第10章 自然语言处理

问题的提出

- 如何让计算机能够自动或半自动地理解自然语言文本，懂得人的意图和心声？
- 如何让计算机实现海量语言文本的自动处理、挖掘和有效利用，满足不同用户的各种需求，实现个性化信息服务？

自然语言处理

Natural Language Processing, NLP

自然语言理解-NLP Understanding

- 自然语言(Natural Language)的复杂性
  - 话说普京与特朗普赛跑，普京第一，特朗普第二
  - 俄罗斯媒体：在国际领导人赛跑中，普京勇夺冠军，特朗普倒数第一！
  - 美国媒体：在国际领导人赛跑中，特朗普勇夺亚军，普京倒数第二！

- 沈向洋-“自然语言是人工智能皇冠上的明珠”
- 人工智能的终极目标！

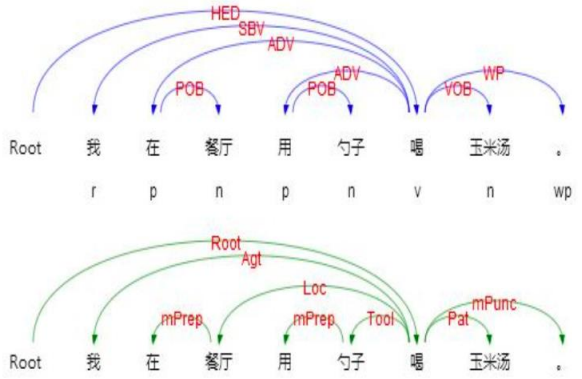
基本概念

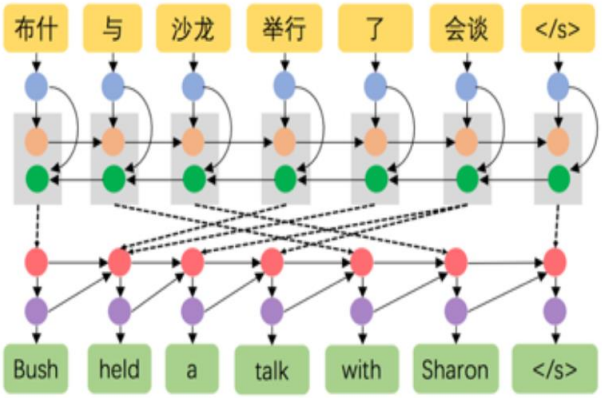
◆自然语言处理 (Natural Language Processing, NLP)

自然语言处理是研究如何利用计算机技术对语言文本（句子、篇章或话语等）进行处理和加工的一门学科，研究内容包括对词法、句法、语义和语用等信息的识别、分类、提取、转换和生成等各种处理方法和实现技术。

《计算机科学技术百科全书》（宗成庆）

自然语言处理





分词、 词性标注、命名实体识别、指代消解、 句法分析、语法分析...

关键词抽取、 文本分类、 情感分析与观点挖掘、信息抽取...

机器翻译、 自动摘要、问答系统、对话系统...

基本问题

句法 (Syntax) 问题

研究句子结构成分之间的相互关系和组成句子序列的规则 。

为什么一句话可以这么说也可以那么说?

如何建立快速有效的句子结构分析方法?

苹果，我吃了。

我吃了苹果。

≠ 苹果吃了我。

基本问题

语义 (Semantics) 问题

研究如何从一个语句中词的意义，以及这些词在该语句中句法结构中的作用来推导出该语句的意义。

这句话说了什么?

(1) 苹果不吃了
(2) 这个人真牛
(3) 这个人眼下没些什么
(4) 火烧圆明园/火烧驴肉

基本问题

语用学(Pragmatics) 问题

研究在不同上下文中语句的应用，以及上下文对语句理解所产生的影响。从狭隘的语言学观点看，语用学处理的是语言结构中有形式体现的那些语境。相反，语用学最宽泛的定义是研究语义学未能涵盖的那些意义。

为什么要说这句话?

（1）火，火！
（2）看看鱼怎么样了？

主要困难

- **困难之一：大量歧义(ambiguity)现象**
  - **词法歧义**

例如: (1) I'll see Prof. Zhang home.

(2)计算机研究所取得的成就

计算机/研究所/取得/的/成就

计算机/研究/所/取得/的/成就

(3)门把手弄坏了

门/ 把/ 手/ 弄/ 坏/ 了

门把手/ 弄/ 坏/ 了

主要困难

文章标题中的歧义比比皆是：

- **上大学子烛光追思钱伟长**

  (新浪网： http://www.sina.com.cn/, 2010.8.8)

- **教育部长跑活动负责人与商家总经理被曝系师生**

  (科学网： http://news.sciencenet.cn/，2010-11-14)

主要困难

❖ **词性歧义**

①介词：像，好似；②动词：喜欢

(1) Time flies like an arrow.

①动词：飞，飞翔，飞驰
②名词：苍蝇，飞虫

✧ 时间像箭一样飞驰（光阴似箭）。

✧ 时间苍蝇喜欢箭（有一种苍蝇叫“时间”）。

(2) “动物保护警察”明年上岗

(《环球时报》2010年9月25日，第10版)

主要困难

❖ 结构歧义

(1) 喜欢乡下的孩子。

(2) 关于鲁迅的文章。

(3) 今天中午吃**馒头**。 (4) 今天中午吃**食堂**。

(5) 今天中午吃**大碗**。 (6) 今天中午吃了**闭门羹**。

(7) 写文章/ 写毛笔/ 写黑板

主要困难

(8) I saw a man with a telescope.

➔ I saw [a man with a telescope].

I [saw a man] with a telescope.

➔ I saw a man with a telescope in the park. ?

英语句子歧义组合的开塔兰数(Catalan Numbers) C_n:

$$C_n = \binom{2n}{n}\frac{1}{n+1} \qquad \text{其中：} \binom{2n}{n} = \frac{(2n)!}{n!\times n!}$$

***n* 为句子中介词短语的个数。**

主要困难

❖ 语义歧义

他说:“她这个人真有意思(funny)”。她说:“他这个人怪有意思的(funny)”。于是人们以为他们有了意思(wish)，并让他向她意思意思(express)。他火了:“我根本没有那个意思(thought)”!她也生气了:“你们这么说是什么意思(intention)”?事后有人说:“真有意思(funny)”。也有人说:“真没意思(nonsense)”。

-《生活报》1994. 11. 13. 第6版

人们的语言表达中大量地使用缩略语和隐喻的表达方式，如：

要把权力装进制度的笼子；老虎苍蝇一起打。

破四旧，除四害；消灭一切牛鬼蛇神。

主要困难

- **多音字及韵律等歧义**

一 语音合成面临的诸多问题

(1) 一字多音

例如：尾巴、亲家、削铅笔、一行

(2) 韵律、声调、语气、重音

例如：药材好药才好。

他的钱包被偷了。

今日说法/小心地滑/聊吧/说吧

主要困难

- **困难之二：大量未知语言现象**
  - 新词、人名、地名、术语等，如：裸退、非典、夏天、高山、温馨、时光、吉林、不来梅、失联
  - 新含义

    如：苹果、奔腾、同志、小姐、老虎、苍蝇等
  - 新用法和新句型等，尤其在口语中或部分网络语言中，不断出现一些“非规范的”新的语句结构。如：被长工资，很中国，百度一下

新冠

主要困难

- **归纳起来，NLU 所面临的挑战：**
  - **普遍存在的不确定性**：词法、句法、语义、语用和语音各个层面
  - **未知语言现象的不可预测性**：新的词汇、新的术语、新的语义和语法无处不在
  - **始终面临的数据不充分性**：有限的语言集合永远无法涵盖开放的语言现象
  - **语言知识表达的复杂性**：语义知识的模糊性和错综复杂的关联性难以用常规方法有效地描述，为语义计算带来了极大的困难

主要困难

- **机器翻译中映射单元的不对等性：** 词法表达不相同、句法结构不一致、语义概念不对等

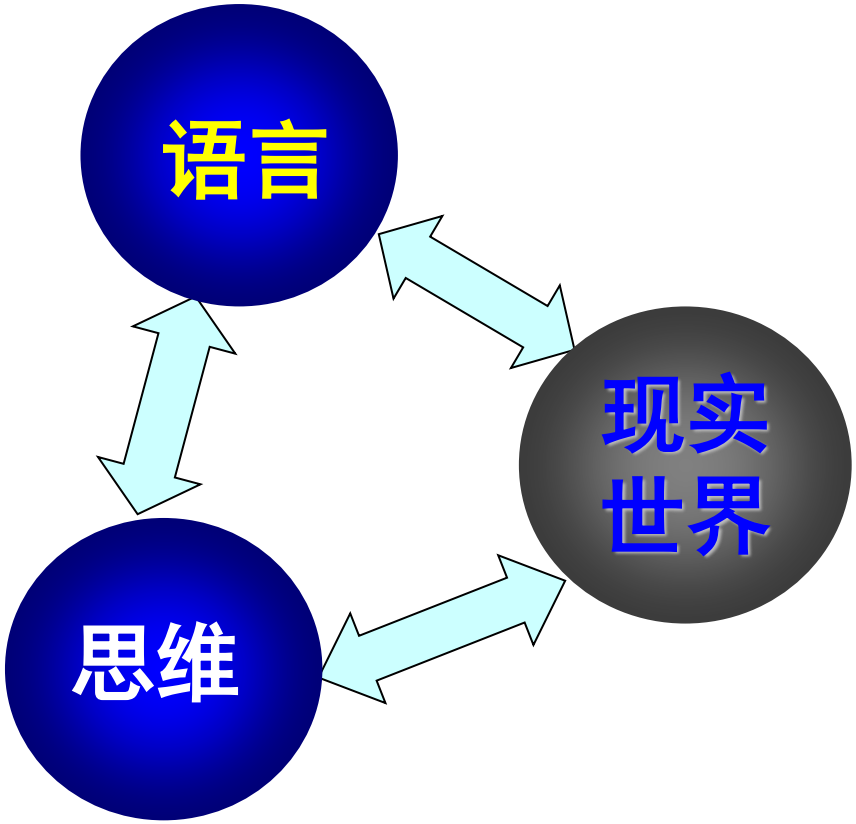
从大量复杂多样的不确定性中寻找确定性结论

主要困难

◆人脑理解语言是一个复杂的思维过程

- 语言学、心理学
- 逻辑学、认知科学
- 计算机科学
- 统计学、信息论
- 背景知识、常识等

……



主要困难

人脑的语言认知过程到底怎样？

什么意思？

基本研究方法

◆**理性主义**：通常通过一些特殊的语句或语言现象的研究来得到对人的语言能力的认识，而这些语句和语言现象在实际的应用中并不常见。

- **问题求解的基本思路：基于规则的分析方法建立符号处理系统**
  - **规则库开发：N + N → NP**
  - **词典标注：　#工作，N(uc)；V；**
  - **推导算法设计：归约、推导、歧义消解方法…**

知识库 + 推理系统 → NLP 系统

理论基础：Chomsky 的文法理论

基本研究方法

◆**经验主义**：偏重于对大规模语言数据中人们所实际使用的普通语句的统计。

- **求解问题的思路：基于大规模真实语料(语言数据)建立计算方法**
  - **大规模真实数据的收集、标注**：真实性、代表性、标注信息 ……
  - **统计模型建立**：模型的复杂性、有效性、参数训练方法 ……

语料库 ＋ 统计模型 → NLP 系统

理论基础：统计学、信息论、机器学习

研究现状

- **各种理论问题：**
  从词法(汉语分词)到语义 ……
- **各种应用系统：**
  从机器翻译到信息抽取 ……

哪个问题已经解决了？

哪个问题都没彻底解决！

研究现状

◆ 基本现状

- 部分问题得到了解决，可以为人们提供辅助性帮助，如：专业领域文档翻译，电子词典，搜索引擎，文字录入等；
- 基础问题研究仍任重而道远，如：语义表示和计算、高质量的自动翻译等；
- 社会需求日益迫切：信息服务、通讯、网络内容管理、情报处理、国家安全等；
- 许多技术离真正实用的目标还有相当的距离，尚未建立起有效、完善的理论体系。

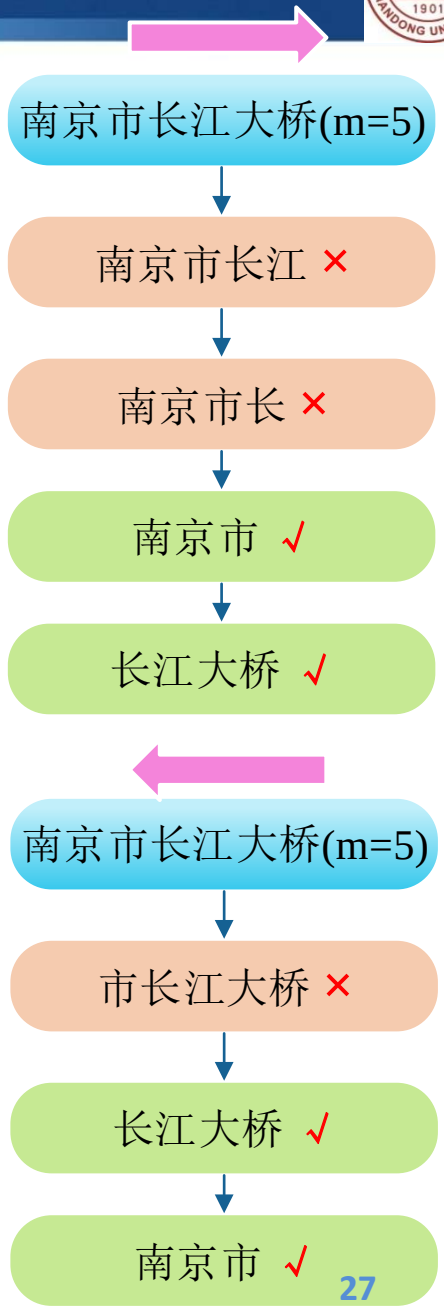
研究现状

中文分词

- 基于词典(规则)的方法
  - 按照一定策略将待分析的汉字串与一个“词典”中的词条进行匹配，如果匹配成功，那么该汉字串就是一个词。
- 基于统计的方法
  - 训练：根据观测到的数据(人工标注好的语料)的统计特征对模型参数进行估计。
  - 分词：通过模型计算各种分词出现的概率，将概率最大的分词结果作为最终结果。

中文分词-基于词典(规则)的方法

- 按照扫描方向：正向匹配和逆向匹配
- 按照扫描长度：最大匹配和最小匹配
- 正向最大匹配
  - 从左向右取待切分汉语句的*m*个字符作为匹配字段(*m*为词典中最长词条个数)；
  - 查找词典并进行匹配；
    - 若匹配成功，则将这个匹配字段作为一个词切分出来；
    - 若匹配不成功，则将这个匹配字段的最后一个字去掉，剩下的字符串作为新的匹配字段，进行再次匹配，重复以上过程，直到切分出所有词为止。
  - 例：南京市长江大桥 (*m*=5)
- 逆向最大匹配
- 双向最大匹配
  - 双向最大匹配法是将正向最大匹配法得到的分词结果和逆向最大匹配法的到的结果进行比较，把所有可能的最大词都分出来。



中文分词-基于统计的方法

- *n*-gram：基于假设，第*n*个词的出现只与前面*n*-1个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。
  - 一个句子$S=\{t_1 t_2 t_3 \ldots t_N\}$
  - 句子出现的概率：y是分词序列 $y^* = \arg\max_y P(y \mid S) = \prod_{i=1}^{N} P(t_i \mid t_1 \ldots t_{i-1})$
  - Unigram：$P_{\text{uni}}(y|S) = P(t_1)P(t_2)P(t_3)P(t_4)\ldots P(t_N)$
  - Bigram：只考虑前一个词项的出现情况，

    $P_{\text{bi}}(y|S) = P(t_1)P(t_2|t_1)P(t_3|t_2)\ldots P(t_N|t_{N-1})$
  - Trigram

南京市长江大桥

基本词表

南 | 京 | 市 | 长 | 江 | 大 | 桥

Bigram示例

南 | 京 | 市 | 长 | 江 | 大 | 桥	$P(S)=P(南)P(京\|南)P(市\|京)P(长\|市)P(江\|长)P(大\|江)P(桥\|大)$ ×
南京 | 市长 | 江 | 大桥	$P(S)=P(南京)P(市长\|南京)P(江\|市长)P(大桥\|江)$ ×
南京市 | 长江 | 大桥	$P(S)=P(南京市)P(长江\|南京市)P(大桥\|长江)$ ×
南京市 | 长江大桥	$P(S)=P(南京市)P(长江大桥\|南京市)$ √ **Max**

隐马尔可夫模型
Hidden Markov Model, HMM

隐马尔可夫模型

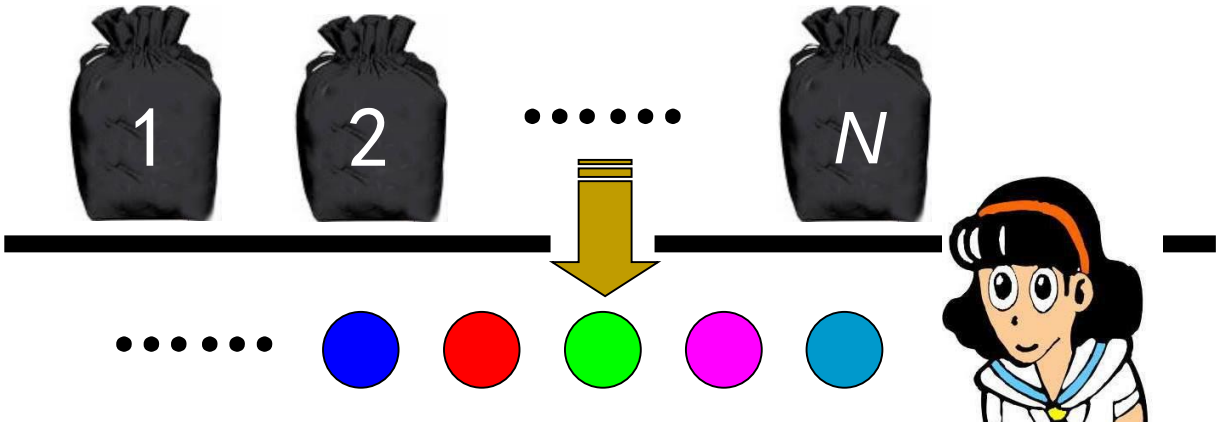
◆隐马尔可夫模型(Hidden Markov Model, HMM)

创建于20世纪70年代，是美国数学家鲍姆(Leonard E. Baum)等人提出来的。

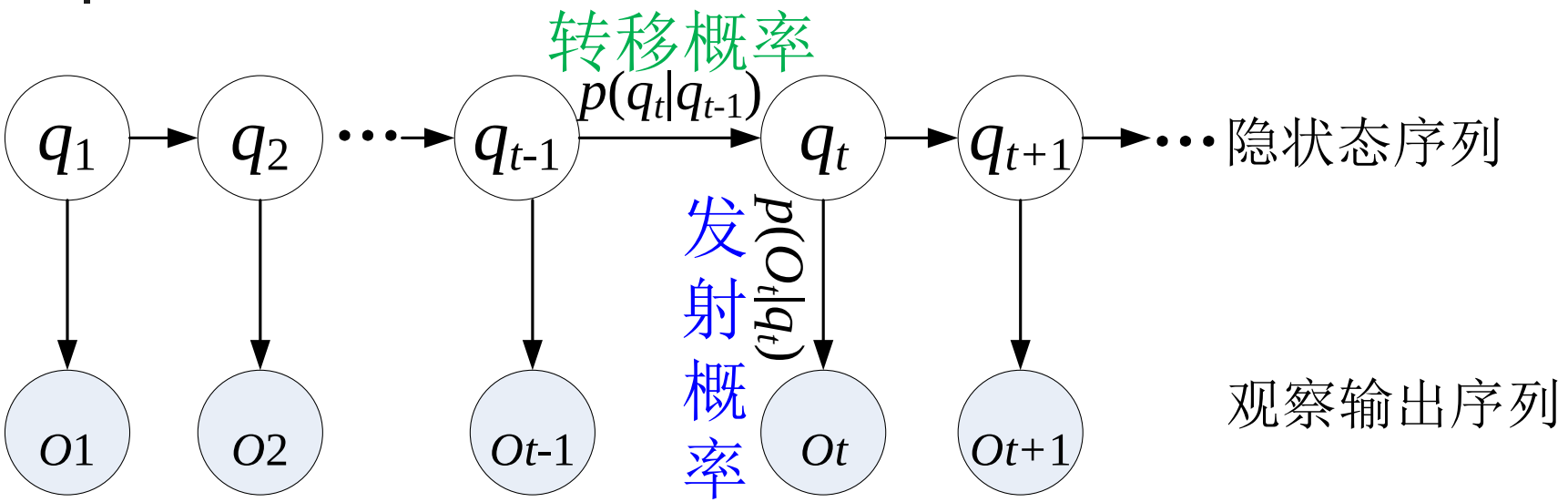
描写：该模型是一个双重随机过程，我们不知道具体的状态序列，只知道状态转移的概率，即模型的状态转换过程是不可观察的（隐蔽的），而可观察事件的随机过程是隐蔽状态转换过程的随机函数。

隐马尔可夫模型

例如：N 个袋子，每个袋子中有 M 种不同颜色的球。一实验员根据某一概率分布选择一个袋子，然后根据袋子中不同颜色球的概率分布随机取出一个球，并报告该球的颜色。对局外人：可观察的过程是不同颜色球的序列，而袋子的序列是不可观察的。每只袋子对应HMM中的一个状态；球的颜色对应于 HMM 中状态的输出。



隐马尔可夫模型



HMM 图解

隐马尔可夫模型

◆HMM 的组成

1. 模型中的状态数为 N (袋子的数量)
2. 从每一个状态可能输出的不同的符号数 M (不同颜色球的数目)

隐马尔可夫模型

3. 状态转移概率矩阵 $A=a_{ij}$ (a_{ij} 为实验员从一只袋子(状态 S_i) 转向另一只袋子(状态 S_j) 的概率)。其中,

$$\begin{cases} a_{ij} = p(q_{t+1} = S_j \mid q_t = S_i), \quad 1 \le i, j \le N \\ a_{ij} \ge 0 \\ \sum_{j=1}^{N} a_{ij} = 1 \end{cases} \qquad \ldots (6.6)$$

隐马尔可夫模型

4. 从状态 S_j 观察到某一特定符号 v_k 的概率分布矩阵为：

$B=b_j(k)$

其中，$b_j(k)$为 实验员从第 j 个袋子中取出第 k 种颜色的球的概率。那么，

$$\begin{cases} b_j(k) = p(O_t = v_k \mid q_t = S_j), \quad 1 \le j \le N, \quad 1 \le k \le M \\ b_j(k) \ge 0 \\ \sum_{k=1}^{M} b_j(k) = 1 \end{cases}$$

隐马尔可夫模型

5. 初始状态的概率分布为：$\pi = \pi_i$，其中，

$$\begin{cases} \pi_i = p(q_1 = S_i), \quad 1 \le i \le N \\ \pi_i \ge 0 \\ \sum_{i=1}^{N} \pi_i = 1 \end{cases}$$

为了方便，一般将 HMM 记为：$\mu = (A, B, \pi)$

或者 $\mu = (S, O, A, B, \pi)$ 用以指出模型的参数集合。

隐马尔可夫模型

◆给定HMM求观察序列

给定模型 $\mu=(A,B,\pi)$，产生观察序列 $O=O_1O_2\ldots O_T$：

(1) 令 $t=1$;

(2) 根据初始状态分布 $\pi=\pi_i$ 选择初始状态 $q_1=S_i$；

(3) 根据状态 S_i 的输出概率分布 $b_i(k)$, 输出 $O_t=v_k$；

(4) 根据状态转移概率 a_{ij}，转移到新状态 $q_{t+1}=S_j$;

(5) $t=t+1$, 如果 $t<T$, 重复步骤 (3) (4), 否则结束。

隐马尔可夫模型

◆三个问题：

(1) 在给定模型$\mu=(A, B, \pi)$ 和观察序列 $O=O_1O_2 \ldots O_T$ 的情况下，怎样快速计算概率 $p(O|\mu)$?

