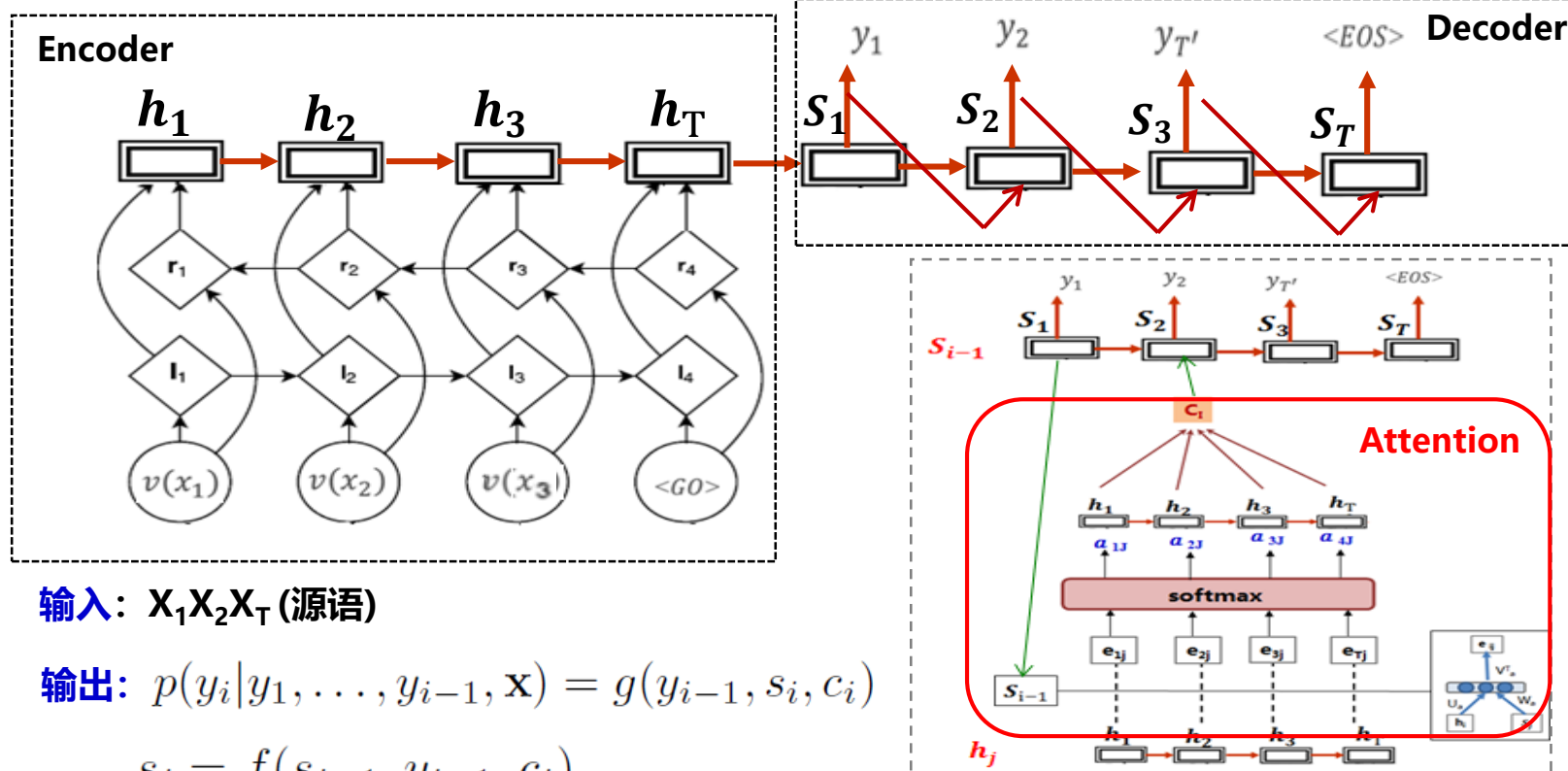
2. Attention+ RNN架构翻译模型

计算 C_i : $e_{ij} = v_a^T \tanh(S_{i-1}W_a + h_jU_a)$ 其中, j 表示Decoder 端节点序号, i 表示Encoder 端节点序号



2. Attention+ RNN架构翻译模型

2. Attention+ RNN架构翻译模型



输入: $x_1 x_2 x_T$ (源语)

输出: $p(y_i | y_1, \dots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, s_i, c_i)$

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

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

$$e_{ij} = v_a^T \tanh(s_{i-1} W_a + h_j U_a)$$

■ Attention+ RNN模型学习

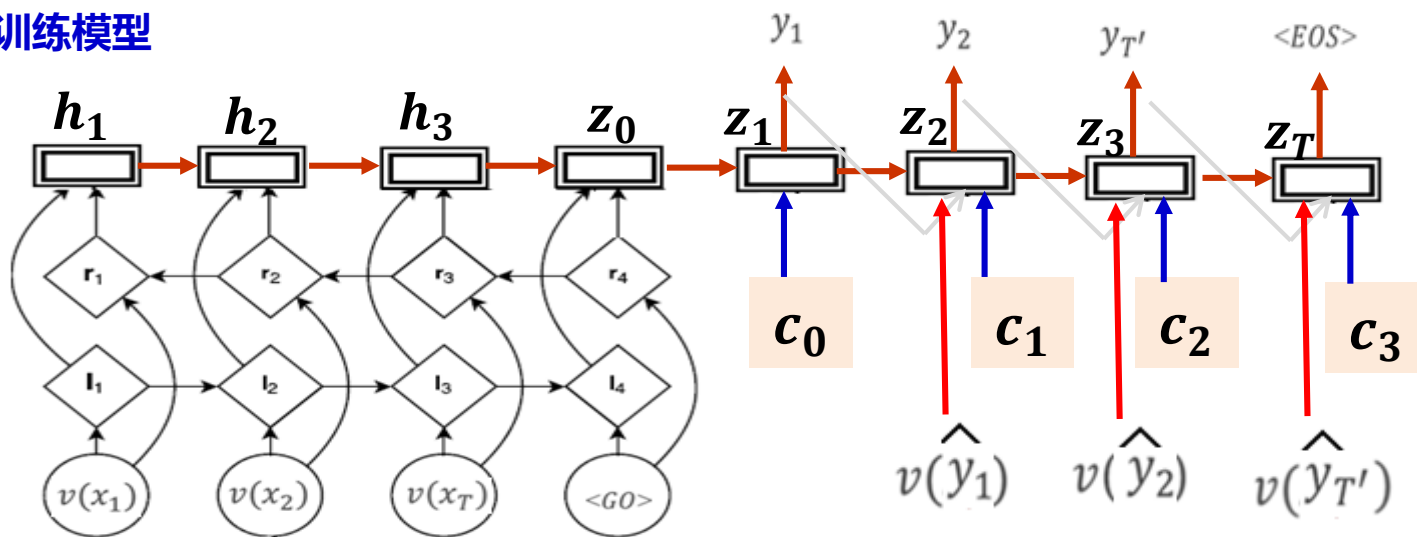
模型训练过程

实例 (x^i, \hat{y}^i)