(2) 在给定模型 $\mu=(A, B, \pi)$ 和观察序列 $O=O_1O_2 \ldots O_T$ 的情况下，如何选择在一定意义下“最优”的状态序列 $Q = q_1 q_2 \ldots q_T$，使得该状态序列“最好地解释”观察序列?

(3) 给定一个观察序列 $O=O_1O_2 \ldots O_T$，如何根据最大似然估计来求模型的参数值？即如何调节模型的参数，使得$p(O|\mu)$ 最大?

基于HMM的中文分词方法

- 给定一个观察序列(句子)$X=x_1x_2\ldots x_t\ldots x_T$, 其中$x_t$是一个字、词等文字单元
- 假设X的状态序列(如词的开始符、词的结束符等)为Y
  - $y_t(i)$有M个状态　　$Y = y_1(i)y_2(i)...y_t(i)...y_T(i),\ 1 \le i \le M$

$$Y^* = \arg\max_Y P(Y \mid X) = \arg\max_Y \frac{P(Y,X)}{P(X)} \propto \arg\max_Y P(X \mid Y)P(Y)$$

1) 独立性假设 $\Rightarrow P(X \mid Y) = \prod_{t=1}^{T} P(x_t \mid y_t)$

2) 马尔可夫(一阶)假设：$P(y_t \mid y_{t-1}y_{t-2}\ldots y_1) = P(y_t \mid y_{t-1}) \Rightarrow P(Y)=P(y_1)\prod_{t=2}^{T} P(y_t \mid y_{t-1})$

$$Y^* = \arg\max_Y P(y_1)\prod_{t=1}^{T} P(x_t \mid y_t)\prod_{t=2}^{T} P(y_t \mid y_{t-1})$$

$$= \arg\max_Y P(y_1)P(x_1 \mid y_1)\prod_{t=2}^{T}[P(y_t \mid y_{t-1})P(x_t \mid y_t)]$$

$$= \arg\max_Y [P(y_1)P(x_1 \mid y_1)][P(y_2 \mid y_1)P(x_2 \mid y_2)][P(y_3 \mid y_2)P(x_3 \mid y_3)]\ldots$$

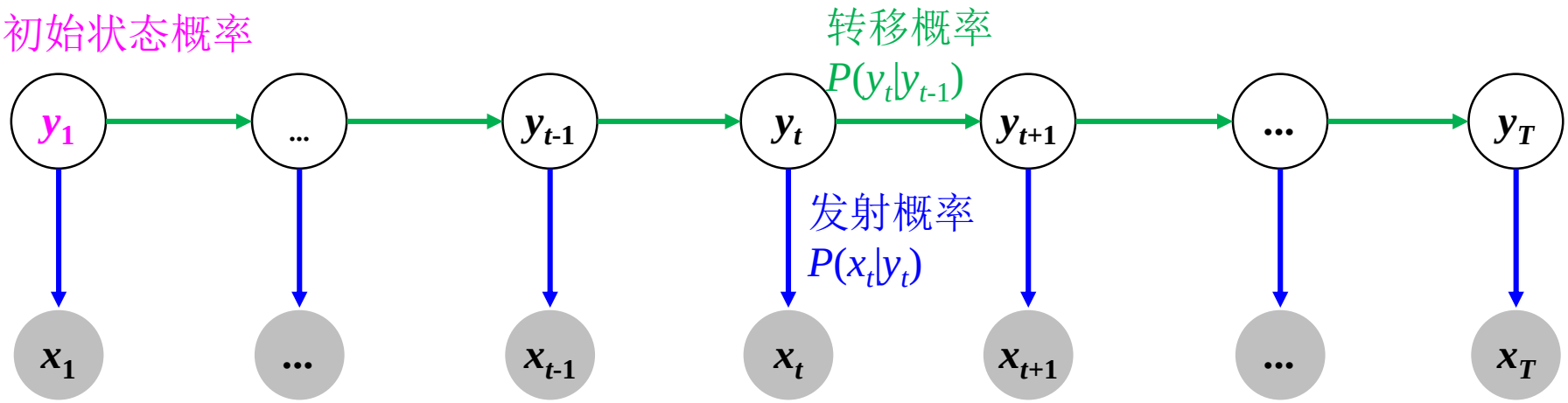
$$Y^* = \arg\max_Y P(y_1)\prod_{t=1}^{T} P(x_t \mid y_t)\prod_{t=2}^{T} P(y_t \mid y_{t-1})$$

$$= \arg\max_Y P(y_1)P(x_1 \mid y_1)\prod_{t=2}^{T}[P(y_t \mid y_{t-1})P(x_t \mid y_t)]$$

$$= \arg\max_Y [P(y_1)P(x_1 \mid y_1)][P(y_2 \mid y_1)P(x_2 \mid y_2)][P(y_3 \mid y_2)P(x_3 \mid y_3)]...$$

- 发射概率 $P(x_t|y_t)$
- 转移概率 $P(y_t|y_{t-1})$
  - Y共有M个状态 $\sum_{i=1}^{M} P(y_t(i) \mid y_{t-1}) = 1$
- 初始状态概率 $P(y_1(i))$ (i=1...M)
- HMM包括隐层状态Y，观察序列X，状态转移概率A，符号发射概率B，和初始状态概率分布π。
- HMM表示为μ={A,B, π}，参数通过训练集来学习获得。

HMM可以用有向图模型来表示，因为states (Y)与observations (X)之间存在着明显的依赖关系。



中文分词

- 输入观察序列X：南京市长江大桥
- 状态集合$Y(i)$ ={B, M, E, S}：

状态Y	Begin	Middle	End	Single
解释	词的开始字	词的中间字	词的结束字	单字成词
示例	南京的“南”	乒乓球的“乓”	南京的“京”	你

- 输出状态序列Y：BMEBMME

Viterbi 搜索算法

给定模型μ和观察序列$X=x_1x_2\ldots x_t\ldots x_T$的条件下求概率最大的状态序列$Y=y_1y_2\ldots y_t\ldots y_T$：

$$Y^* = \arg\max_{Y} P(Y \mid X, \mu)$$

Viterbi 算法: 动态搜索最优状态序列。

定义：Viterbi 变量$\delta_t(y(i))$**是在时间** t **时，模型沿着某一条路径到达状态** $y(i)$**，并输出观察序列** $X=x_1x_2\ldots x_t$**的最大概率：**

$$\delta_t(y(i)) = \max_{y_1y_2\cdots y_{t-1}} P(y_1y_2\cdots y_{t-1}y_t(i), x_1x_2\cdots x_{t-1}x_t \mid \mu)$$

从状态$y_t(j)$转移到状态$y_{t+1}(i)$的概率

t+1步状态$y_{t+1}(i)$发射观察值x_{t+1}的概率

$$\delta_{t+1}(y(i)) = \max_{y(j)}\{\delta_t(y(j)) \cdot P(y_{t+1}(i) \mid y_t(j))\} \cdot P(x_{t+1} \mid y_{t+1}(i))$$

where $y(i), y(j) \in \{B, M, E, S\}$

递归计算：

算法描述

(1)初始化：$\delta_1(i) = y_1(i)P(x_1 \mid y_1(i)), \quad 1 \le i \le M$

概率最大的**路径**变量：$\psi_1(i) = 0$

(2)递推计算：

$$\delta_t(i) = \max_{1 \le j \le M}\{\delta_{t-1}(j) \cdot P(y_t(i) \mid y_{t-1}(j))\} \cdot P(x_t \mid y_t(i)), 2 \le t \le T, 1 \le j, i \le M$$

第t步从所有可能的$j \rightarrow i$的M条路径中取最大值

$$\psi_t(i) = \arg\max_{1 \le j \le M}[\delta_{t-1}(j) \cdot P(y_t(i) \mid y_{t-1}(j))] \cdot P(x_t \mid y_t(i)), \quad 2 \le t \le T, \quad 1 \le i, j \le M$$

arg是从1...t路径的累积概率最大

(3)结束：

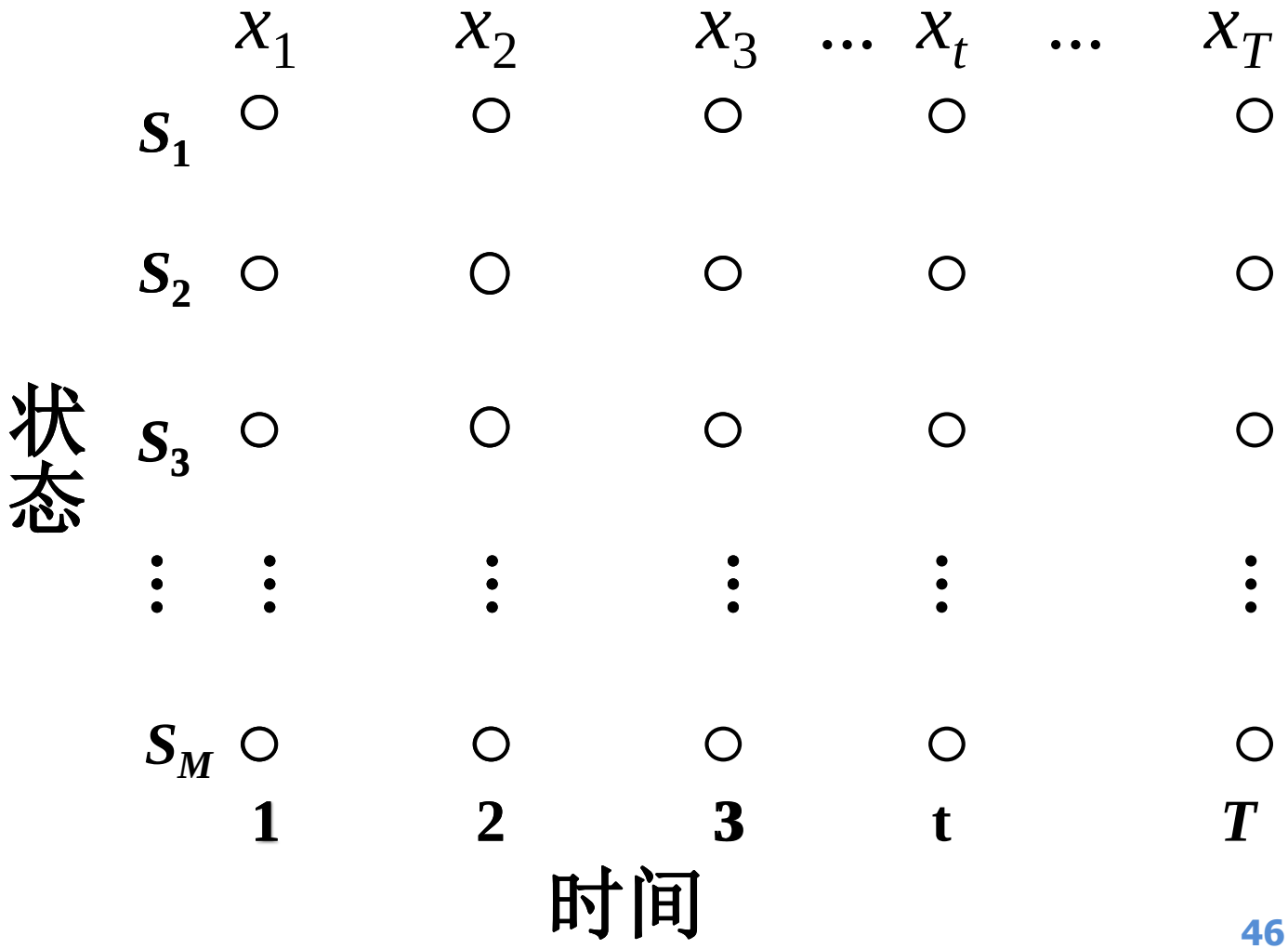
$$Y = \arg\max_{1\le i\le M}[\delta_T(i)], \qquad P(Y) = \max_{1\le i\le M}\delta_T(i)$$

(4)通过回溯得到路径（状态序列）：

$$y_t = \psi_{t+1}(y_{t+1}), \quad t = T-1,\ T-2,\ \cdots,\ 1$$

算法的时间复杂度：$O(M^2T)$

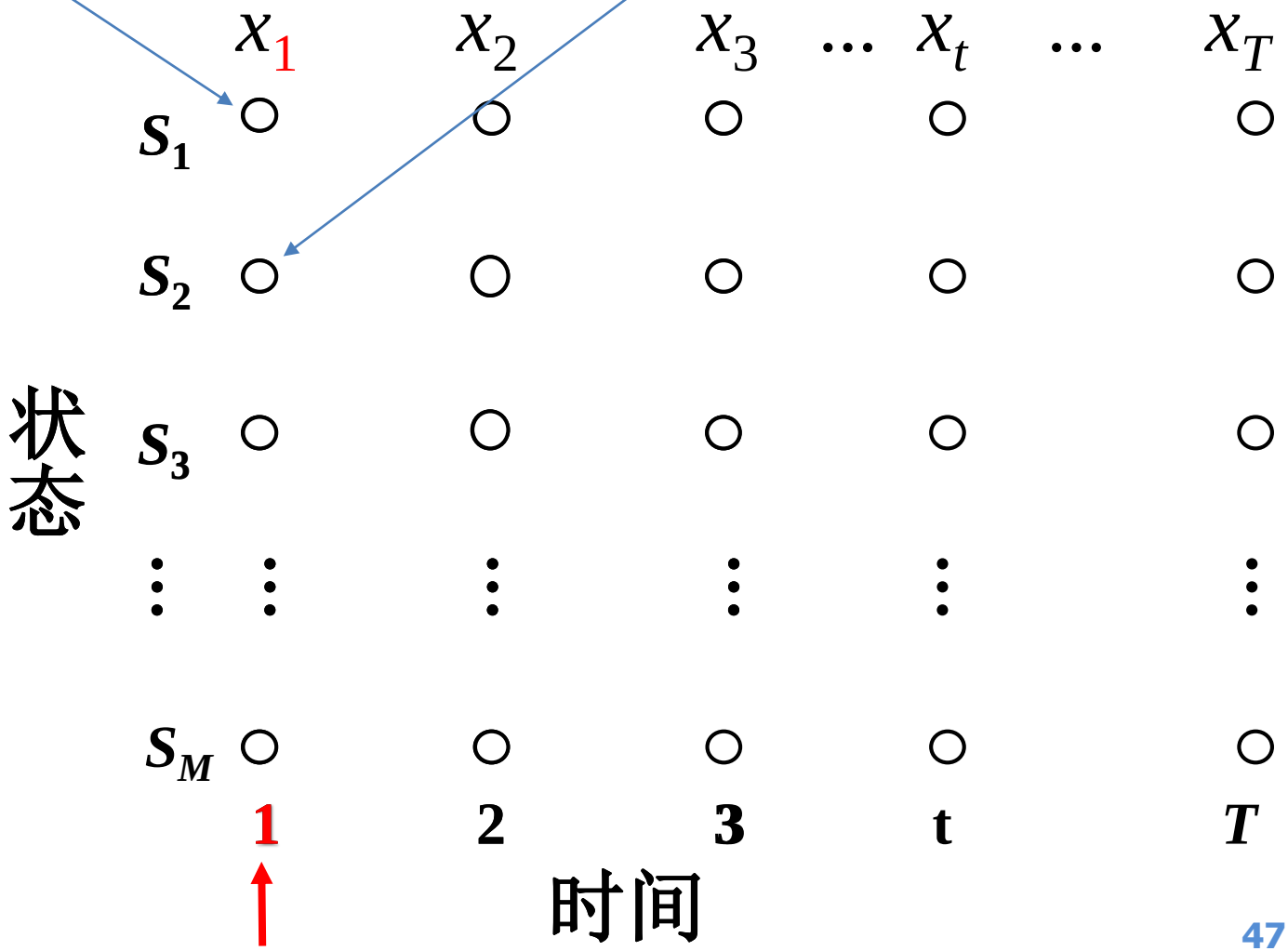
图解Viterbi搜索过程



$$\delta_1(i) = y_1(i)P(x_1 \mid y_1(i)), \quad 1 \le i \le M$$

$$\delta_1(1) = y_1(1)P(x_1 \mid y_1(1)) \qquad \delta_1(2) = y_1(2)P(x_1 \mid y_1(2))$$

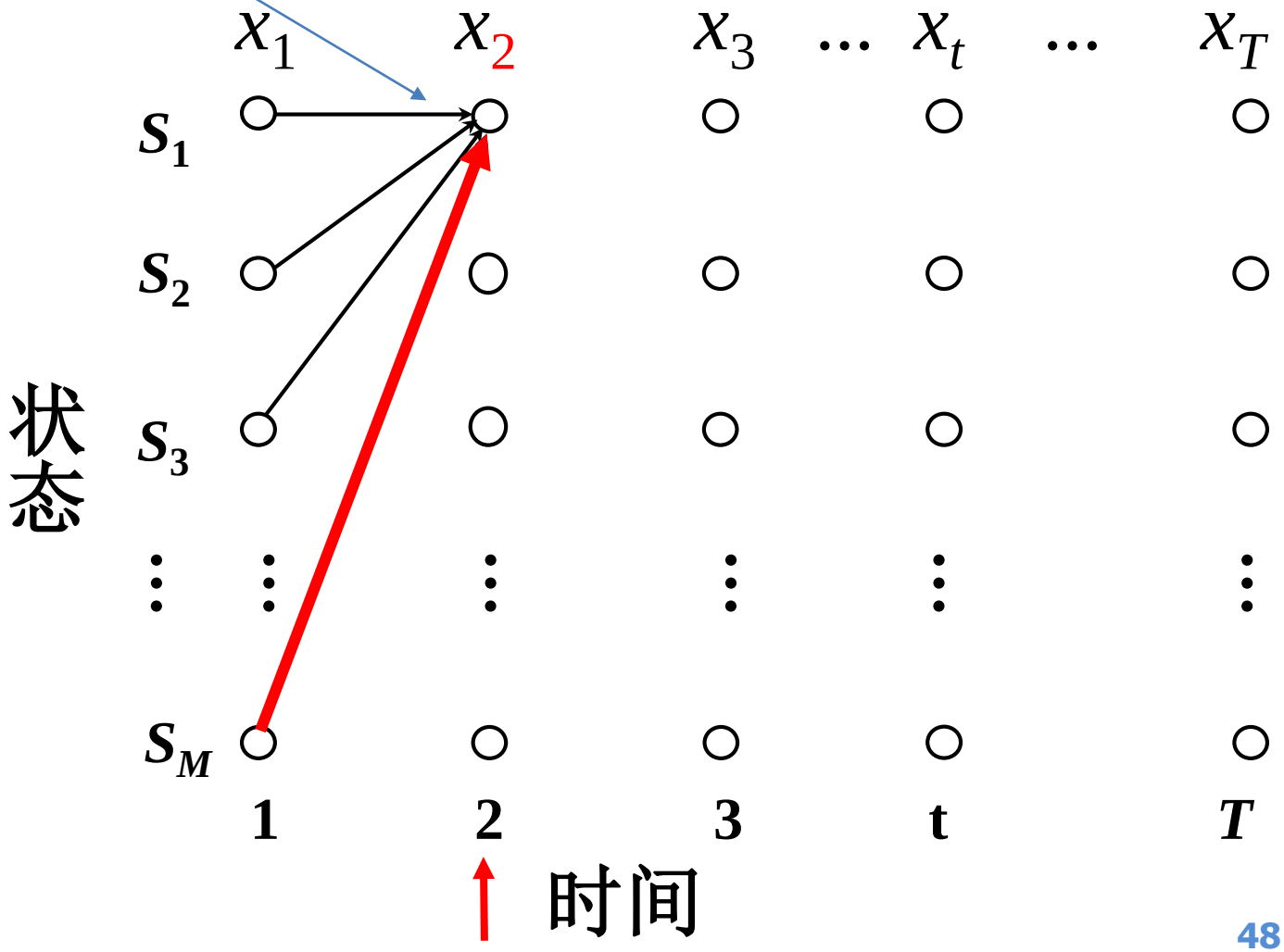
图解 Viterbi 搜索过程



$$\delta_t(i) = \max_{1\le j\le M}\{\delta_{t-1}(j)\cdot P(y_t(i)\mid y_{t-1}(j))\}\cdot P(x_t\mid y_t(i)),\ 2\le t\le T,\ 1\le j,i\le M$$

$$\delta_2(1) = \max_{1\le j\le M}\{\delta_1(j)\cdot P(y_2(1)\mid y_1(j))\}\cdot P(x_2\mid y_2(1))$$