源语: $x_1 x_2 x_T$

目标语: $\hat{y}_1 \hat{y}_2 \hat{y}_{T'}$

训练模型



2. Attention+ RNN架构翻译模型

■ Attention+ RNN模型学习:

给定训练集 $\{[x^n, y^n]\}_{n=1}^N$

- 优化目标: 最大化 $J(\theta) = \sum_{i=1}^N \log P(y^n | x^n; \theta)$

其中, 参数 θ 是包括:

- Encoder RNN cell 所有参数
- Decoder RNN cell所有参数
- Attention module的参数 (W_a 、 U_a 、 v_a)

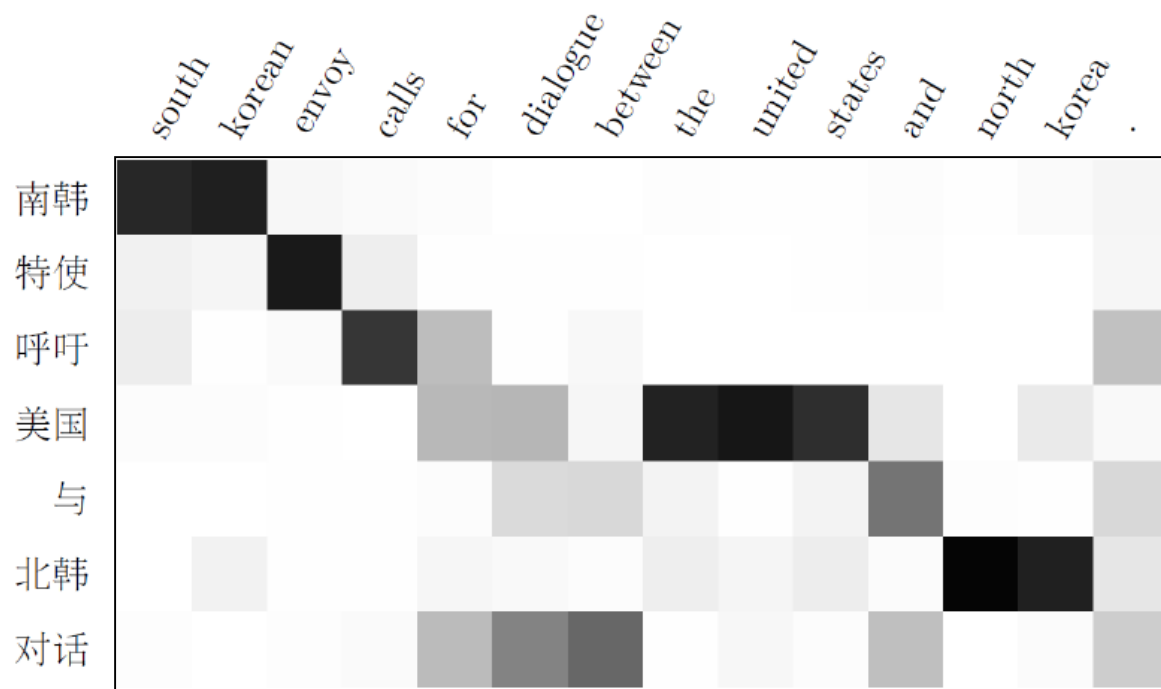
- 参数学习: 梯度下降, BPTT

$$\theta := \theta - \alpha \cdot \frac{\partial J(\theta)}{\partial \theta}$$

学得参数: θ

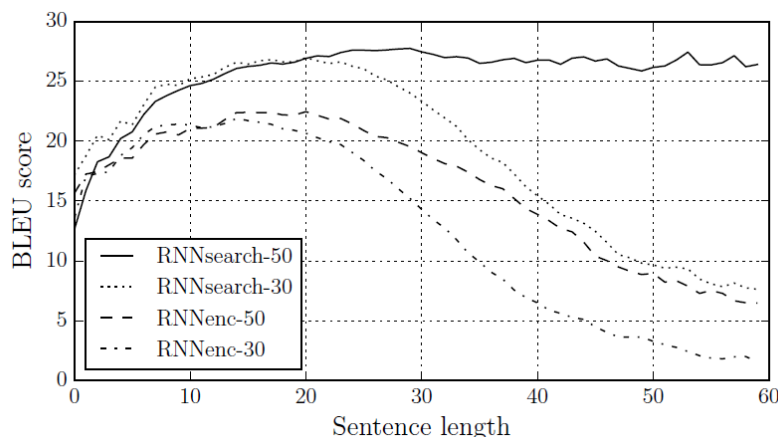
2. Attention+ RNN架构翻译模型

注意力机制可视化效果



2. Attention+ RNN架构翻译模型

注意力机制实验结果

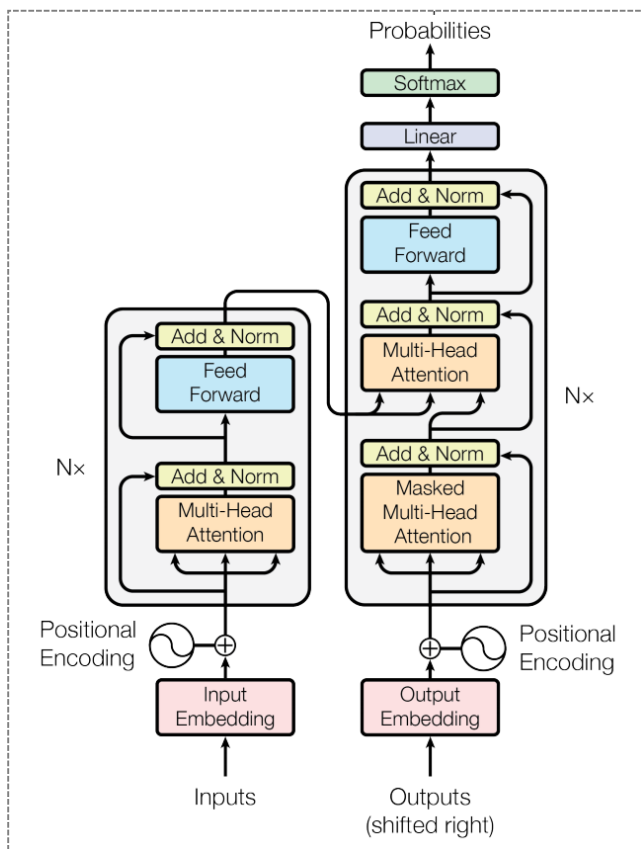


Model	All	No UNK ^o
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50*	28.45	36.15
Moses	33.30	35.63