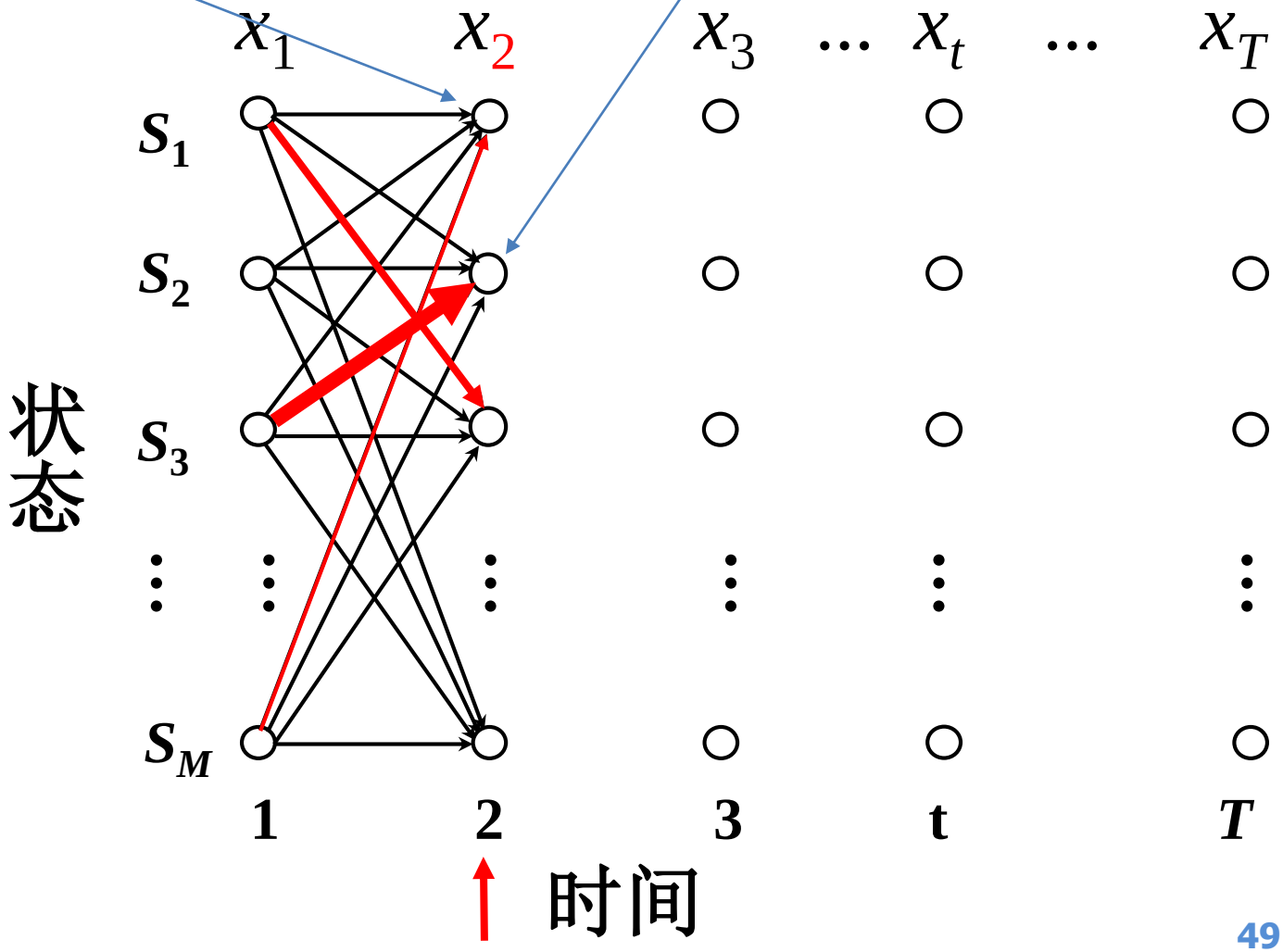
图解 Viterbi 搜索过程



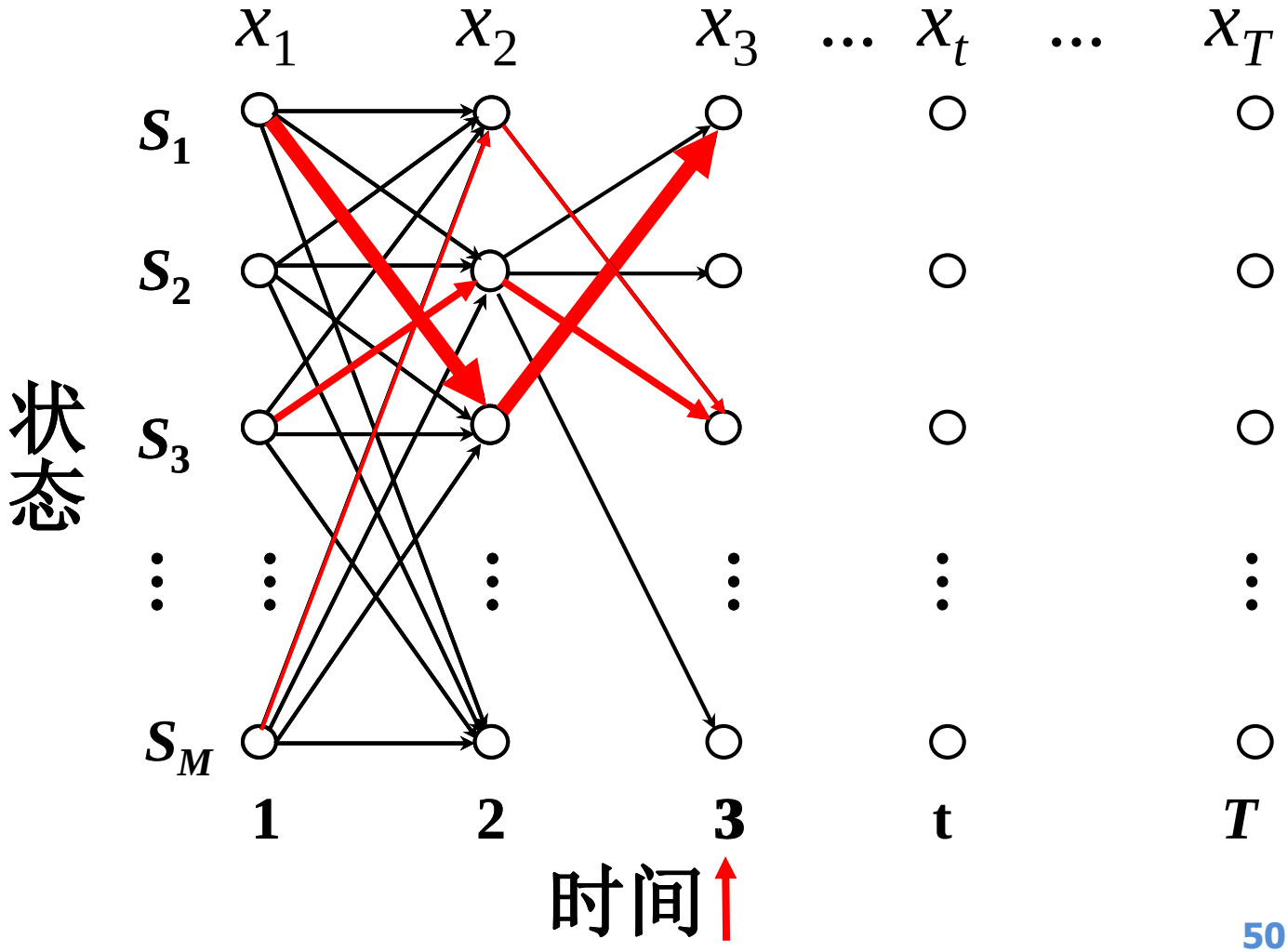
$$\delta_2(1) = \max_{1 \le j \le M} \{\delta_1(j) \cdot P(y_2(1) \mid y_1(j))\} \cdot P(x_2 \mid y_2(1))$$

$$\delta_2(2) = \max_{1 \le j \le M} \{\delta_1(j) \cdot P(y_2(2) \mid y_1(j))\} \cdot P(x_2 \mid y_2(2))$$

图解 Viterbi 搜索过程



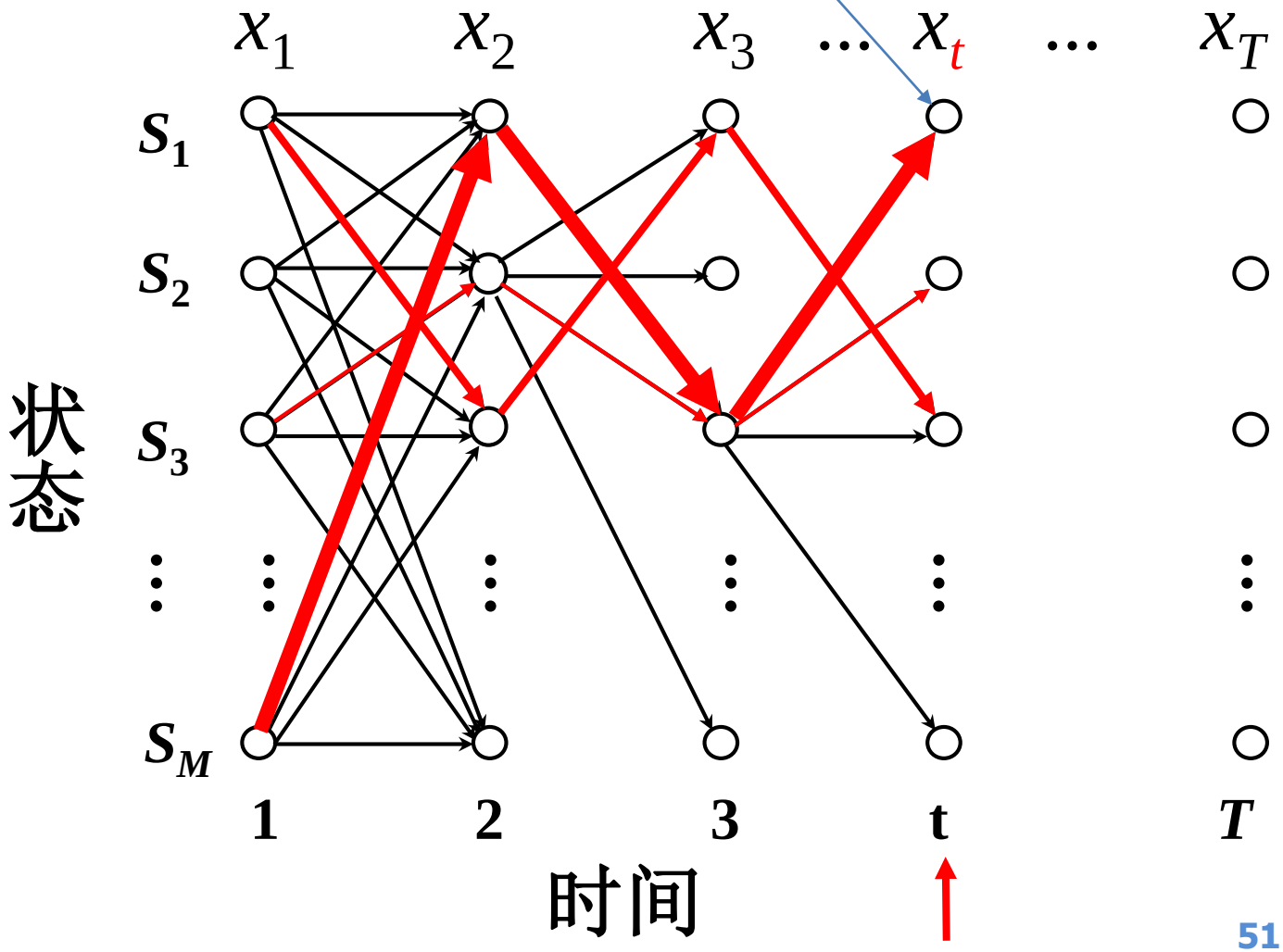
图解Viterbi搜索过程



$$\delta_t(i) = \max_{1\le j\le M}\{\delta_{t-1}(j)\cdot P(y_t(i)\mid y_{t-1}(j))\}\cdot P(x_t\mid y_t(i)),\ 2\le t\le T,\ 1\le j,i\le M$$

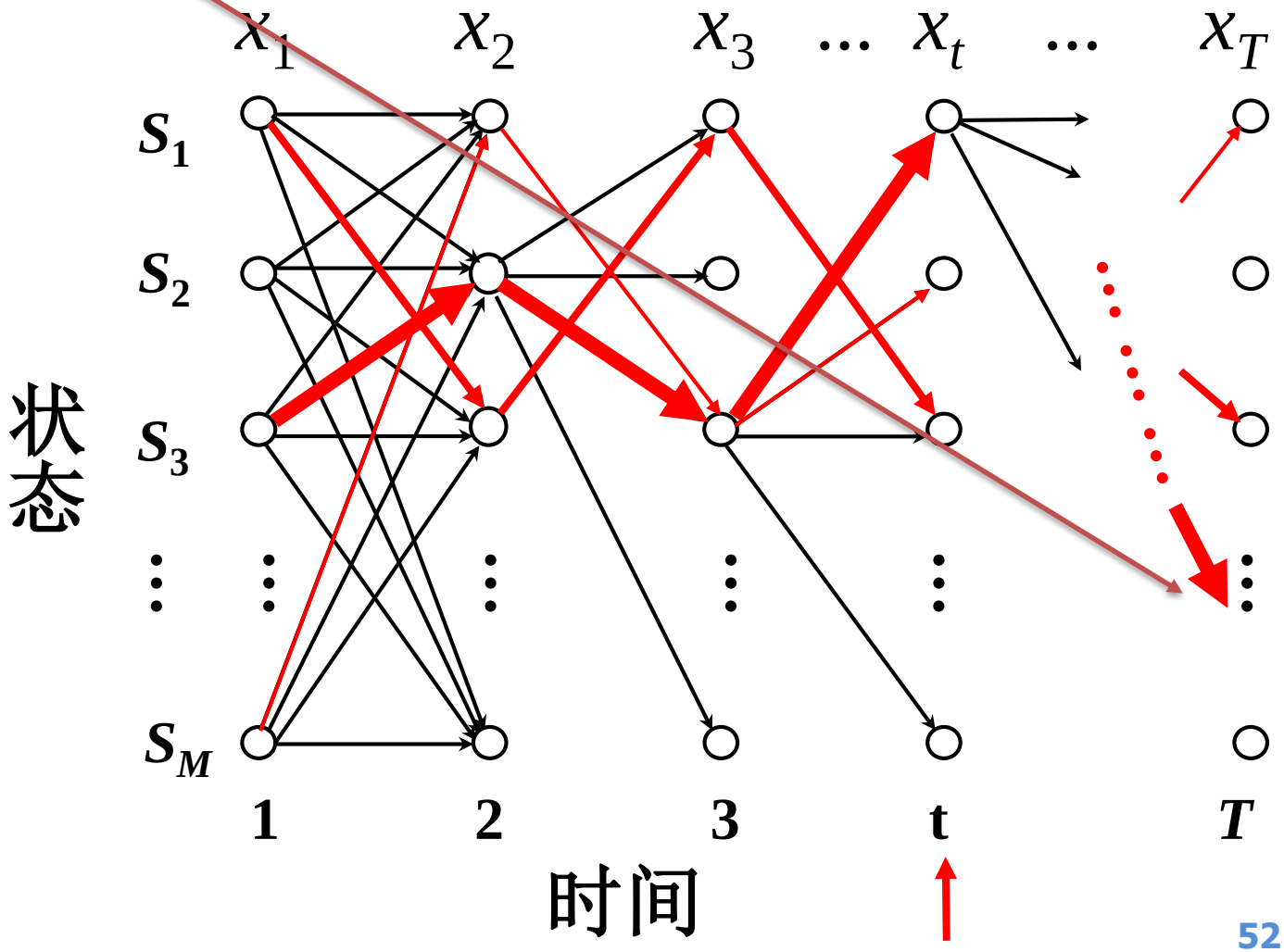
$$\delta_t(1) = \max_{1\le j\le M}\{\delta_{t-1}(j)\cdot P(y_t(1)\mid y_{t-1}(j))\}\cdot P(x_t\mid y_t(1))$$

图解 Viterbi 搜索过程



找到最大值 $\max_{1\le i\le M}\delta_T(i)$，然后通过回溯得到路径 $Y=\arg\max_{1\le i\le M}[\delta_T(i)]$

图解Viterbi搜索过程



- 输入观察序列X：南京市长江大桥
- 状态集合：

状态Y	Begin	Middle	End	Single
解释	词的开始字	词的中间字	词的结束字	单字成词
示例	南京的“南”	乒乓球的“乓”	南京的“京”	你

- 输出状态序列Y
- 给定HMM模型：{A,B, π}

字	南	京	市	长	江	大	桥
状态Begin	0.3	0.3	0.1	0.2	0.1	0.1	0.2
状态Middle	0.2	0.3	0.2	0.3	0.3	0.2	0.1
状态End	0.1	0.1	0.3	0.2	0.1	0.2	0.6
状态Single	0.2	0.2	0.2	0.1	0.4	0.2	0.1

南B 京M 市E 长B 江M 大M 桥E ⟹ 南京市， 长江大桥

词性标注

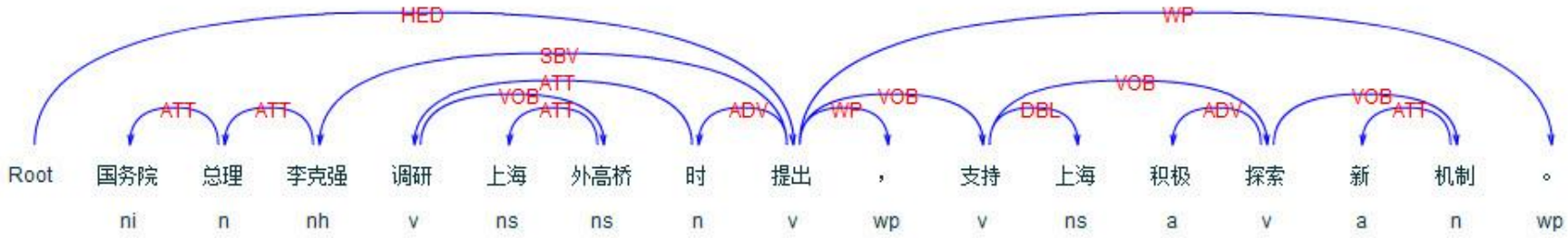
- Part-of-Speech tagging，在给定句子中判定每个词的语法范畴，确定其词性并加以标注的过程，如名词、动词、形容词等。

小明	高兴地	吃着	苹果
名词	形容词	动词	名词

  - 实现算法
    - 基于规则的词性标注方法
      - 按兼类词搭配关系和上下文语境建造词类消歧规则
    - 基于统计模型的词性标注方法
      - 基于HMM、CRF等的词性标注方法
    - 基于规则和统计相结合的词性标注方法

句法分析

- 依存句法分析 (Dependency Parsing, DP)
  - 通过分析语言单位内成分之间的依存关系揭示其句法结构。
  - 识别句子中的“主谓宾”、“定状补” 这些语法成分，并分析各成分之间的关系。



机器翻译

基本翻译方法

- 直接转换法
- 基于规则的翻译方法
- 基于中间语言的翻译方法
- 基于语料库的翻译方法
  - 统计翻译方法
  - 神经网络机器翻译

直接转换法

从源语言句子的表层出发，将单词、短语或句子直接置换成目标语言译文，必要时进行简单的词序调整。对原文句子的分析仅满足于特定译文生成的需要。这类翻译系统一般针对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。例如：

I like Mary. → Me(I) gusta(like) Maria(Mary).