- 在句子限长为30和50的情况下，加AM模型效果优于不加AM模型
- 句子长度增加时，加AM模型效和不加AM模型的效果均变差，但AM模型鲁棒性较好

3. Transformer 翻译模型

3. Transformer 翻译模型



前向全连接网络+多头注意力机制

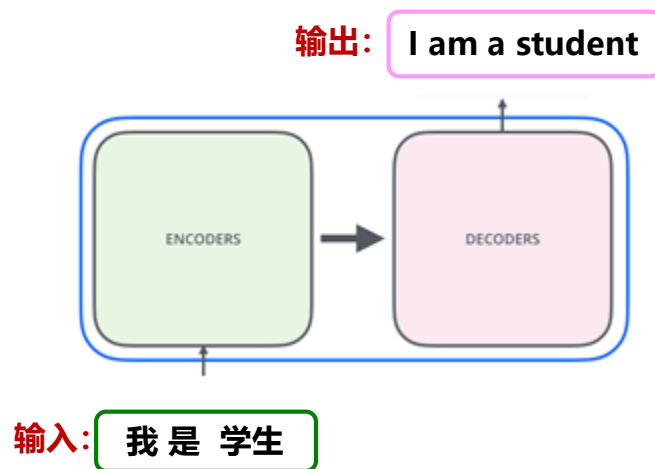
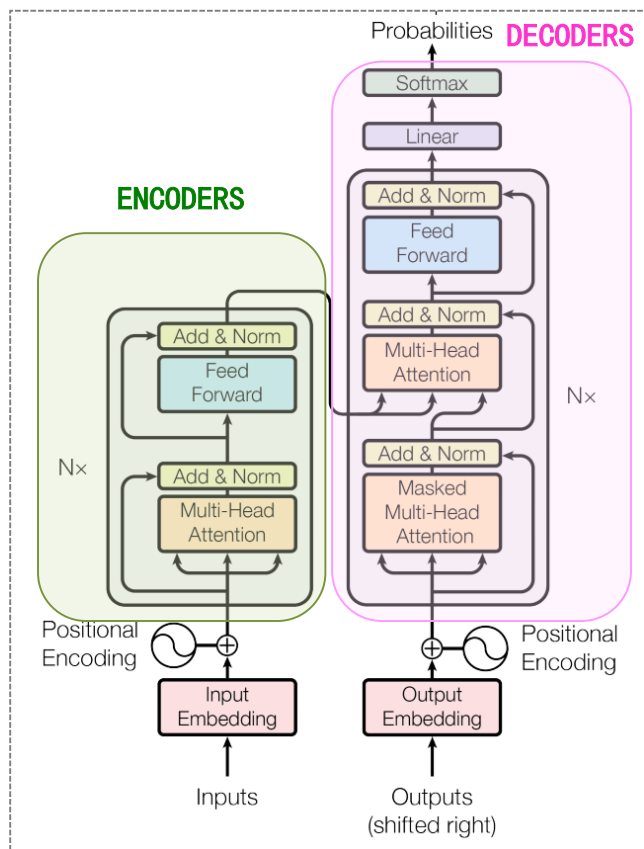
- 编码层——6层Attention堆叠，包含2个子层
- 解码层——6层Attention堆叠，包含3个子层
- 子结构——Attention