X like Y → X gusta Y

基于规则的翻译方法(Rule-based)

1957年美国学者V. Yingve在《句法翻译框架》(Framework for Syntactic Translation)一文中提出了对源语言和目标语言均进行适当描述、把翻译机制与语法分开、用规则描述语法的实现思想，这就是基于规则的翻译方法。

基于规则的翻译过程分成6个步骤：

(*a*) 对源语言句子进行词法分析

(*b*) 对源语言句子进行句法/语义分析

(*c*) 源语言句子结构到译文结构的转换

(*d*) 译文句法结构生成

(*e*) 源语言词汇到译文词汇的转换

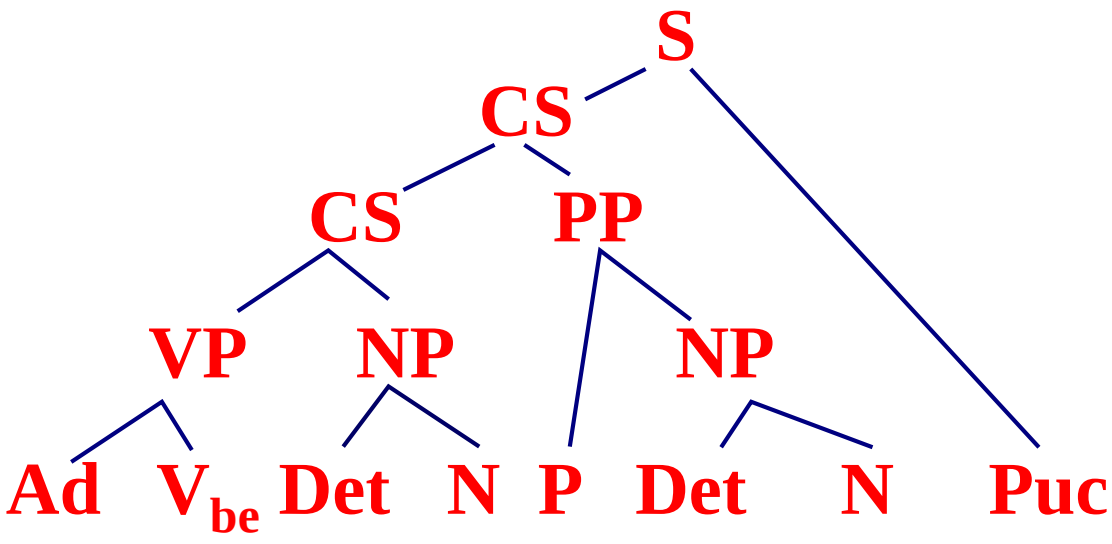
(*f*) 译文词法选择与生成

给定源语言句子：**There is a book on the desk.**

■**词法分析：**

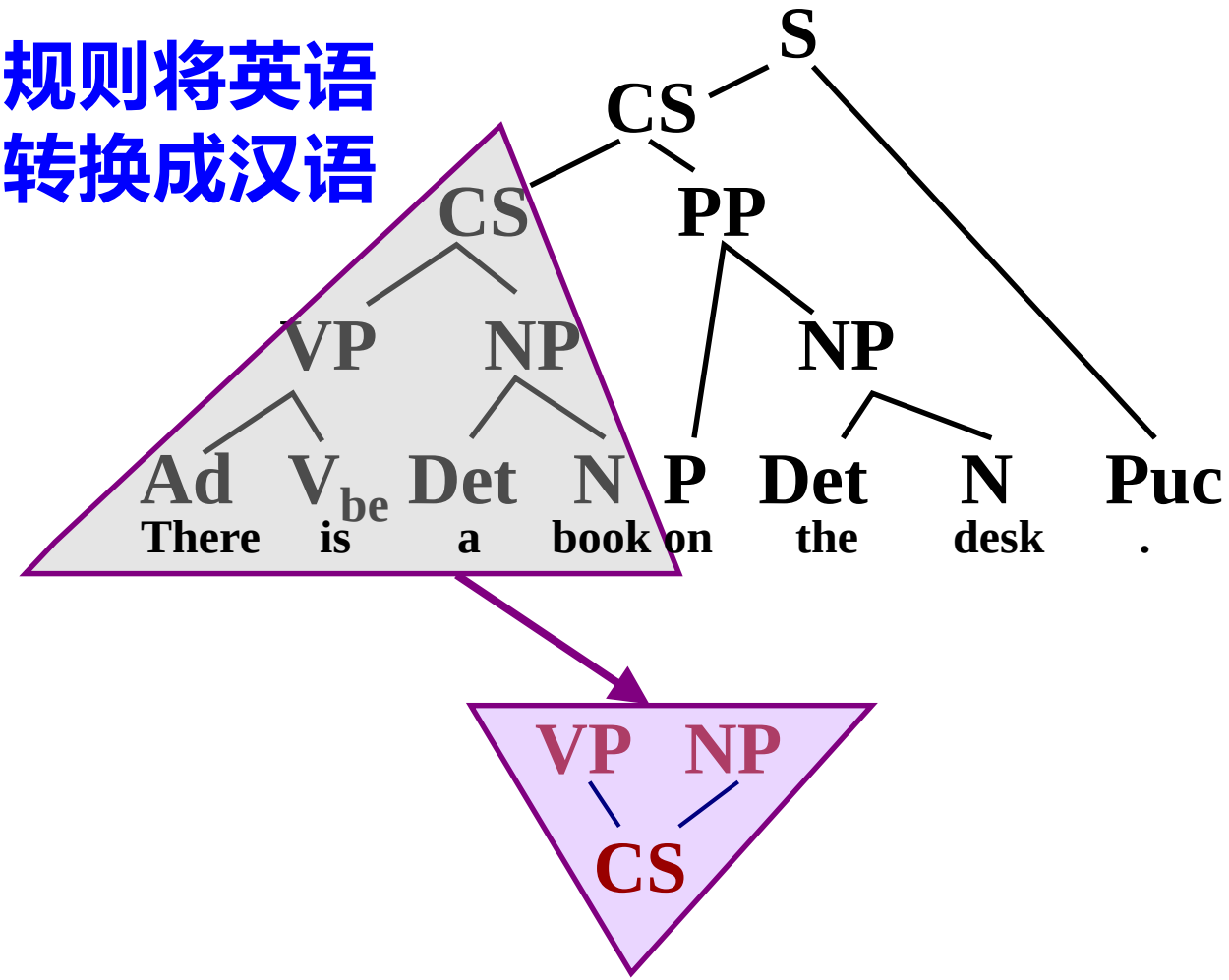
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

■**利用句法规则进行句法结构分析：**

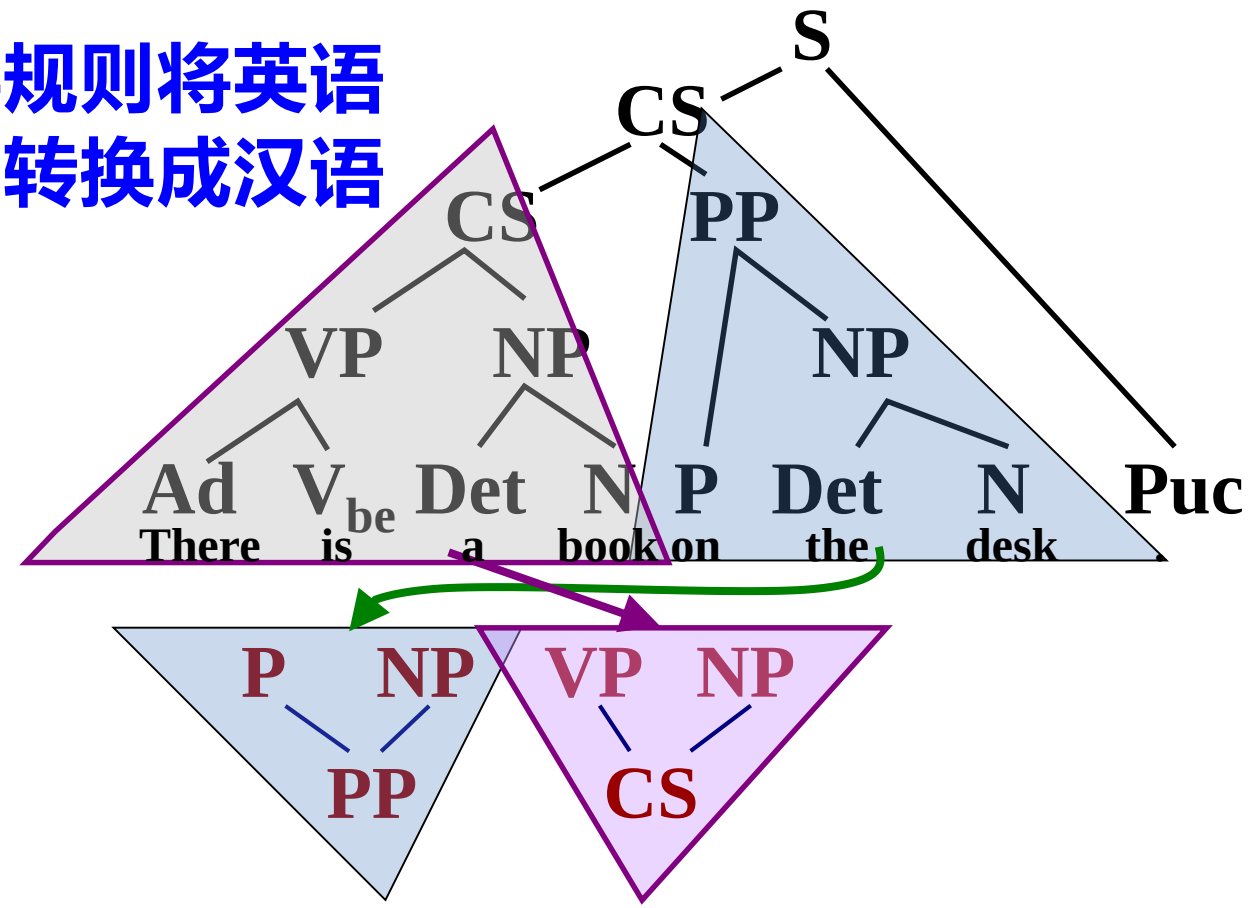


动词短语（verb phrase, VP）
名词短语（noun phrase, NP）
介词短语（preposition, PP）
2类连词（conjunction, 分别记作： CC, CS）

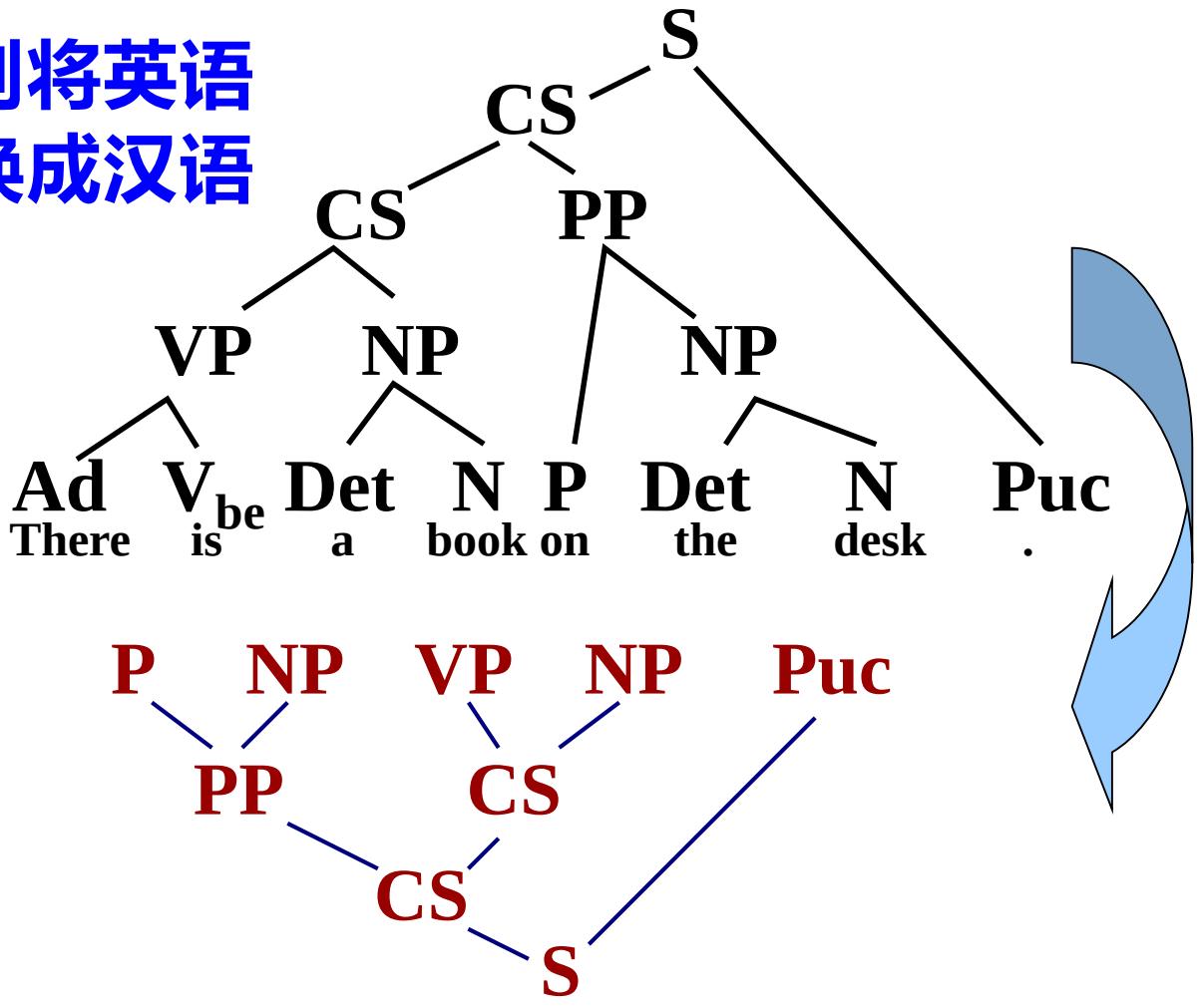
- **利用转换规则将英语句子结构转换成汉语句子结构**



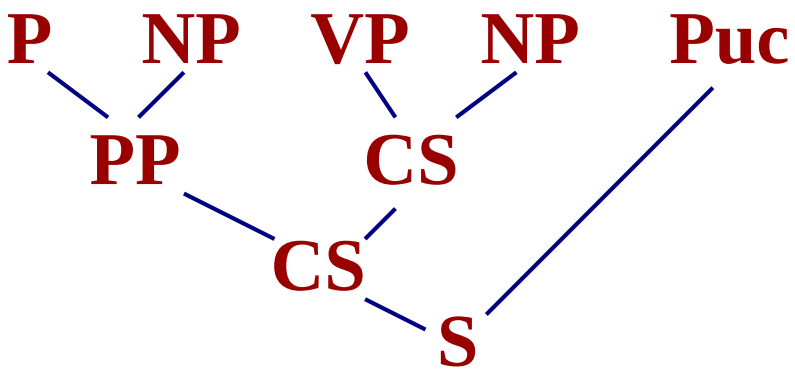
- **利用转换规则将英语句子结构转换成汉语句子结构**



- 利用转换规则将英语句子结构转换成汉语句子结构



- **根据转换后的句子结构，利用词典和生成规则生成翻译的结果句子**



#a, Det, 一

#book, N, 书; V, 预订

#desk, N, 桌子

#on, P, 在 X 上

#There be, V, 有

输出译文：

在桌子上有一本书。

由于基于规则的翻译方法执行过程为：

“独立分析 - 独立生成 - 相关转换”

因此，又称基于转换的翻译方法。

其代表系统是法国格勒诺布尔(Grenoble)机器翻译研究所(GETA)开发的ARIANE翻译系统。

1976年加拿大蒙特利尔大学与加拿大联邦翻译局联合开发的实用性机器翻译系统 TAU-METEO：天气预报信息服务。

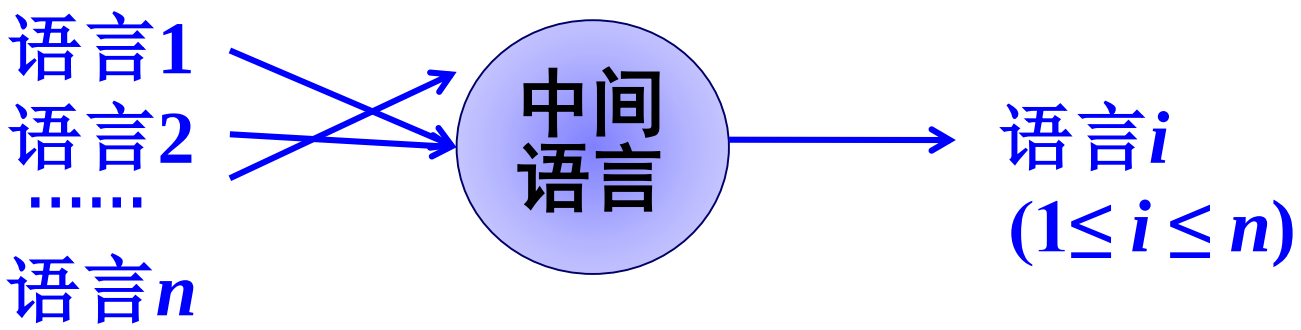
对基于规则的翻译方法的评价：

优点：可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知的或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效果。

弱点：规则一般由人工编写，工作量大，主观性强，一致性难以保障，不利于系统扩充，对非规范语言现象缺乏相应的处理能力。

基于中间语言的翻译方法(Interlingua-based)

- 方法：输入语句→中间语言→ 翻译结果
- 代表系统：JANUS（CMU) 早期版本
  - ✯ 源语言解析器
  - ✯ 比较准确的中间语言(Interlingua)
  - ✯ 目标语言生成器(Target Language Generator)



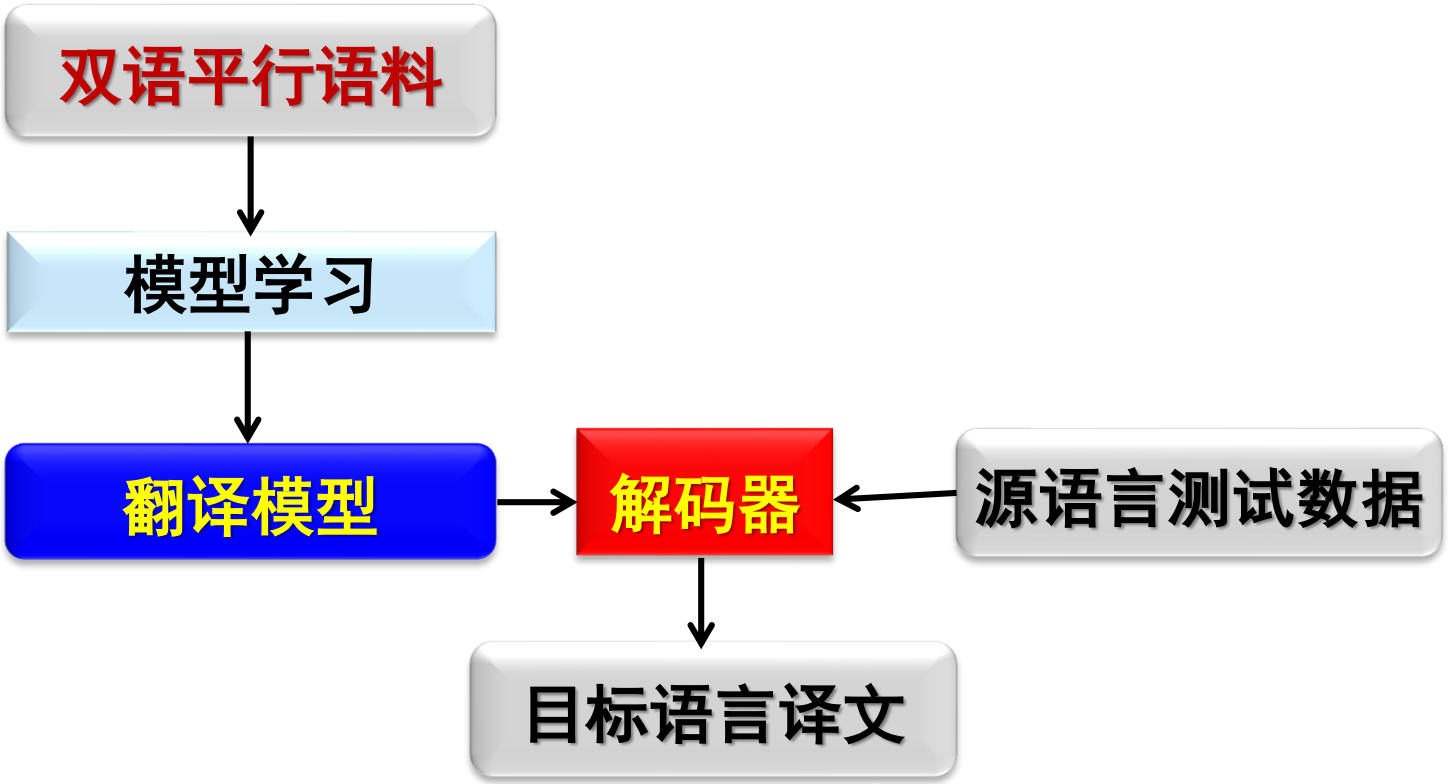
对基于中间语言的翻译方法评价：

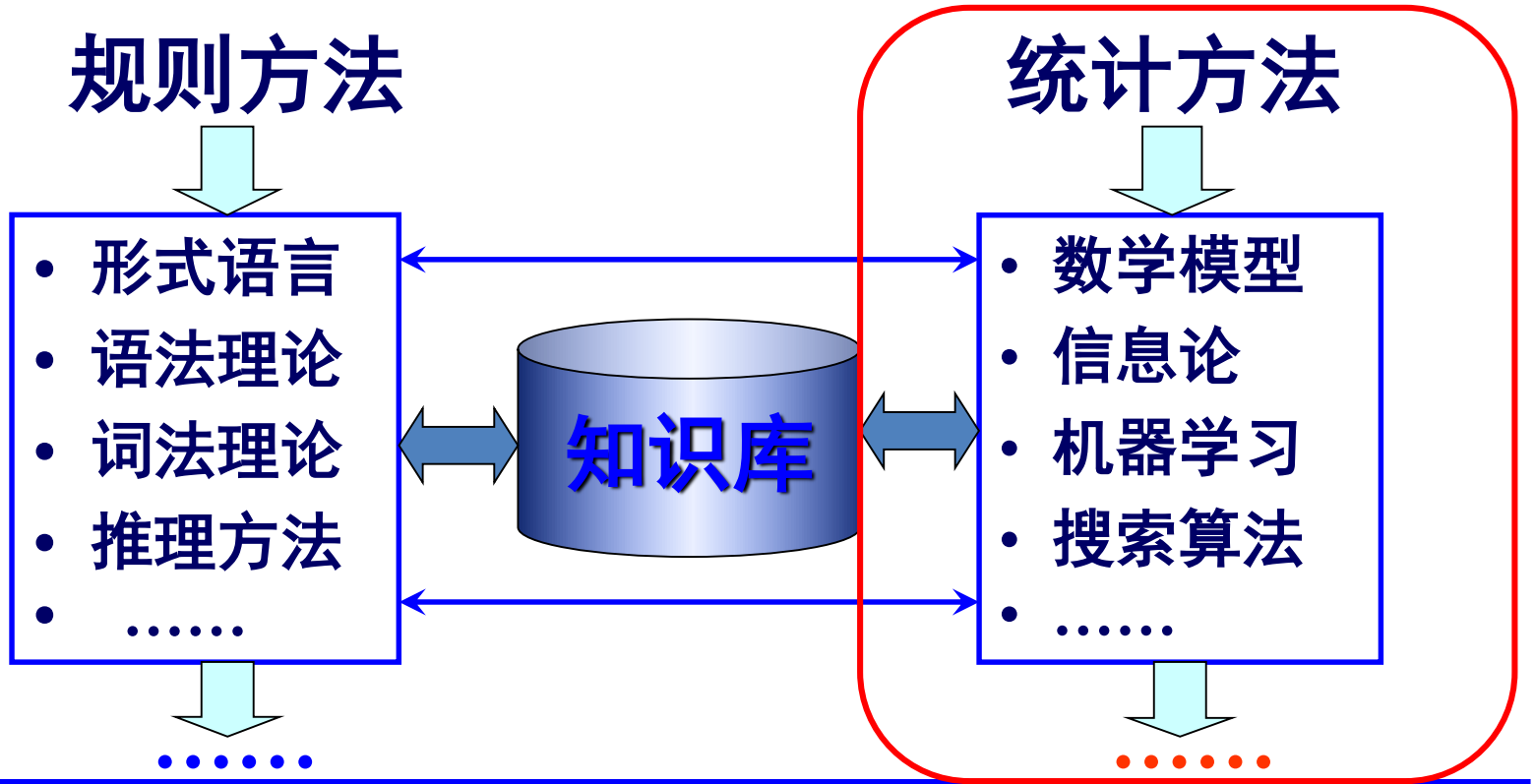
优点：中间语言的设计可以不考虑具体的翻译语言对，因此，该方法尤其适合多语言之间的互译。

弱点：如何定义和设计中间语言的表达方式，以及如何维护并不是一件容易的事情，中间语言在语义表达的准确性、完整性等很多方面，都面临若干困难。

基于语料库的翻译方法

- **数据驱动的翻译方法**（如SMT和 NMT）





理性主义与经验主义的合谋 —
符号智能 + 计算智能，建立融合方法

Garcia and associates .

Garcia y asociados .

Carlos Garcia has three associates .

Carlos Garcia tiene tres asociados .

his associates are not strong .

sus asociados no son fuertes .

Garcia has a company also .

Garcia tambien tiene una empresa .

its clients are angry .

sus clientes estan enfadados .

the associates are also angry .

los asociados tambien estan enfadados .

the clients and the associates are enemies .

los clientes y los asociados son enemigos .

the company has three groups .

la empresa tiene tres grupos .

its groups are in Europe .

sus grupos estan en Europa .

the modern groups sell strong pharmaceuticals .

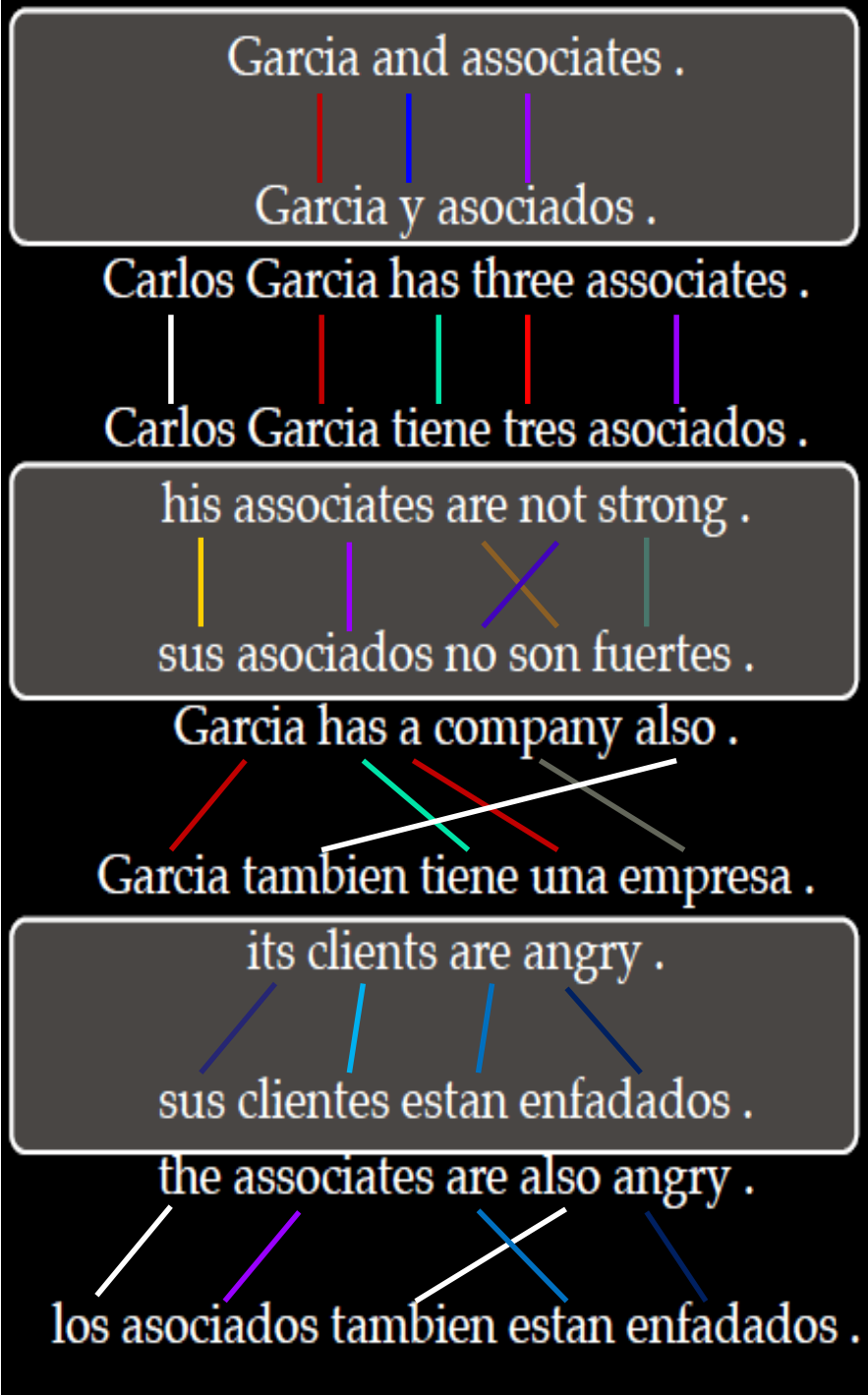
los grupos modernos venden medicinas fuertes .

the groups do not sell zanzanine .

los grupos no venden zanzanina .

the small groups are not modern .

los grupos pequenos no son modernos .


Garcia and associates .
Garcia y asociados .
Carlos Garcia has three associates .
Carlos Garcia tiene tres asociados .
his associates are not strong .
sus asociados no son fuertes .
Garcia has a company also .
Garcia tambien tiene una empresa .
its clients are angry .
sus clientes estan enfadados .
the associates are also angry .
los asociados tambien estan enfadados .
the clients and the associates are enemies .
los clientes y los asociados son enemigos .
the company has three groups .
la empresa tiene tres grupos .
its groups are in Europe .
sus grupos estan en Europa .
the modern groups sell strong pharmaceuticals .
los grupos modernos venden medicinas fuertes .
the groups do not sell zanzanine .
los grupos no venden zanzanina .
the small groups are not modern .
los grupos pequenos no son modernos .

统计翻译的诞生

- 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”[Brown, 1990]；1993年他们在该杂志发表论文“统计机器翻译的数学：参数估计”[Brown, 1993]，两篇文章奠定了统计机器翻译的理论基础。

左：Robert Mercer

右：Peter F. Brown

统计翻译基本原理

◆噪声信道模型

一种语言T 由于经过一个噪声信道而发生变形，从而在信道的另一端呈现为另一种语言 S (信道意义上的输出，翻译意义上的源语言)。翻译问题实际上就是如何根据观察到的 S，恢复最为可能的T 问题。这种观点认为，任何一种语言的任何一个句子都有可能是另外一种语言中的某个句子的译文，只是可能有大有小[Brown *et. al*, 1990]。

➢基于统计的方法

给定源语言句子： $S = s_1^m \equiv s_1 s_2 \cdots s_m$

将其翻译成目标语言句子： $T = t_1^l \equiv t_1 t_2 \cdots t_l$

根据贝叶斯公式：

$$P(T|S) = \frac{P(T)P(S|T)}{P(S)}$$

求解使P值最大的T

$$\hat{T} = \arg\max_T P(T)P(S|T)$$

语言模型
(Language model, LM)

翻译模型
(Translation model, TM)

构建解码器(decoder)，快速搜索最优翻译候选：

◆三个关键问题：

- 估计语言模型概率 $P(T)$；
- 估计翻译模型概率 $P(S|T)$；
- 快速有效地搜索候选译文T，使$P(T)\times P(S|T)$最大。

◆主要任务：

- 收集大规模双语句子对、目标语言句子
- 参数训练与模型优化

估计语言模型概率 $P(T)$

给定句子： $T = t_1^l = t_1 t_2 \cdots t_l$

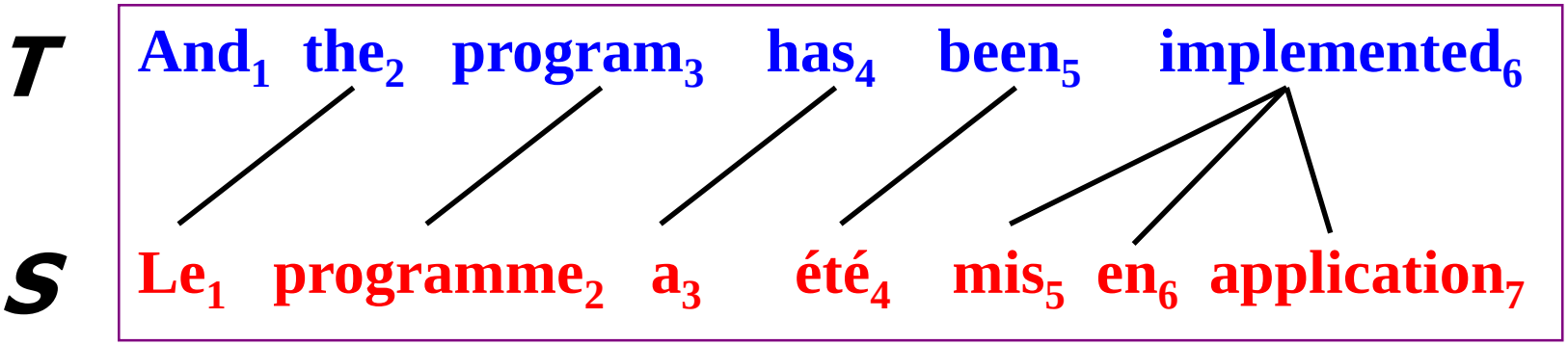
句子概率： $P(T) = P(t_1)P(t_2|t_1)\cdots P(t_l|t_1 t_2 \cdots t_{l-1})$

***n*-gram 问题**，不再赘述。

翻译概率 $p(S|T)$ 的计算

关键问题是怎样定义目标语言句子中的词与源语言句子中的词之间的对应关系。

假设英语与法语的翻译对：



统计翻译基本原理

不妨，我们用$\mathcal{A}(S,T)$ 表示源语言句子S 与目标语言句子T 之间所有对位关系的集合。在目标语言句子T 的长度（单词的个数）为 l，源语言句子 S 的长度为 m 的情况下，T 和 S 的单词之间有 $2^{l\times m}$ 种不同的对应关系。

$$|\mathcal{A}(S,T)| = 2^{l\times m} \qquad A(S,T) \in \mathcal{A}(S,T)$$

用来刻画这些对应关系$A(S,T)$的模型叫做**对位模型（alignment model）**。

统计翻译基本原理

将对位模型 A 视为隐含变量，则：

$$P(S|T) = \sum_{A} P(S, A|T)$$

按照约定，源语言句子 $S = s_1^m = s_1 s_2 \cdots s_m$ 有m个单词

目标语言句子 $T = t_1^l = t_1 t_2 \cdots t_l$ 有l个单词

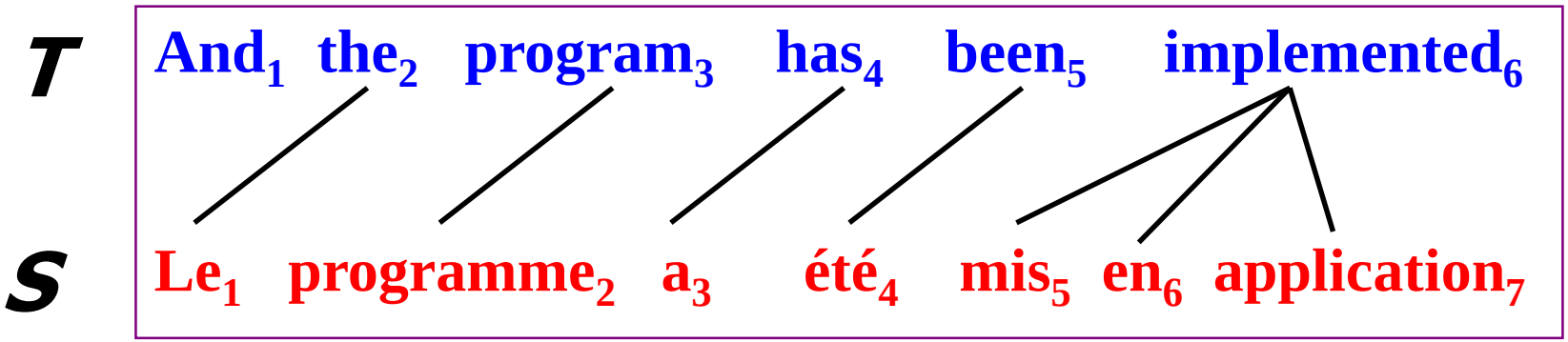
每一种对位序列表示成：

$$A = a_j^m = a_1 a_2 \cdots a_m \qquad a_j \in [0, 1, \cdots, l]$$

$a_j=i$表示从S的第j个词 到T的第i个词 的 对位关系

j从1到m，i从1到l

统计翻译基本原理



$$l = 6 \qquad m = 7$$

$$a_1 = 2, a_2 = 3, a_3 = 4, a_4 = 5, a_5 = 6, a_6 = 6, a_7 = 6$$

$$A = a_1^m = (2, 3, 4, 5, 6, 6, 6)$$

a_j=i表示从S的第j个词 到T的第i个词 的 对位关系

j从1到m，i从1到l

翻译概率 $P(S|T)$ 的计算

$$P(S|T) = \sum_{A} P(S, A|T)$$

$$\Rightarrow \quad P(S, A|T) \quad ?$$

$$P(S, A|T) = p(m|T) \times P(A|T, m) \times P(S|T, A, m)$$

对位模型

词汇翻译模型

统计翻译基本原理

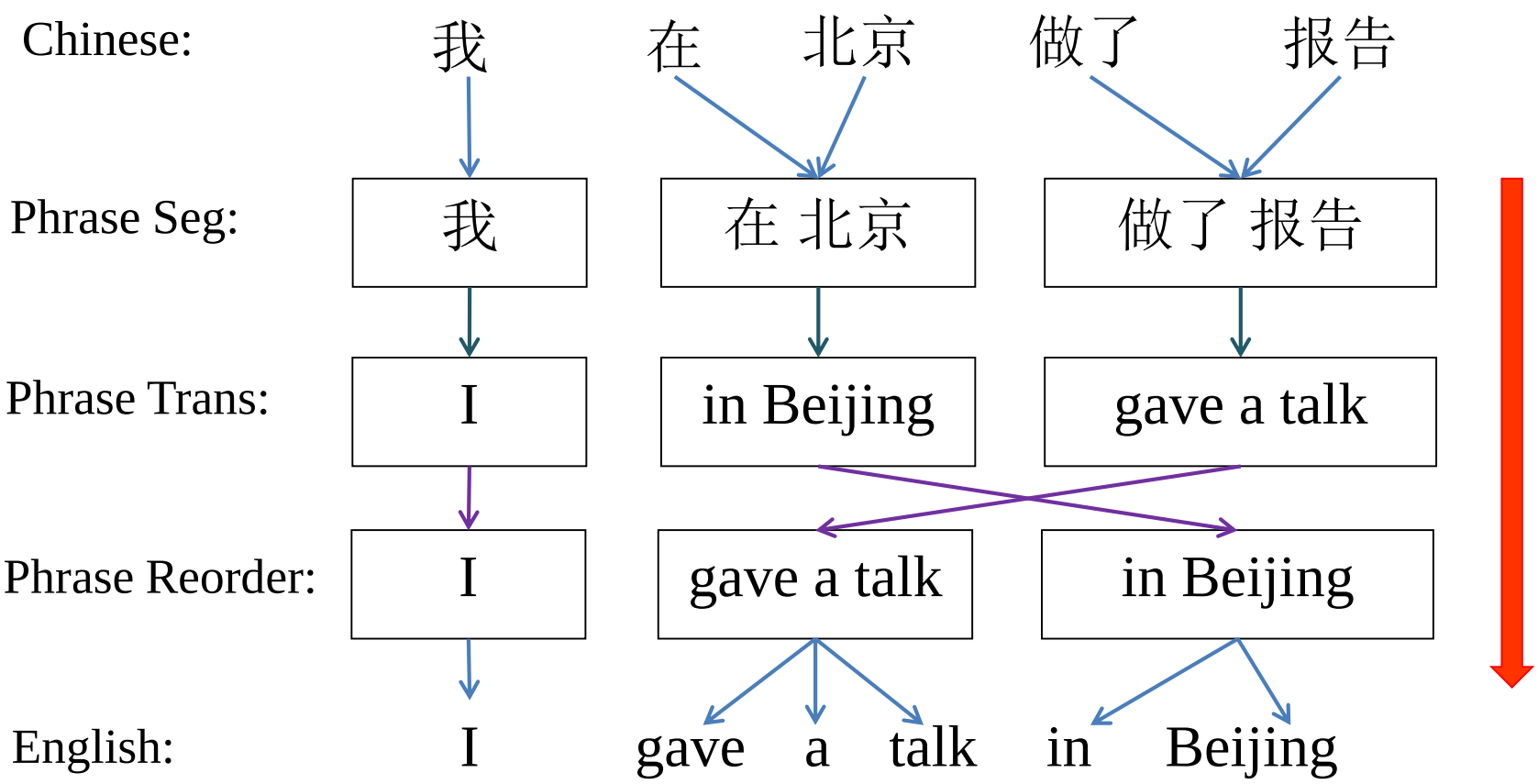
$$P(S,A|T) = p(m|T) \times P(A|T,m) \times P(S|T,A,m)$$

$$= p(m|T) \prod_{j=1}^{m} p\left(a_j|a_1^{j-1}, s_1^{j-1}, m, T\right) \times p\left(s_j|a_1^{j}, s_1^{j-1}, m, T\right)$$