特点：

- 克服了RNN的递归无法并行的缺点，可以高度并行，训练速度快；
- 具有捕捉long distance dependency的能力，有较高的翻译质量

3. Transformer 翻译模型

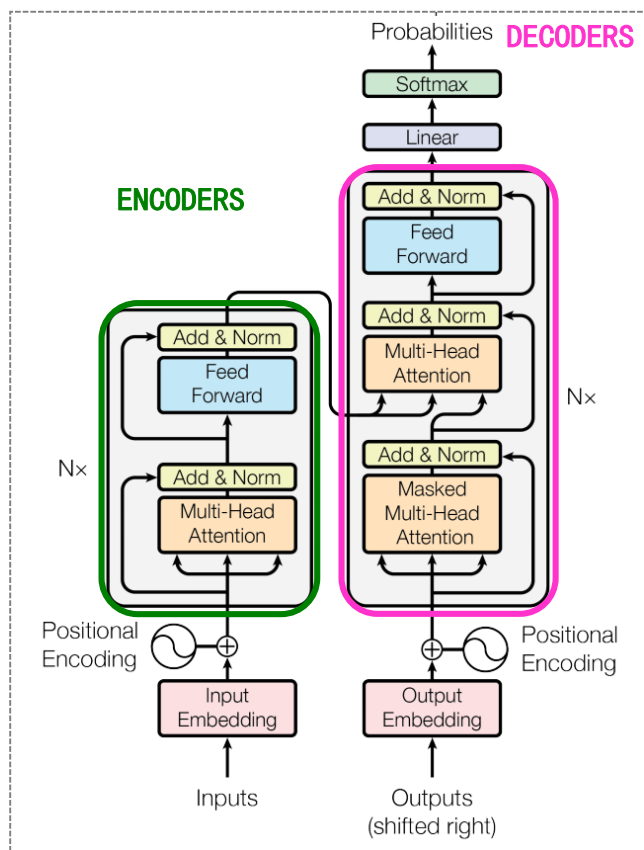
Transformer 模型结构



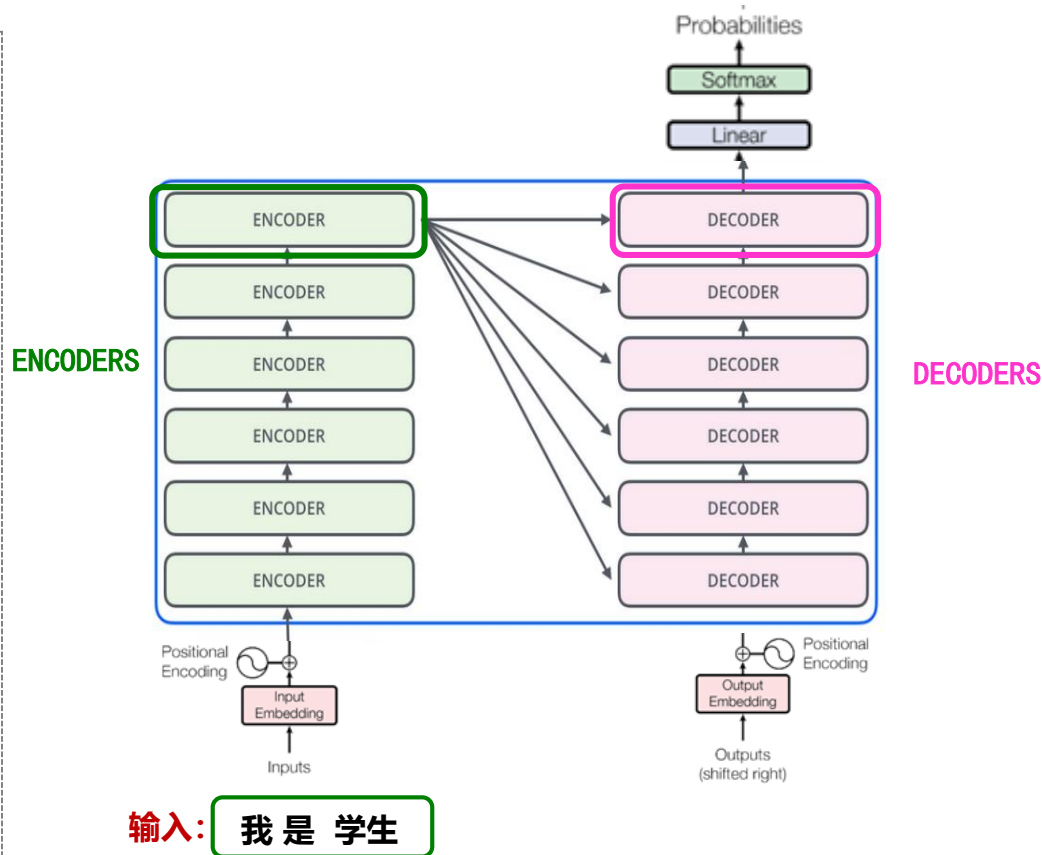
Attention is all you need

3. Transformer 翻译模型