实际上，$p(S, A|T)$ 可以写成多种形式的条件概率的乘积，上式只是其中的一种。在上式的基础上，IBM 的研究人员通过采用不同的假设条件得到了5个翻译模型，分别称作 IBM 翻译模型1、2、3、4 和 5。

神经机器翻译

统计机器翻译


Chinese:
我
在
北京
做了
报告
Phrase Seg:
我
在 北京
做了 报告
Phrase Trans:
I
in Beijing
gave a talk
Phrase Reorder:
I
gave a talk
in Beijing
English:
I
gave
a
talk
in
Beijing

统计机器翻译

①可解释性高

人工设定的模块和特征

②模块随便加

③错误易追踪

统计机器翻译

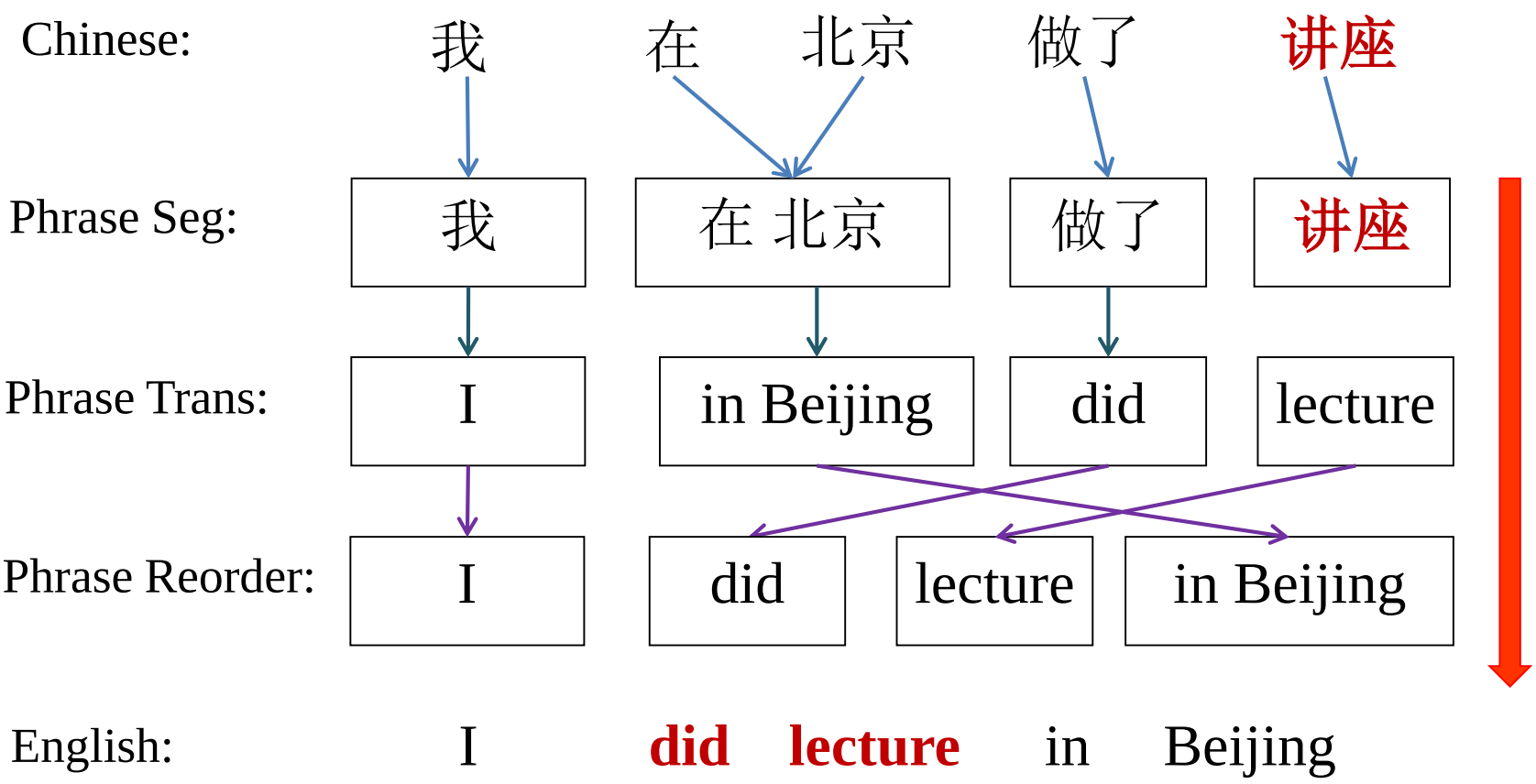
①数据稀疏

人工设定的模块和特征

②复杂结构
无能为力

③强烈依赖先
验知识

统计机器翻译



①数据稀疏

统计机器翻译

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.

②复杂结构无能为力

刘洋，CCL-2016

现实世界 VS. 认知世界

- 现实世界：物体相互独立地存在

现实世界 VS. 认知世界

- 认知世界：概念互相联系、语义连续分布

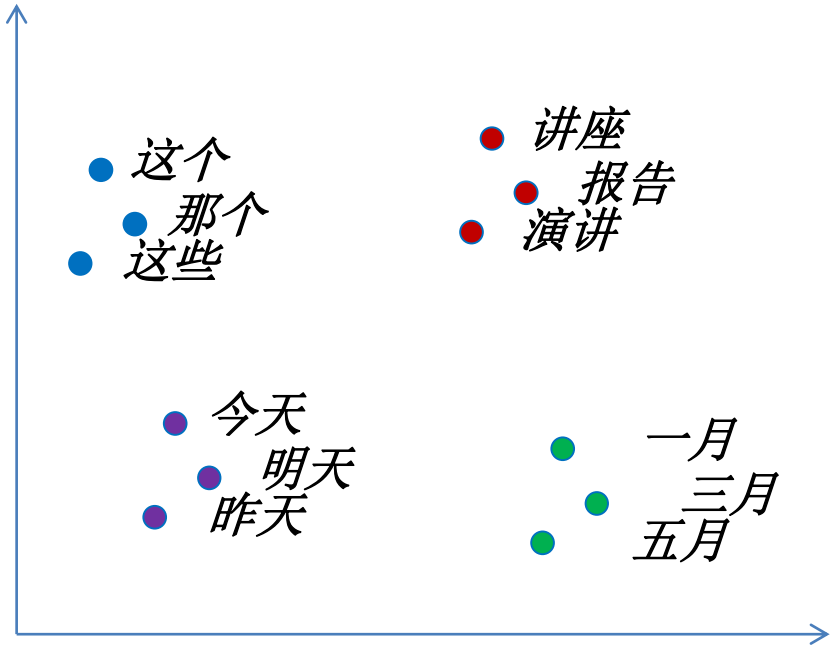
统计机器翻译→神经机器翻译

离散符号表示方法 ⇨ 连续分布式表示方法

讲座 ⊗ 报告 = 0

讲座 报告

$$\begin{bmatrix}0.48\\0.46\\0.26\end{bmatrix} \otimes \begin{bmatrix}0.42\\0.51\\0.21\end{bmatrix} \approx 1$$



低维、稠密的连续实数空间

分布式语义表示
是核心和基础

神经机器翻译

Chinese: 我 在 北京 做了 报告

编码网络

分布式语义表示

解码网络

English: I gave a talk in Beijing

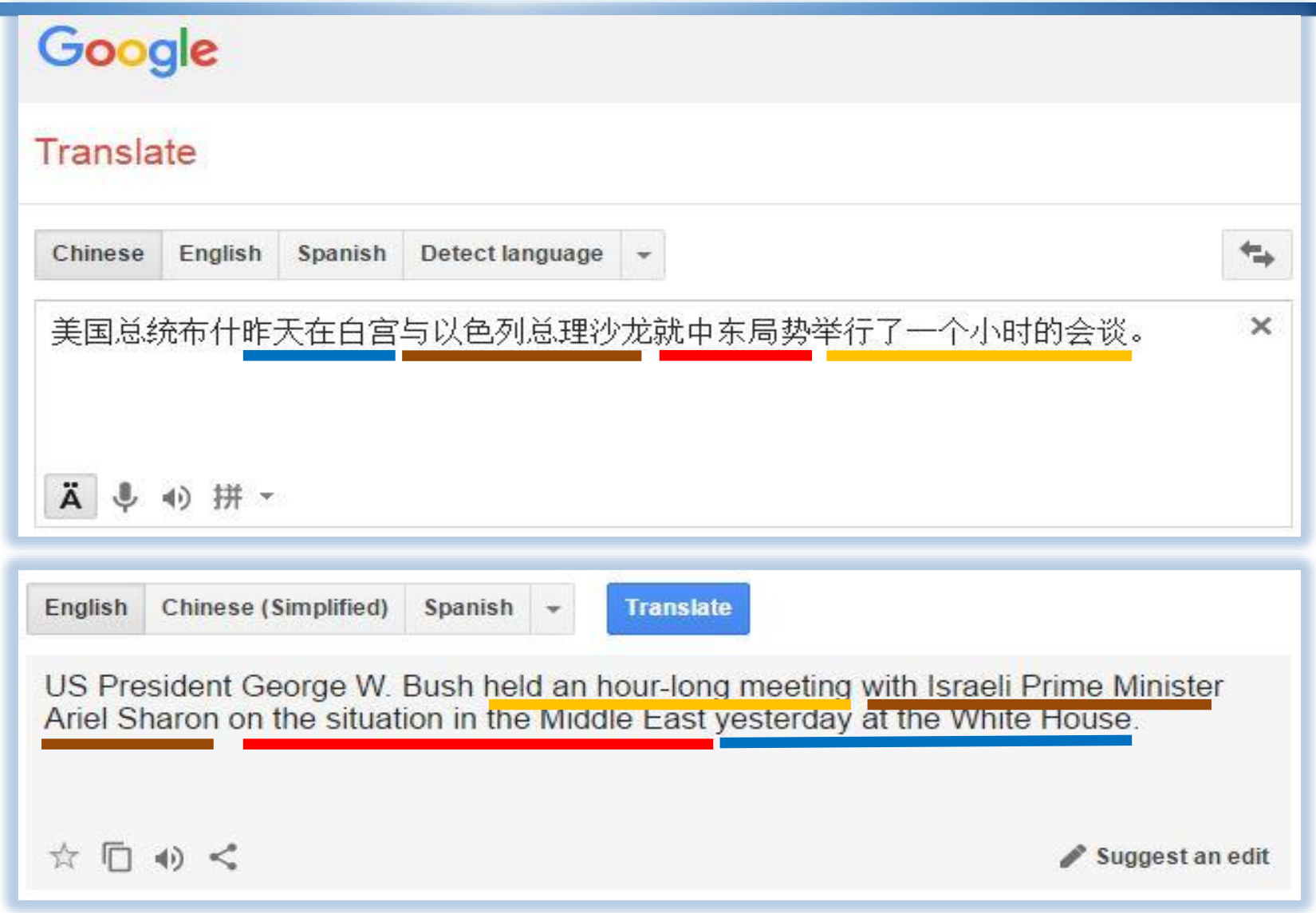
神经机器翻译

Chinese: 我 在 北京 做了 报告

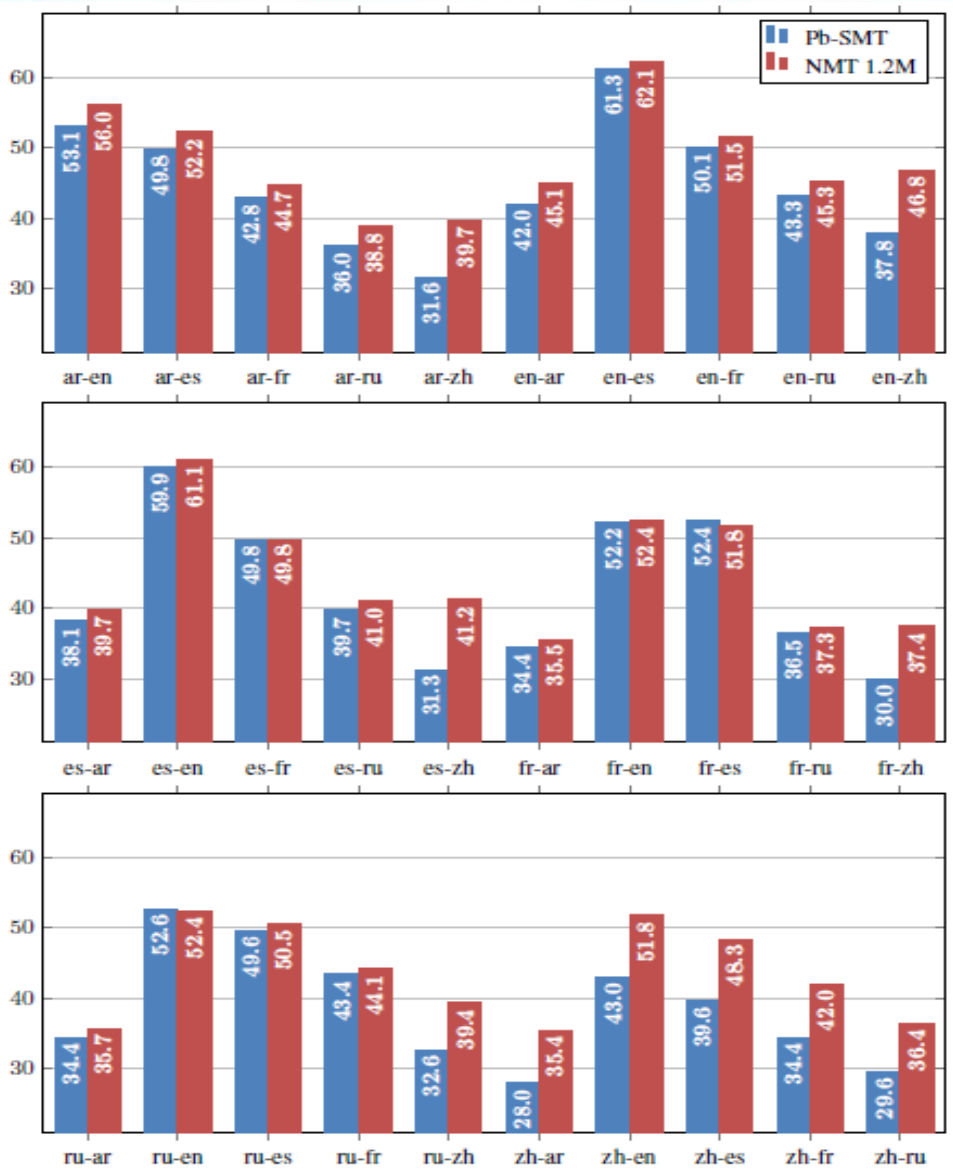
仅需要两个神经网络

English: I gave a talk in Beijing

神经机器翻译


Google
Translate
Chinese
English
Spanish
Detect language
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
Chinese (Simplified)
Spanish
Translate
US President George W. Bush held an hour-long meeting with Israeli Prime Minister Ariel Sharon on the situation in the Middle East yesterday at the White House.
Suggest an edit

神经机器翻译



神经机器翻译大获全胜！

[Junczys-Dowmunt et al, 2016]

统计机器翻译→神经机器翻译

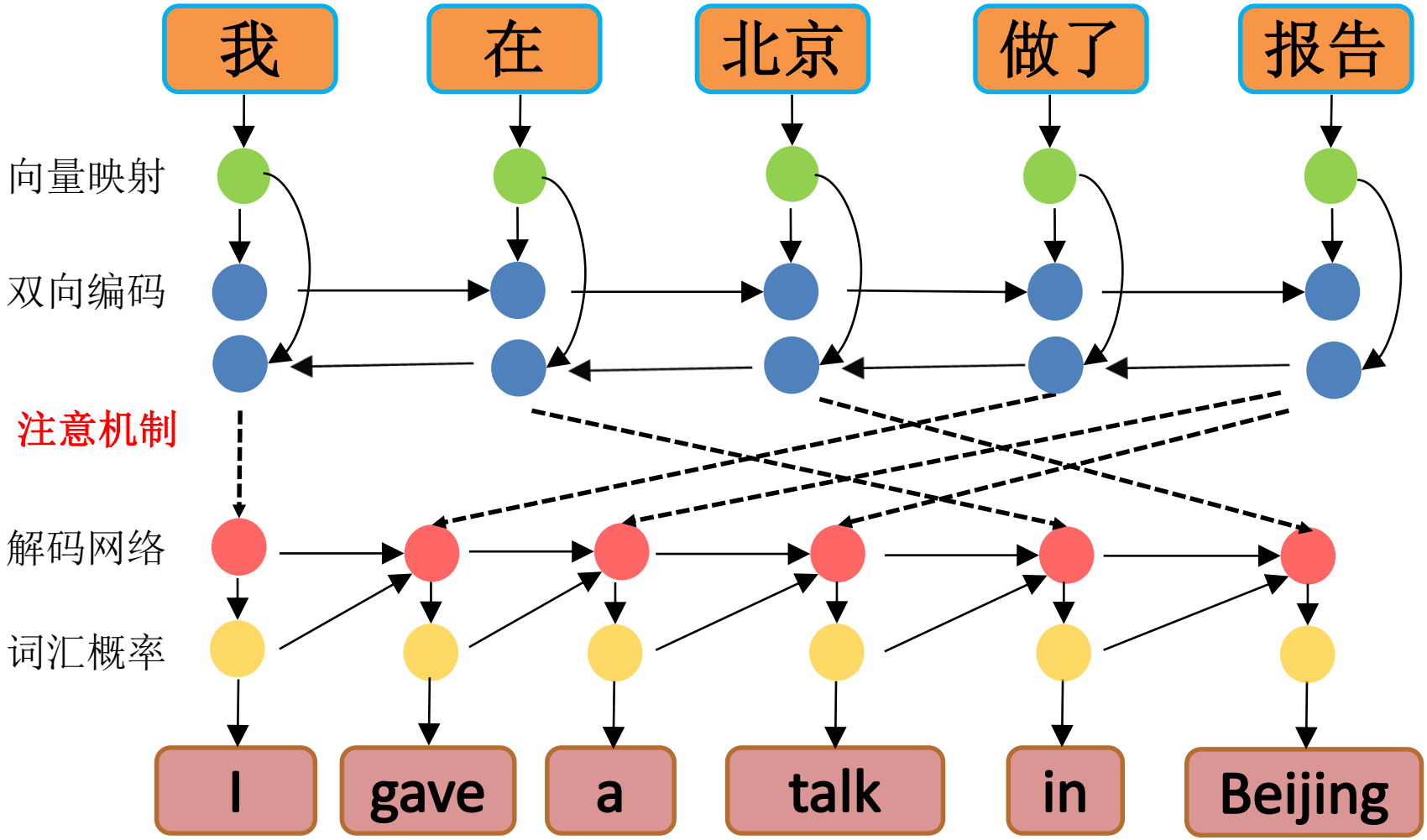
离散符号表示方法 ⇨ 连续分布式表示方法

讲座　　报告

$$\begin{bmatrix}0.48\\0.46\\0.26\end{bmatrix} \otimes \begin{bmatrix}0.42\\0.51\\0.21\end{bmatrix} \approx 1$$

表示是核心
运算是关键

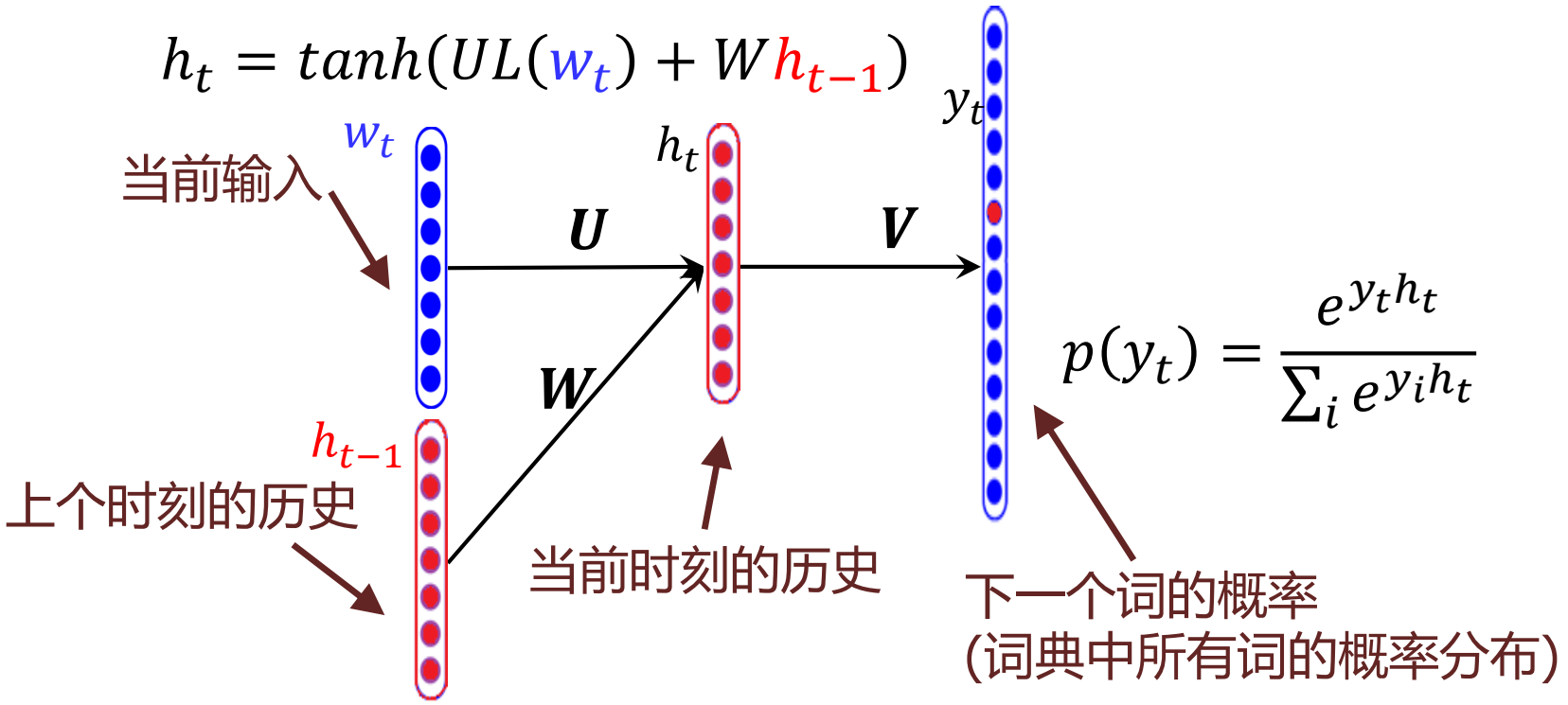
神经机器翻译



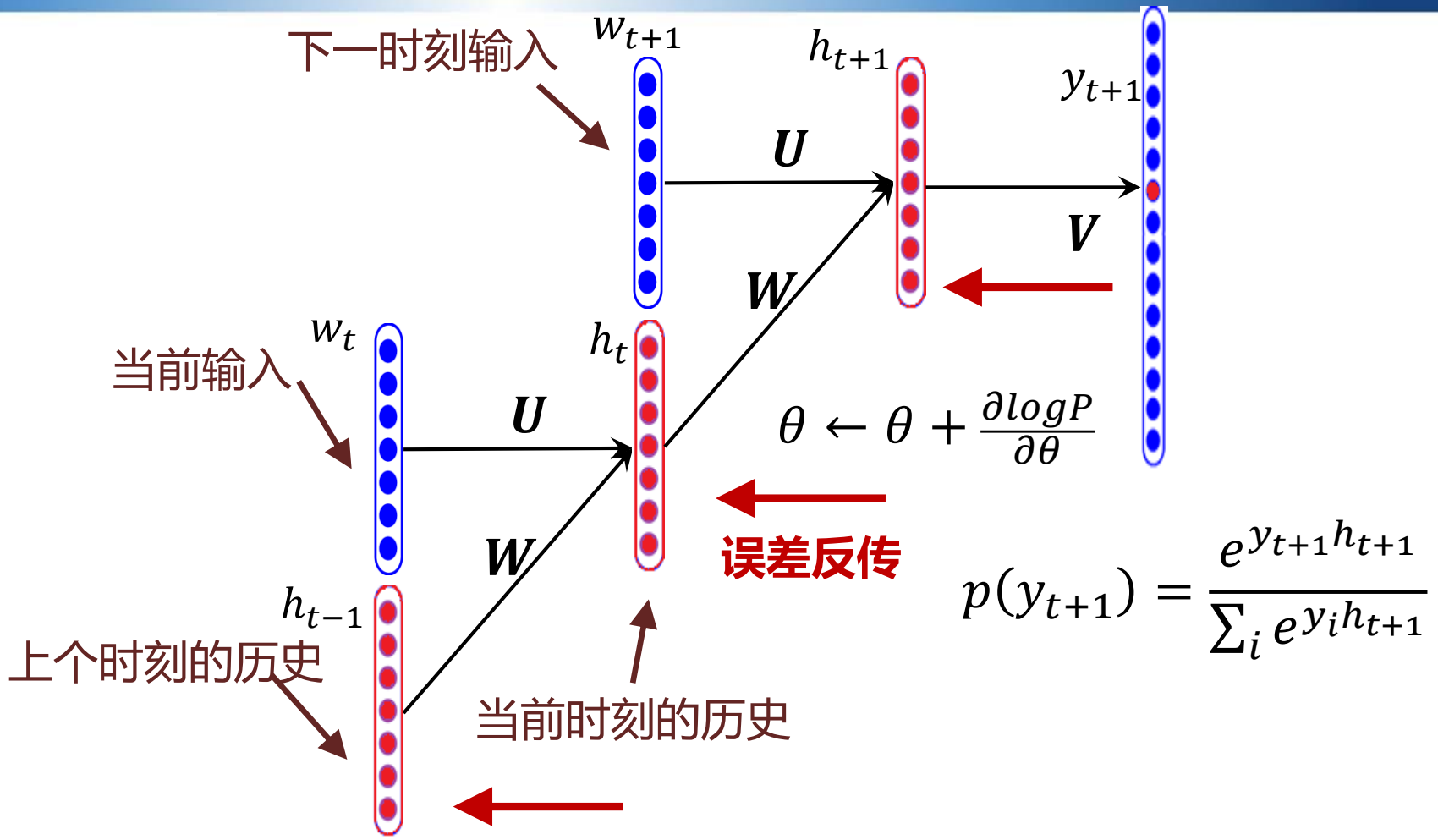
循环神经网络

- 输入: $t-1$时刻历史 h_{t-1} 与 t时刻输入 w_t
- 输出: t时刻历史 h_t 与 下个时刻$t+1$输入y_t 的概率

$$h_t = tanh(UL(w_t) + Wh_{t-1})$$



循环神经网络


下一时刻输入
w_{t+1}
h_{t+1}
y_{t+1}
U
V
W
w_t
h_t
当前输入
U
$\theta \leftarrow \theta + \frac{\partial logP}{\partial \theta}$
误差反传
W
h_{t-1}
上个时刻的历史
当前时刻的历史
$p(y_{t+1}) = \frac{e^{y_{t+1}h_{t+1}}}{\sum_i e^{y_i h_{t+1}}}$

神经机器翻译

$$h_s = tanh(UL(w_s) + W h_{s-1})$$

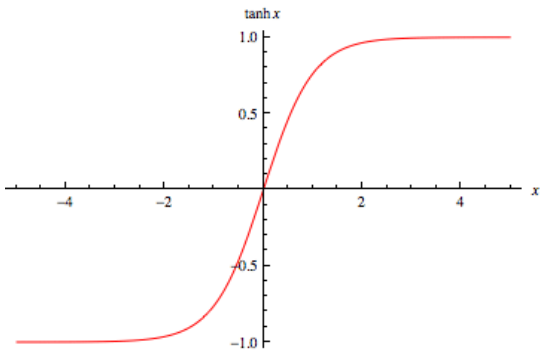
$L(w_s)$: $w_s \longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in R^3$ $w_{我} \longrightarrow \begin{bmatrix} 0.1 \\ 0.9 \\ 0.6 \end{bmatrix}$ 随机初始化

h_{s-1}: 上一时刻的历史信息 $h_0 = \begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$

$$U = \begin{bmatrix} 0.1 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.0 \\ 0.4 & 0.0 & 0.2 \end{bmatrix} \in R^{3\times3} \qquad W = \begin{bmatrix} 0.0 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.3 \\ 0.0 & 0.4 & 0.1 \end{bmatrix} \in R^{3\times3}$$

$$z = UL(w_s) + W h_{s-1} \in R^3$$

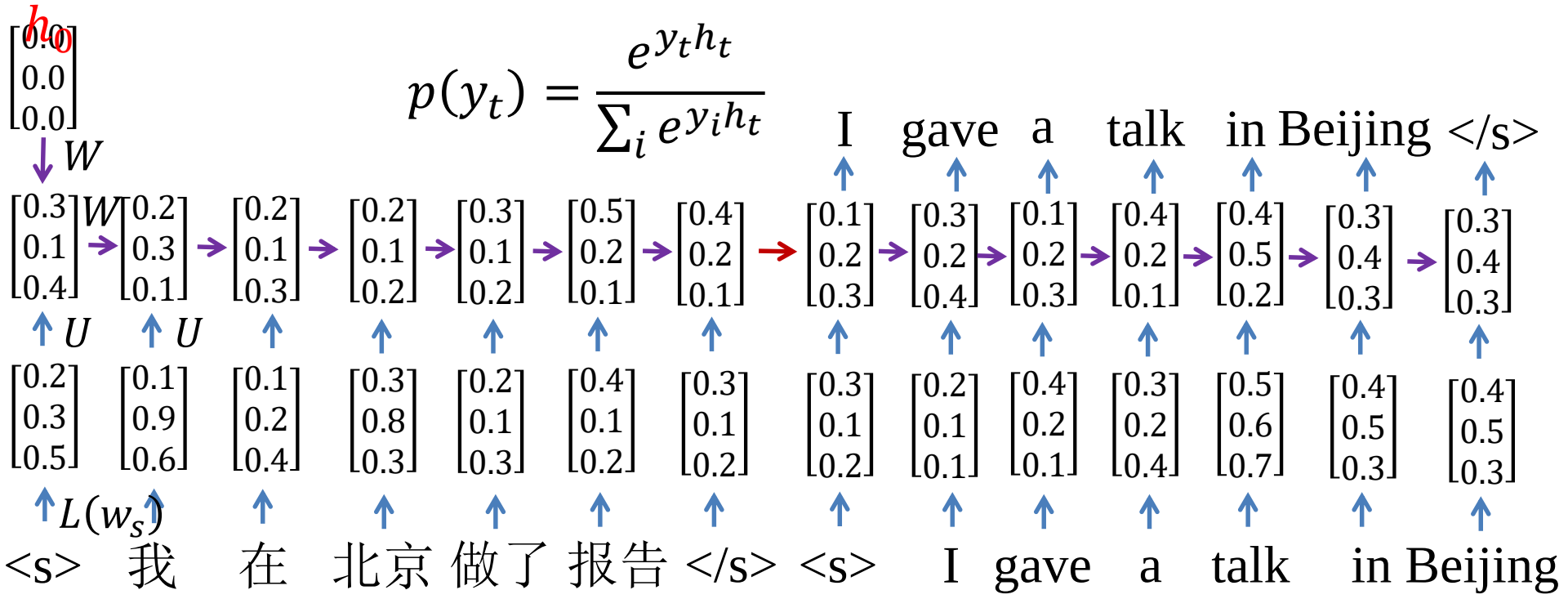
$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \longrightarrow$$



神经机器翻译

$$h_s = tanh(UL(w_s) + Wh_{s-1}) \quad h_t = tanh(UL(w_t) + Wh_{t-1})$$

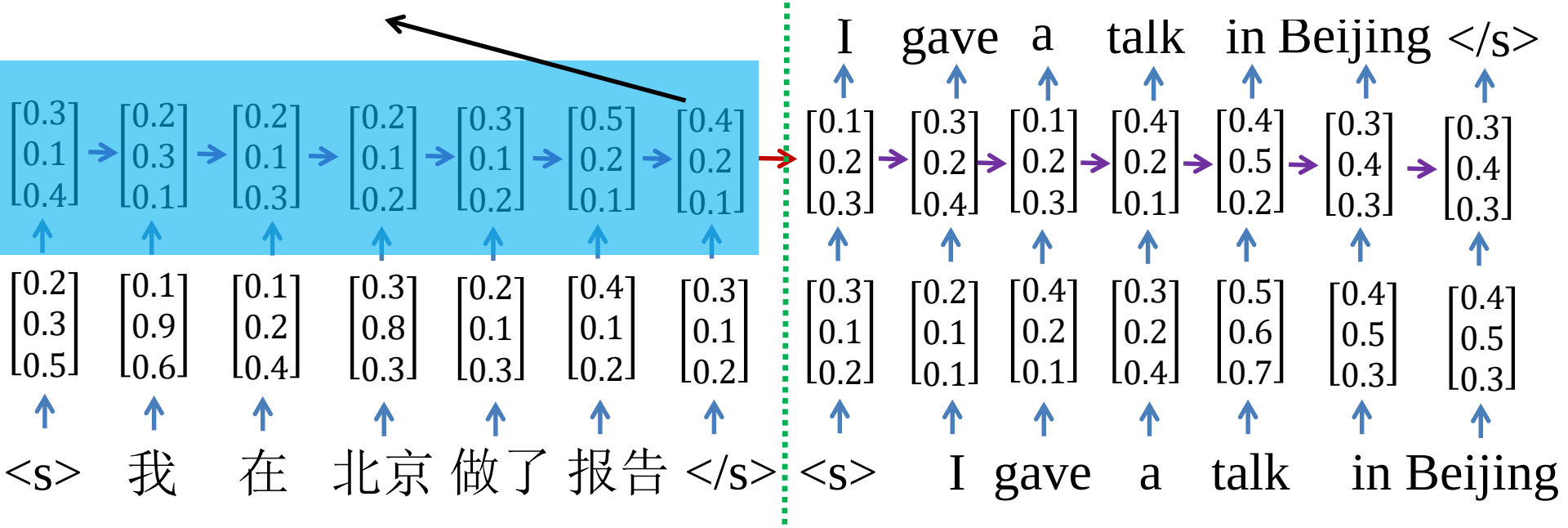
$$p(y_t) = \frac{e^{y_t h_t}}{\sum_i e^{y_i h_t}}$$