Transformer 模型结构



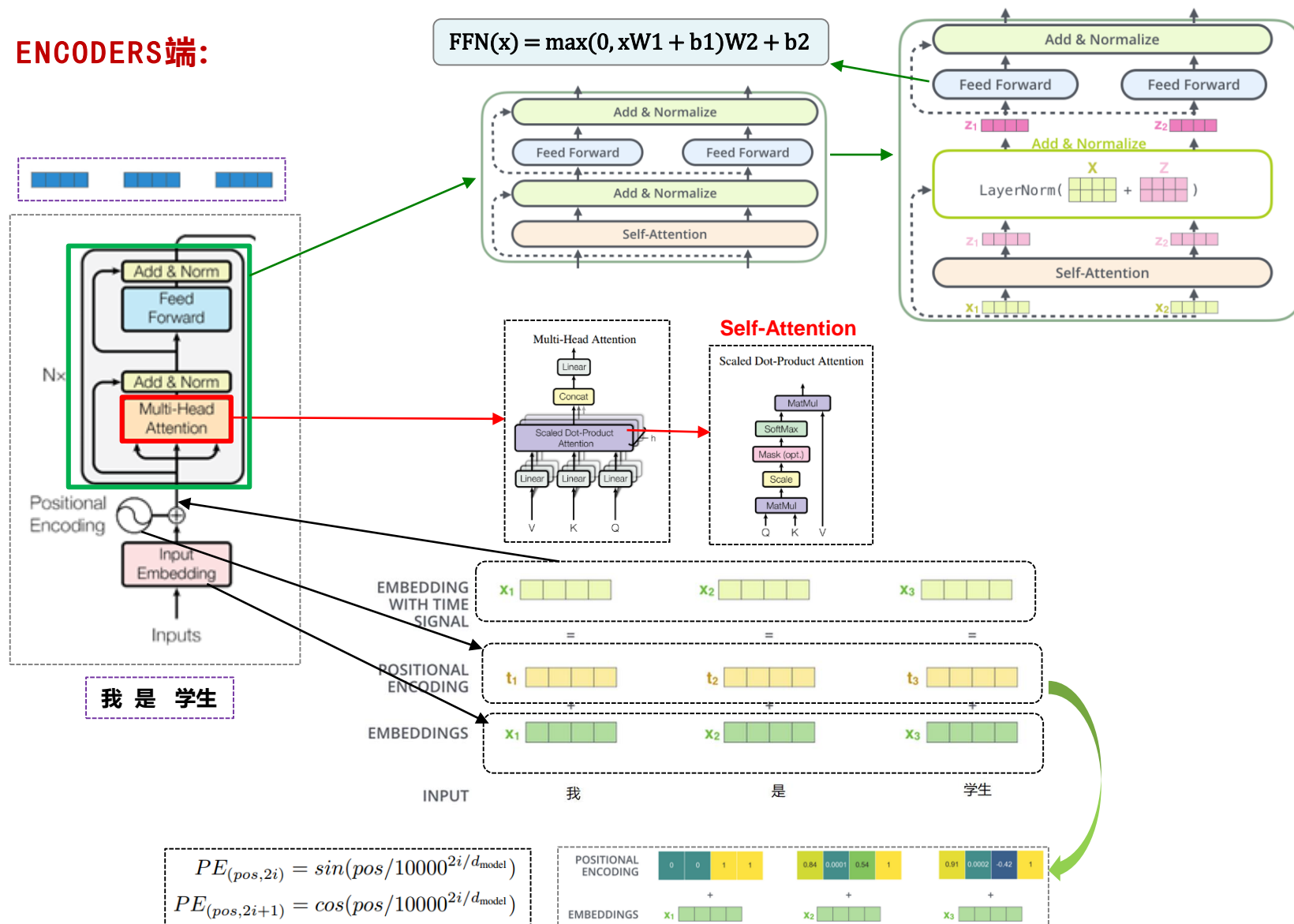
输出: I am a student



输入: 我是学生

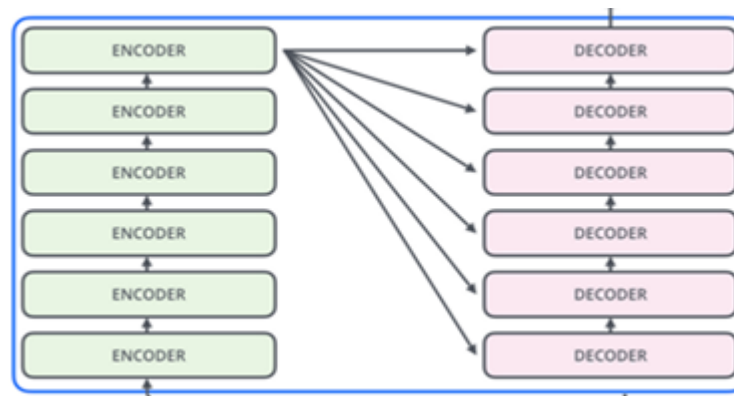
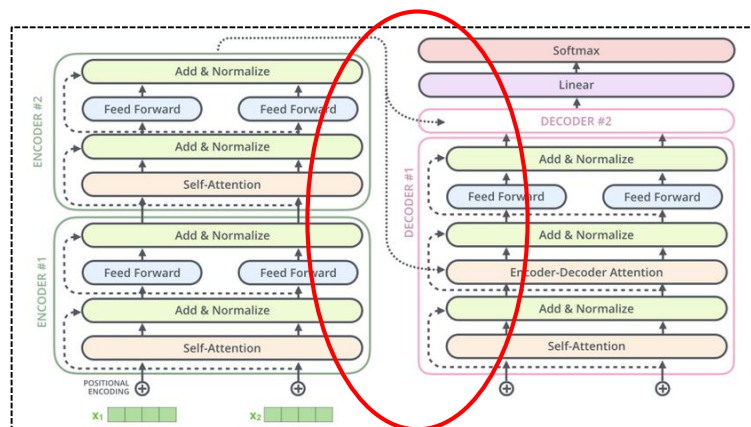
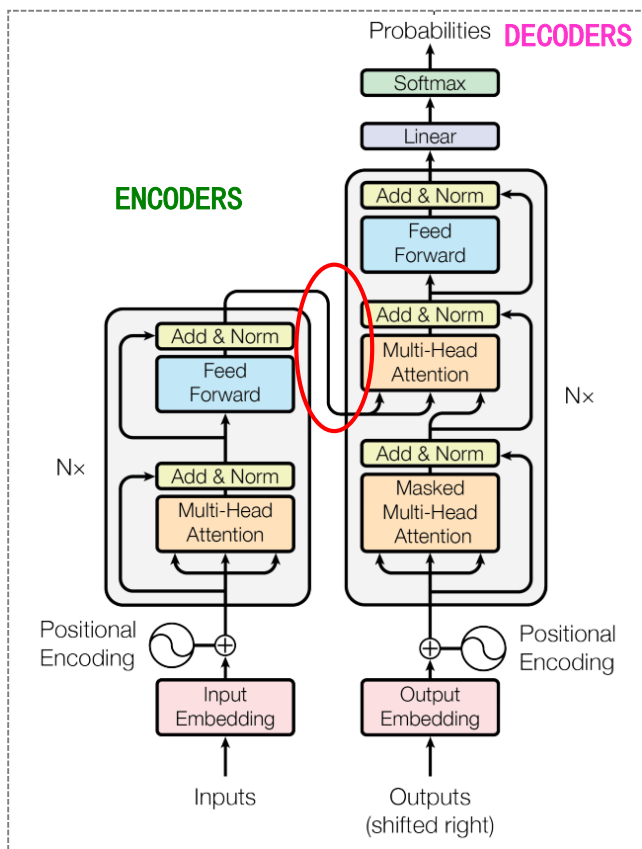
3. Transformer 翻译模型

ENCODERS端:



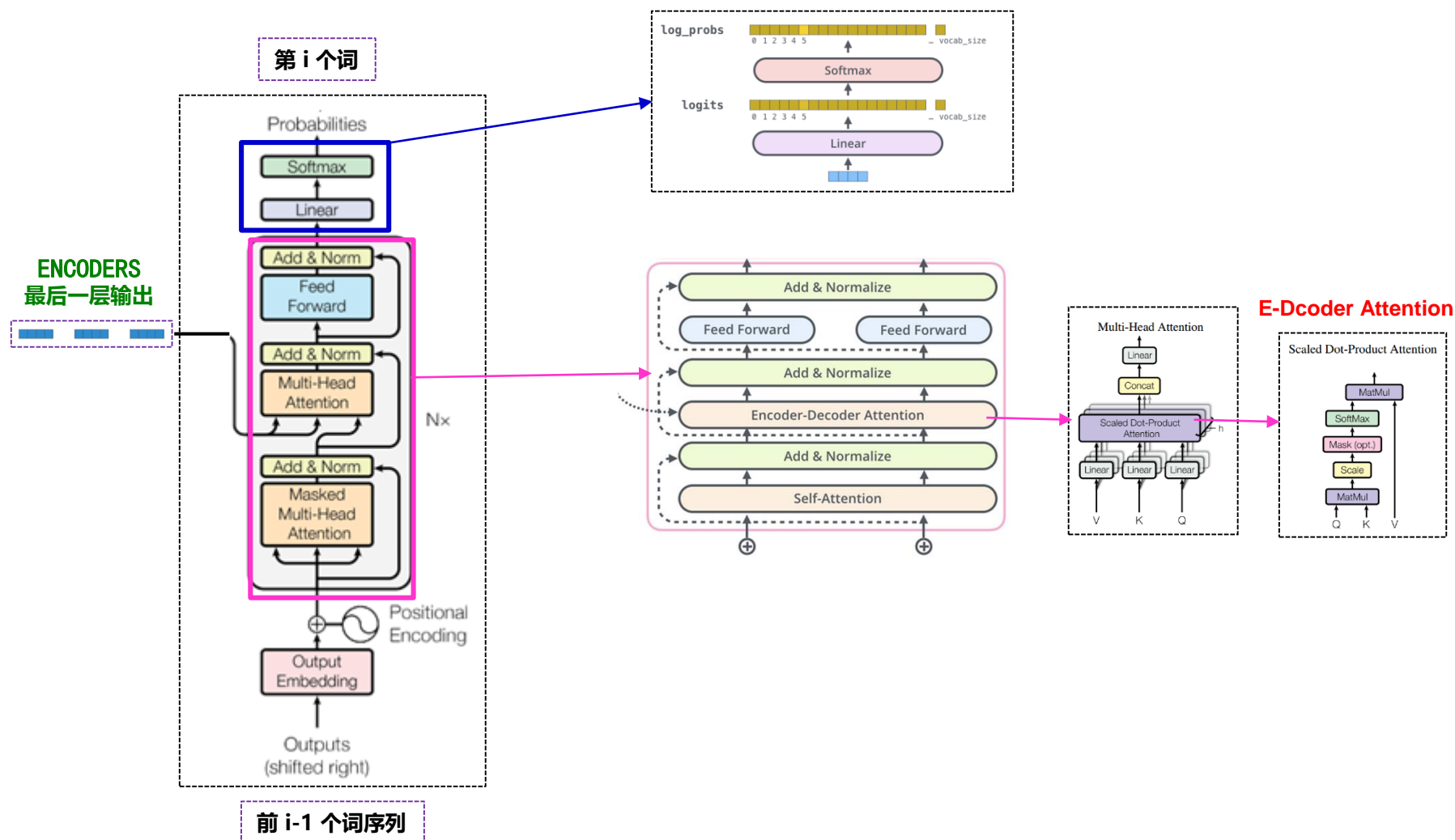
3. Transformer 翻译模型

ENCODERS端与DECODERS端连接:



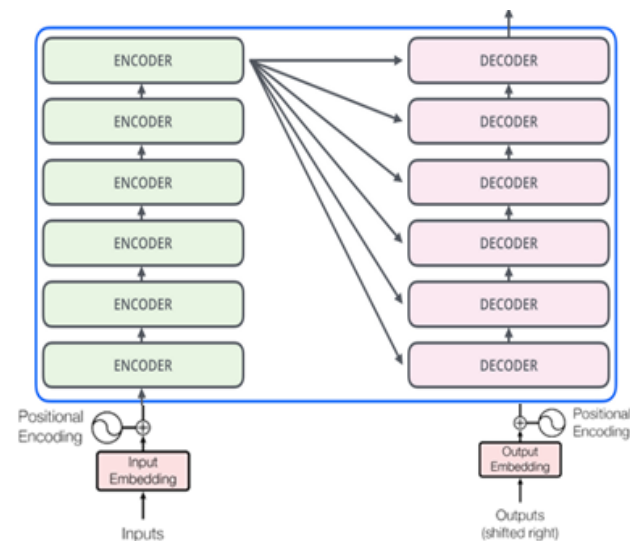
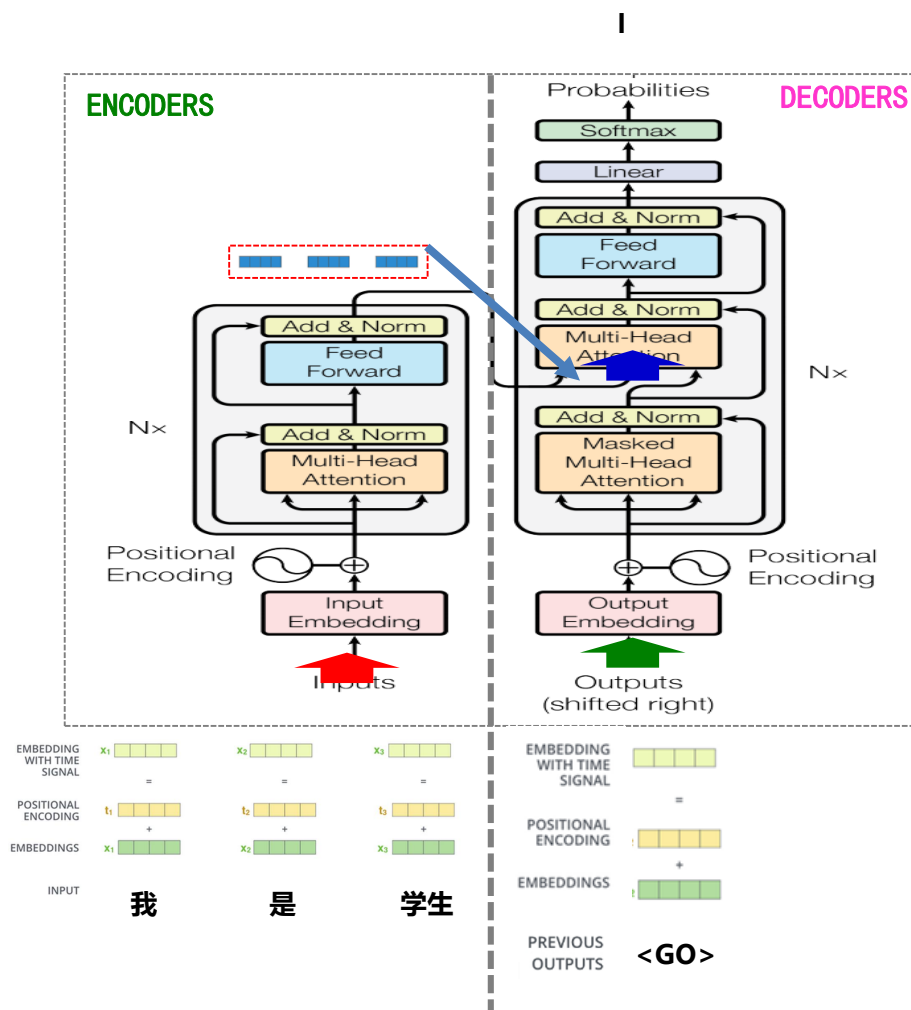
3. Transformer 翻译模型

DECODERS端:



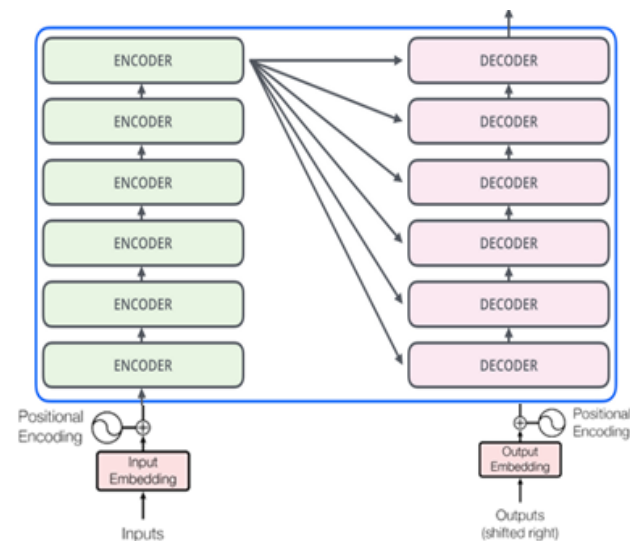
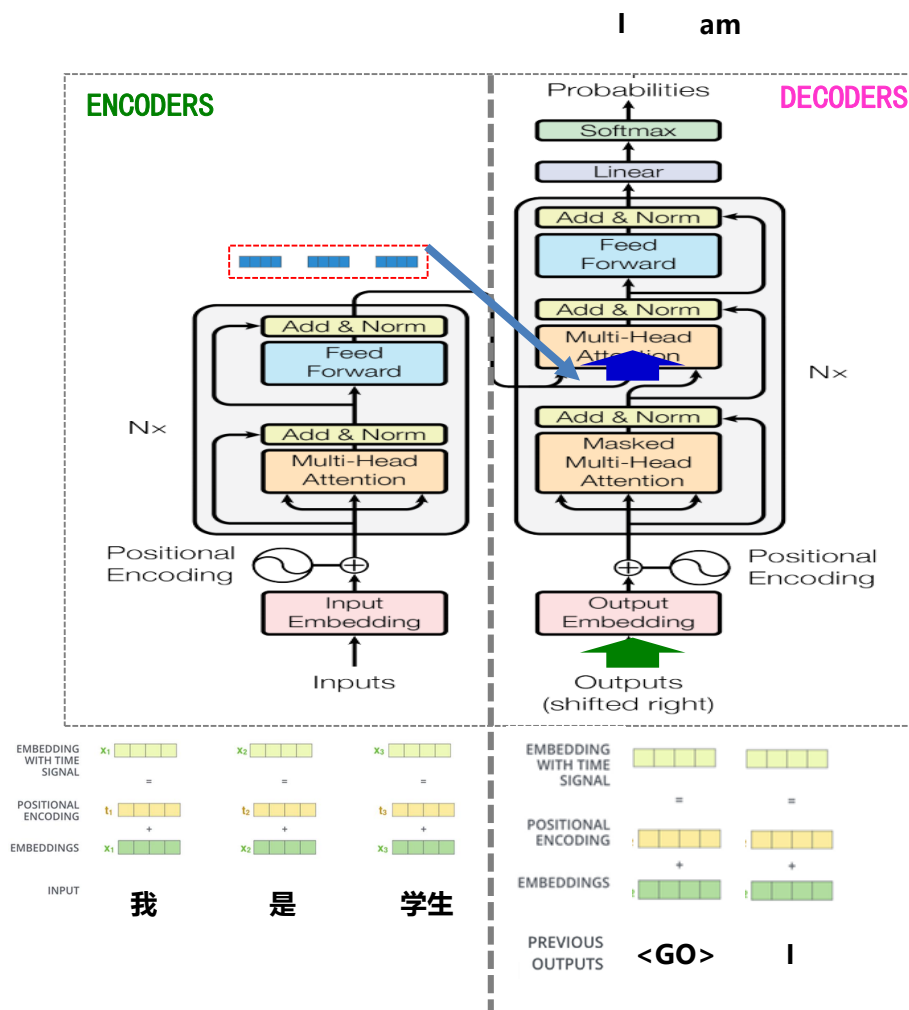
3. Transformer 翻译模型

Transformer 执行过程:



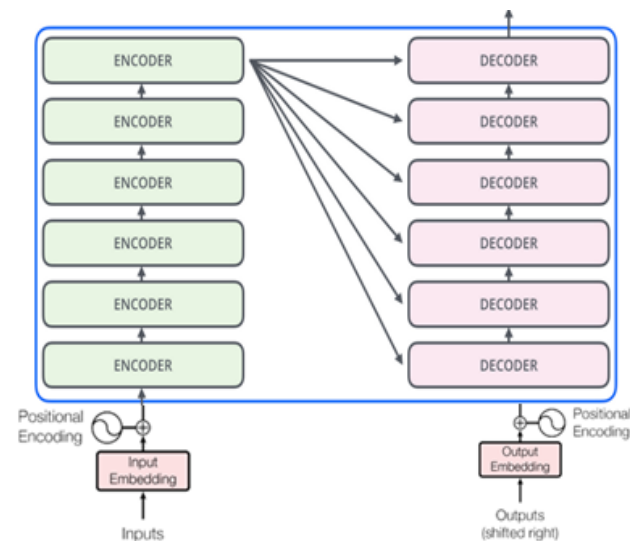
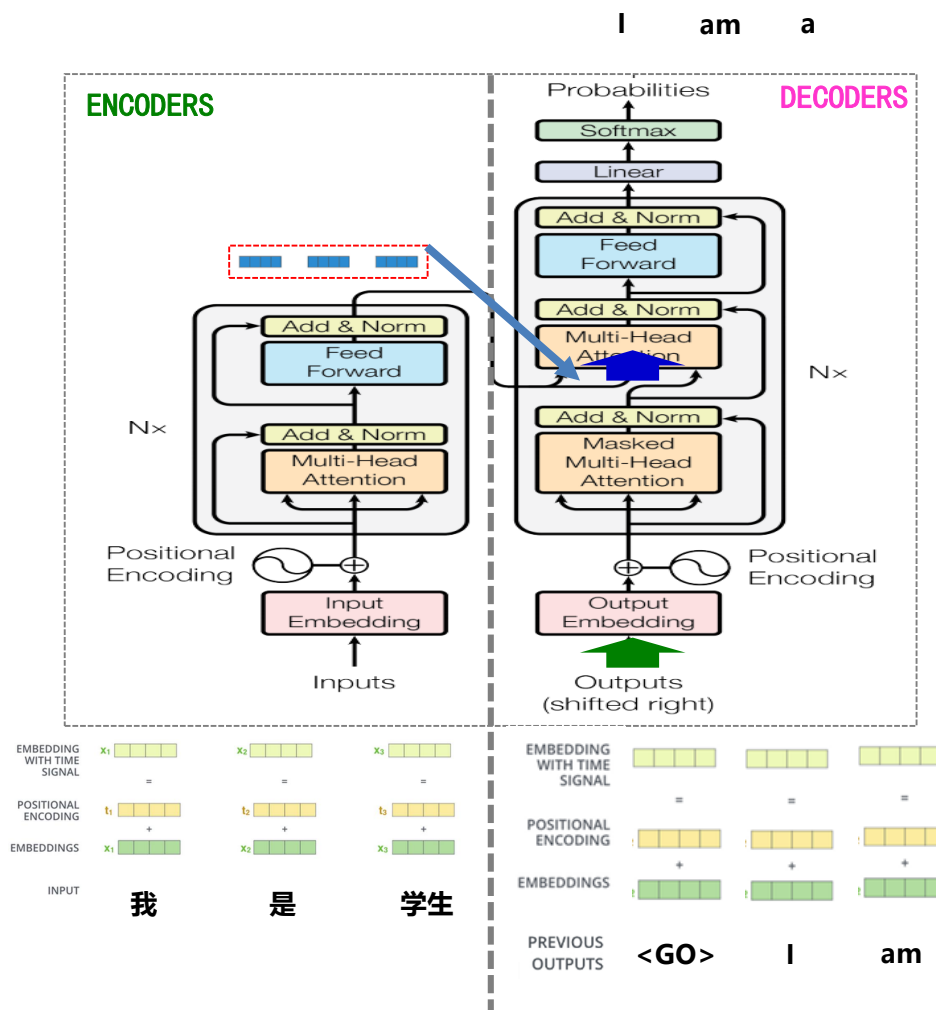
3. Transformer 翻译模型

Transformer 执行过程:



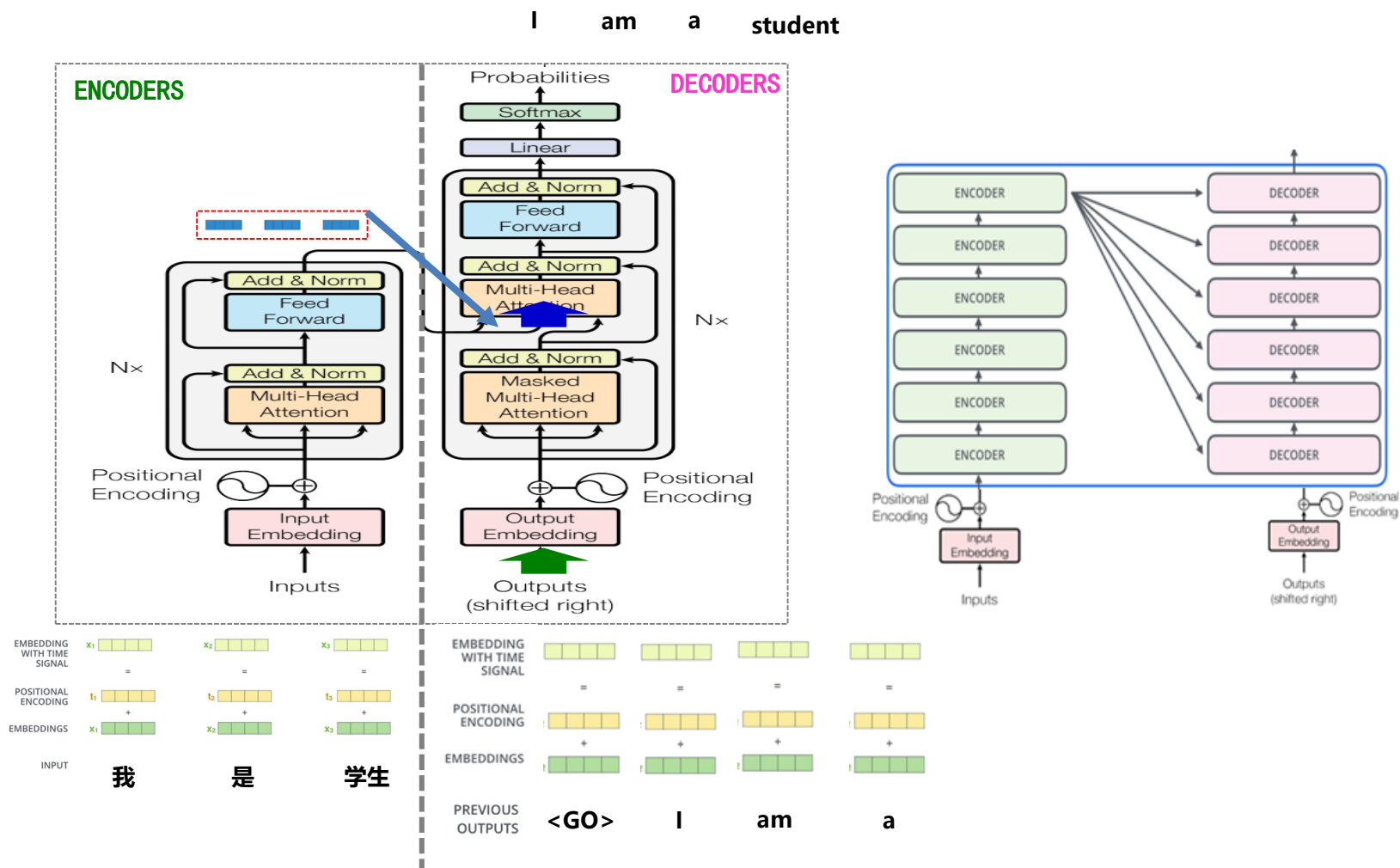
3. Transformer 翻译模型

Transformer 执行过程:



3. Transformer 翻译模型

Transformer 执行过程:



3. Transformer 翻译模型

Transformer 模型实验结果

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [15]	23.75			
Deep-Att + PosUnk [32]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [31]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [8]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [26]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [32]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [31]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [8]	26.36	41.29	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	$3.3 \cdot 10^{18}$	
Transformer (big)	28.4	41.0	$2.3 \cdot 10^{19}$	

参考文献:

宗成庆, 统计自然语言处理 (第2版) 课件

刘群, 机器翻译原理与方法讲义

马永亮, 层次短语翻译模型的实现与分析, 硕士学位论文, 2011

张浩, 面向短语统计机器翻译解码算法的研究, 硕士学位论文, 2012

Neural Machine Translation by Jointly Learning to Align and Translate 2015ICLR

Attention is all you need , 2017NIPS

在此表示感谢!

谢谢各位！