神经机器翻译

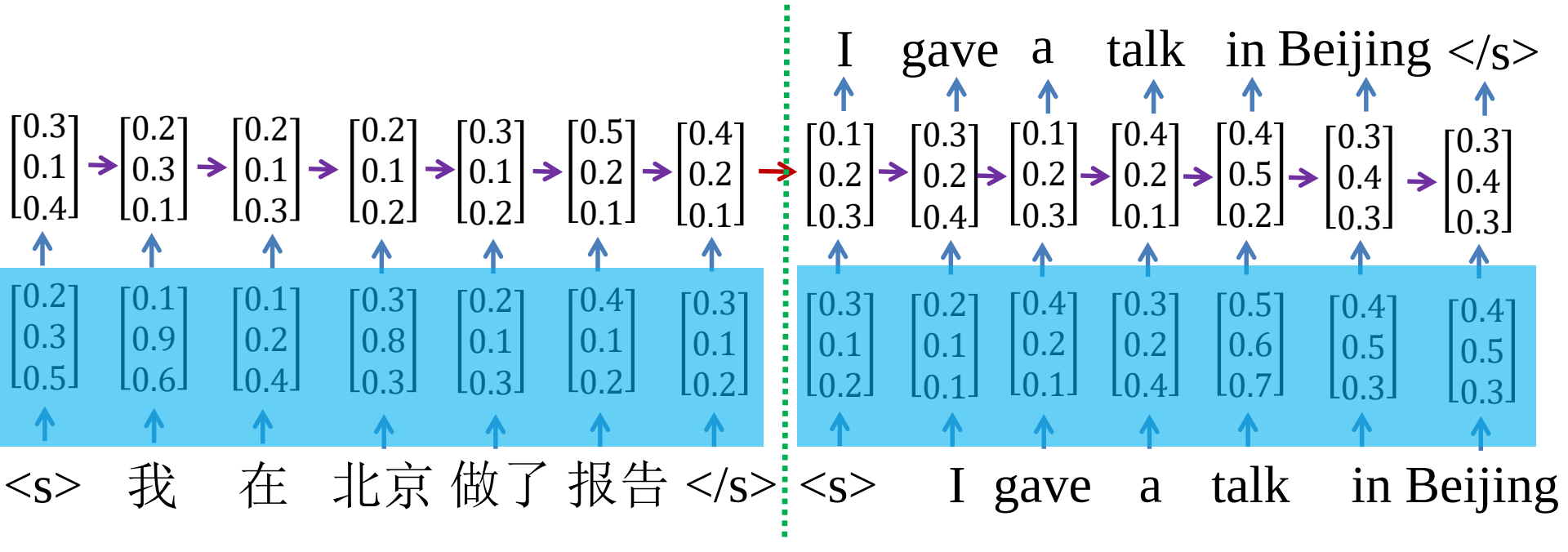
将源语言句子编码成一个
实数向量语义表示

将源语言句子的语义表示
解码生成目标语言句子



神经机器翻译

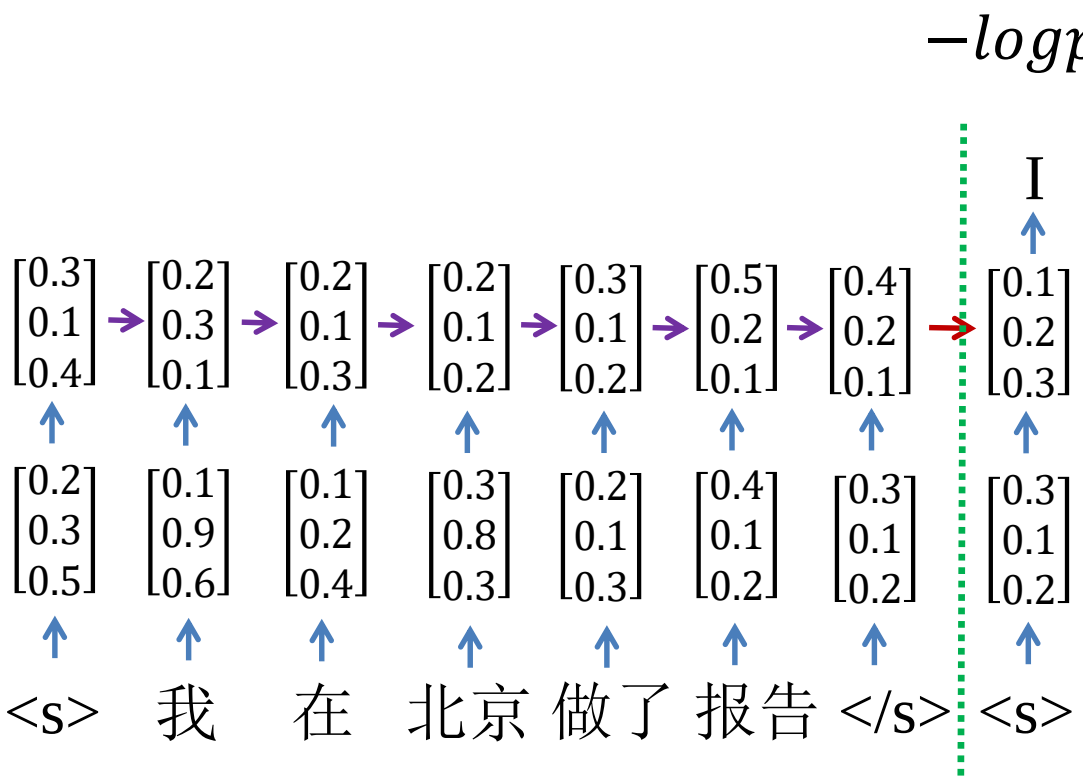
词向量随机初始化，在训练过程中进行优化！



源语言词向量

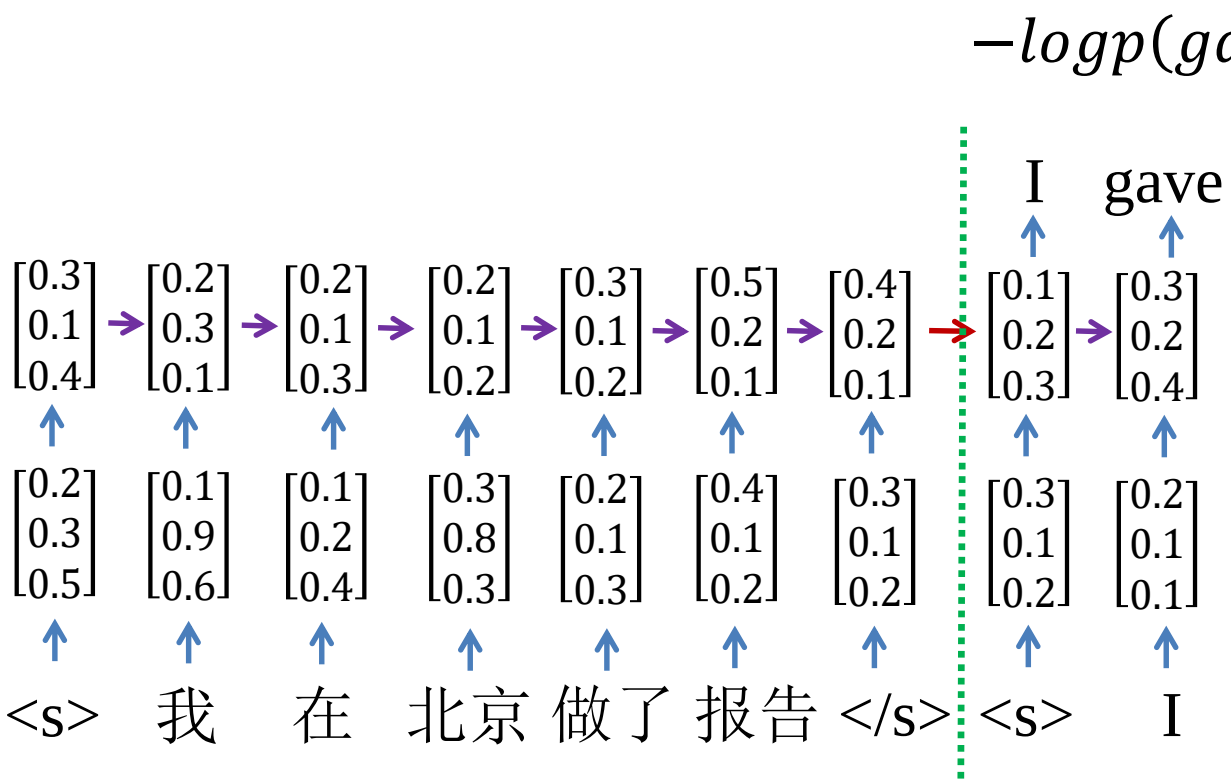
目标语言词向量

神经机器翻译



最大化 $P(target|source)$

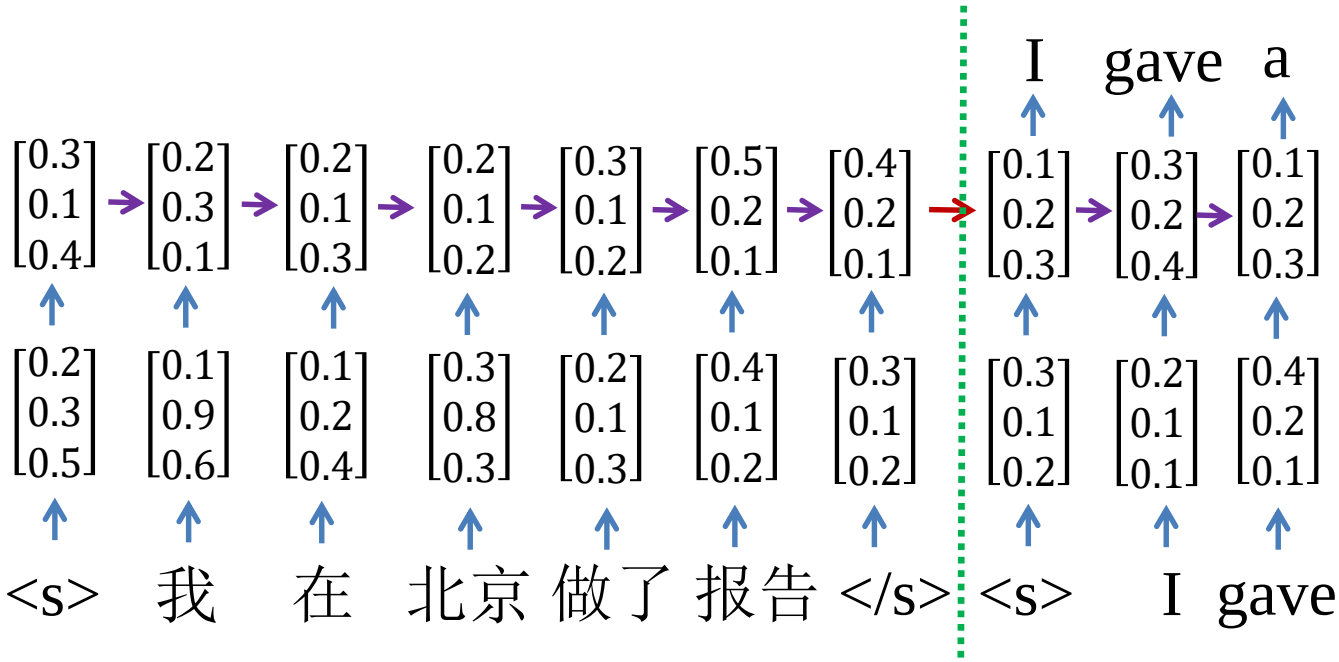
神经机器翻译



最大化 $P(target|source)$

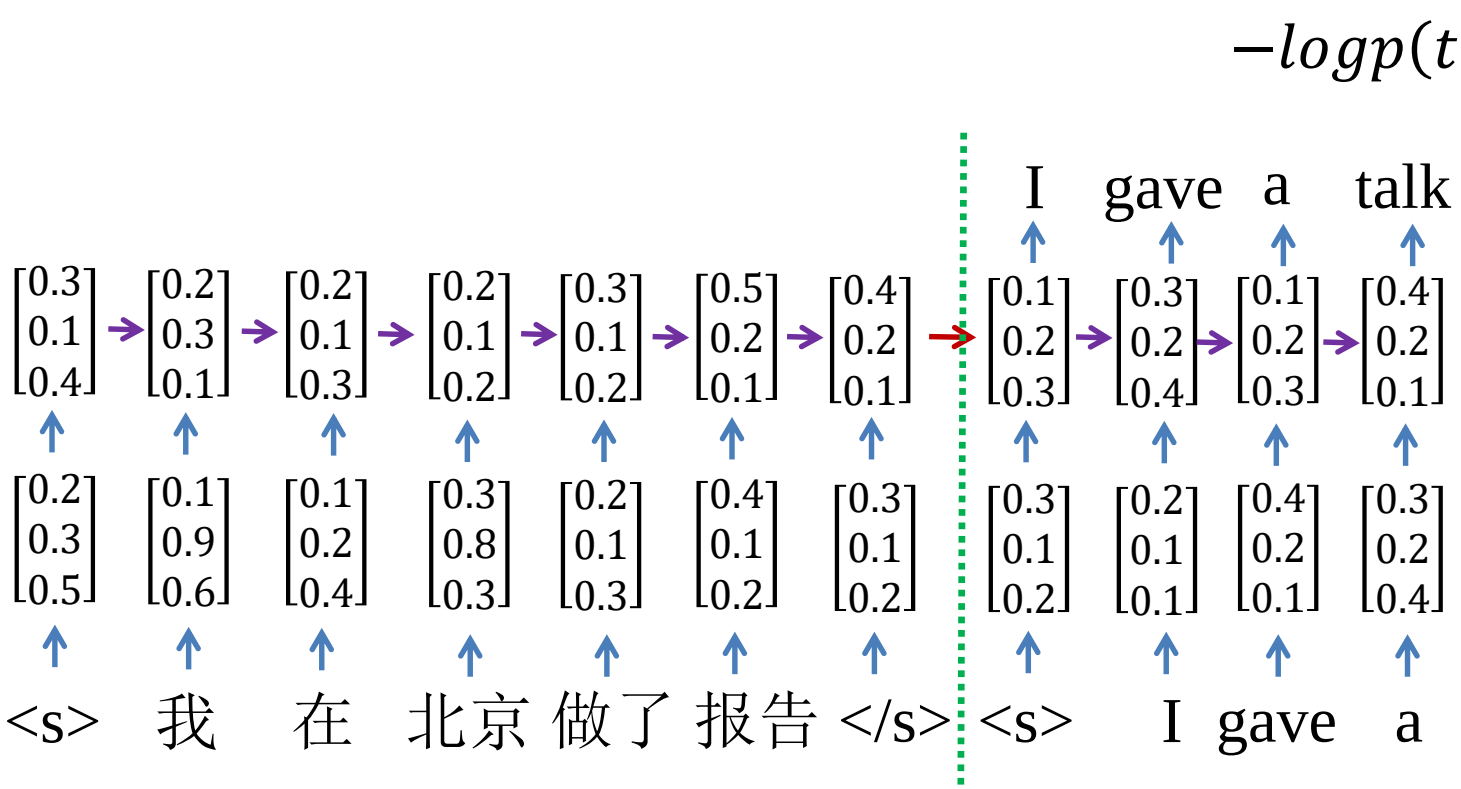
神经机器翻译

$-logp(a)$



最大化 $P(target|source)$

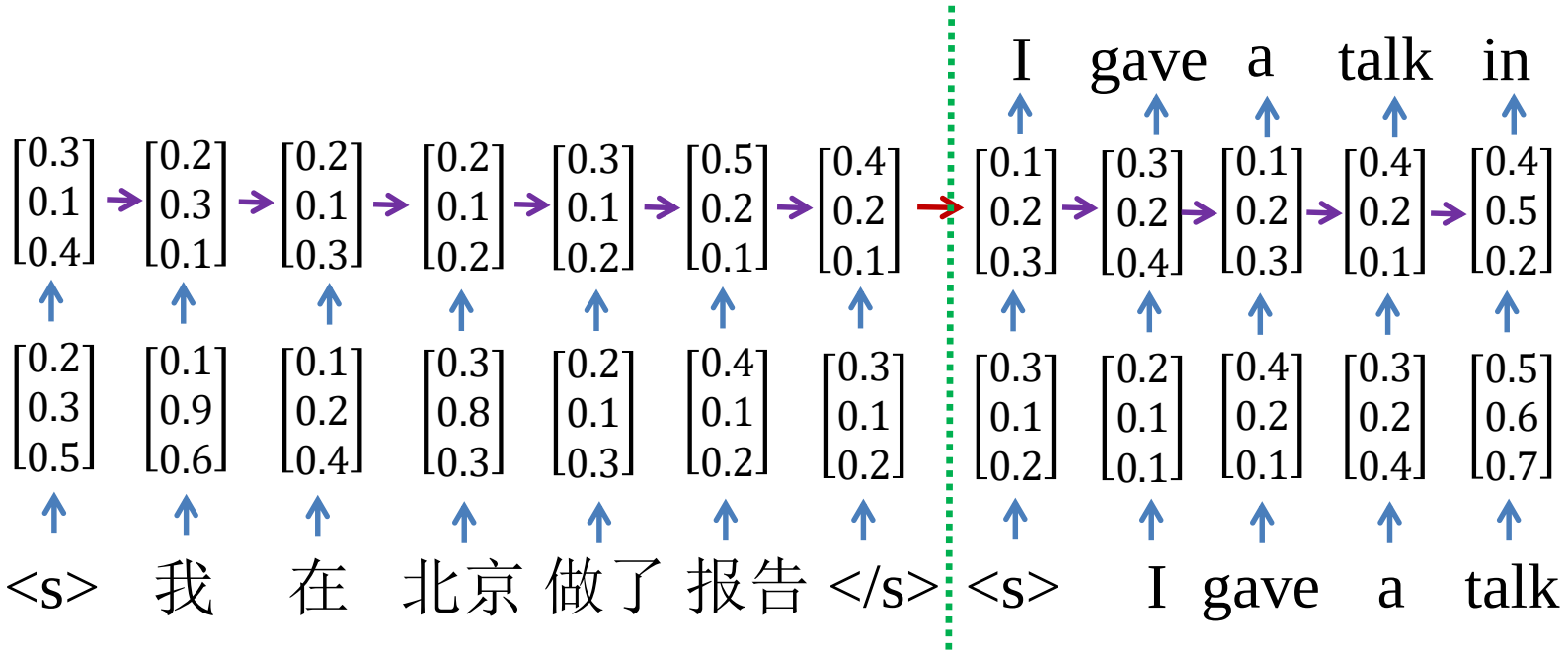
神经机器翻译



最大化 $P(target|source)$

神经机器翻译

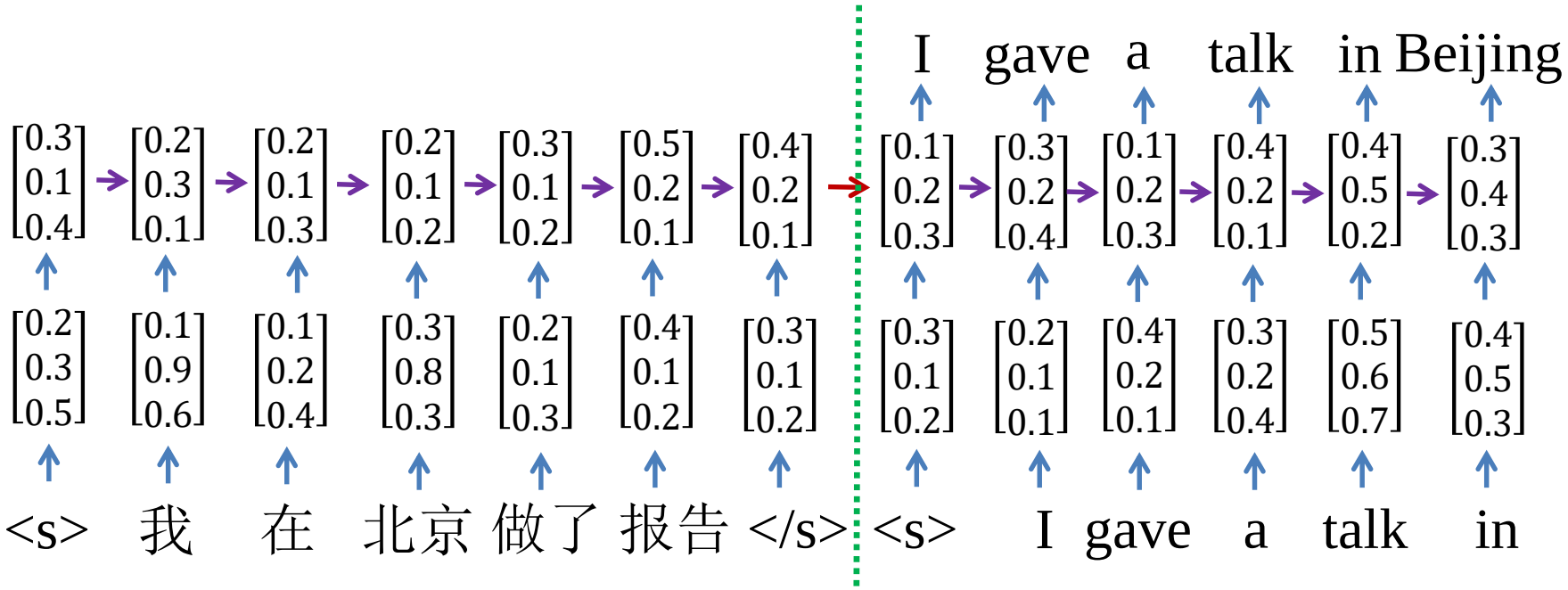
$-logp(in)$



最大化 $P(target|source)$

神经机器翻译

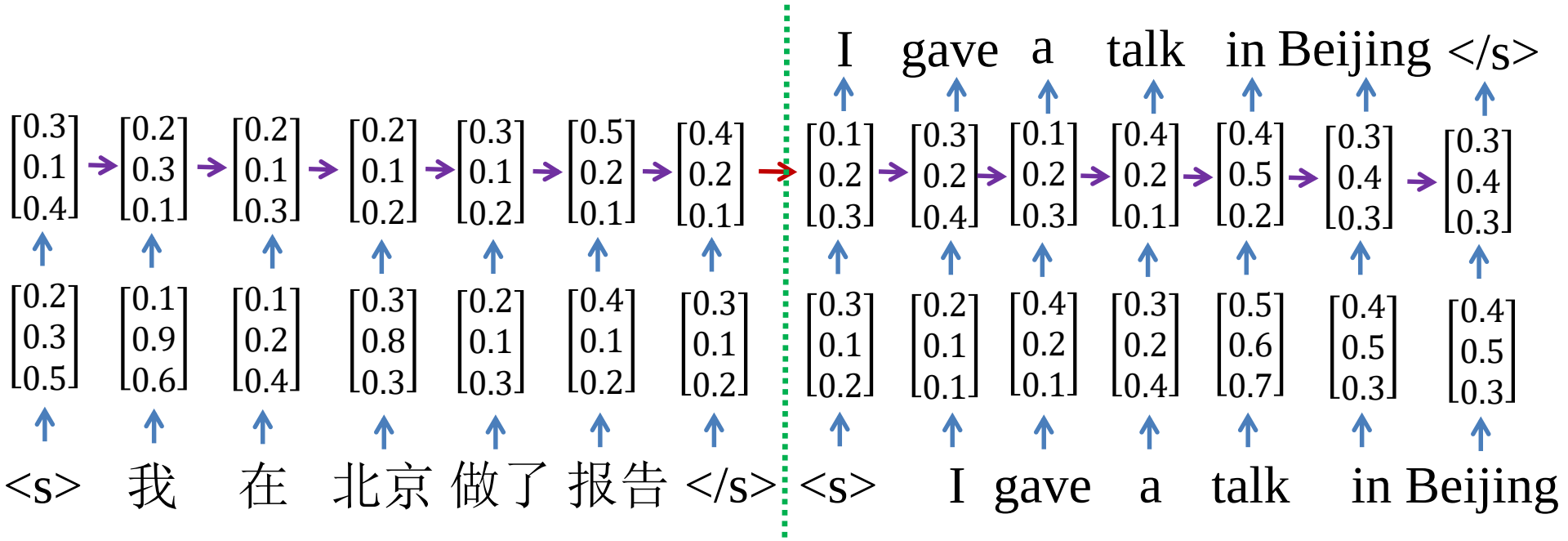
$-logp(Beijing)$



最大化 $P(target|source)$

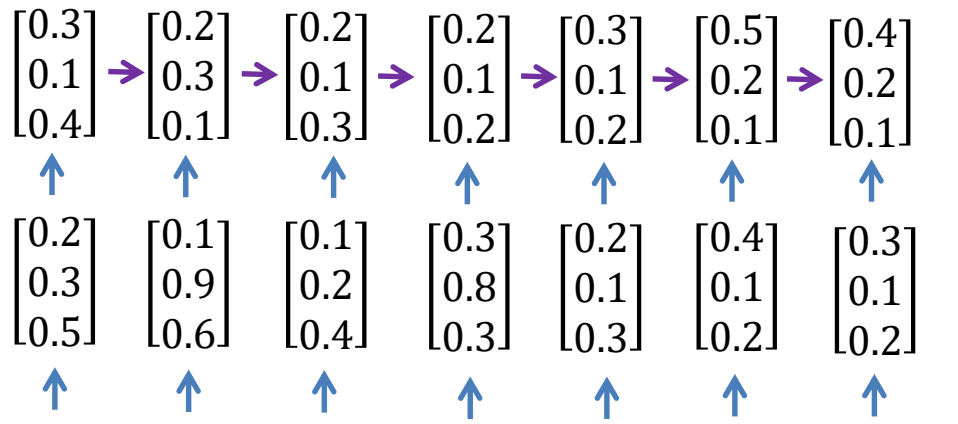
神经机器翻译

$-logp(\langle/s\rangle)$


I gave a talk in Beijing </s>
<s> 我 在 北京 做了 报告 </s>
<s> I gave a talk in Beijing</s>

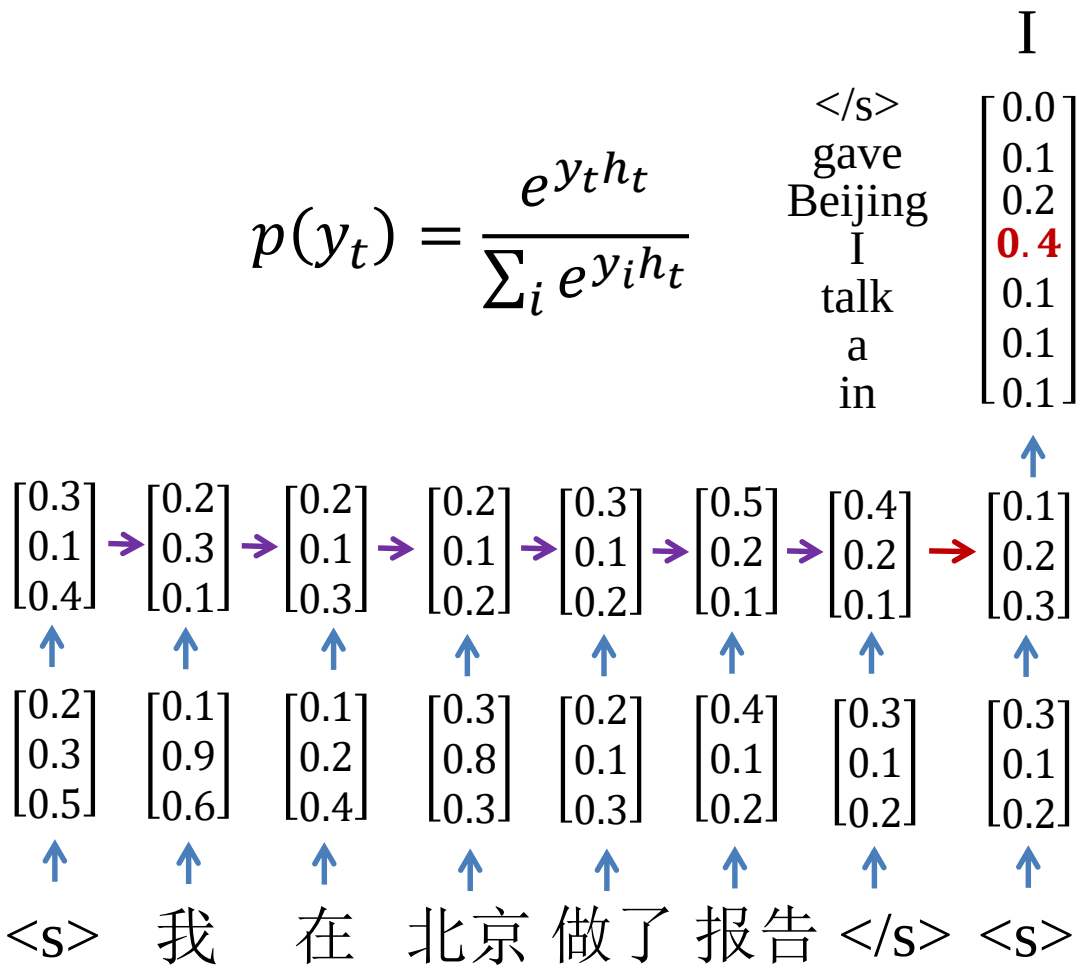
最大化 $P(target|source)$

神经机器翻译-测试

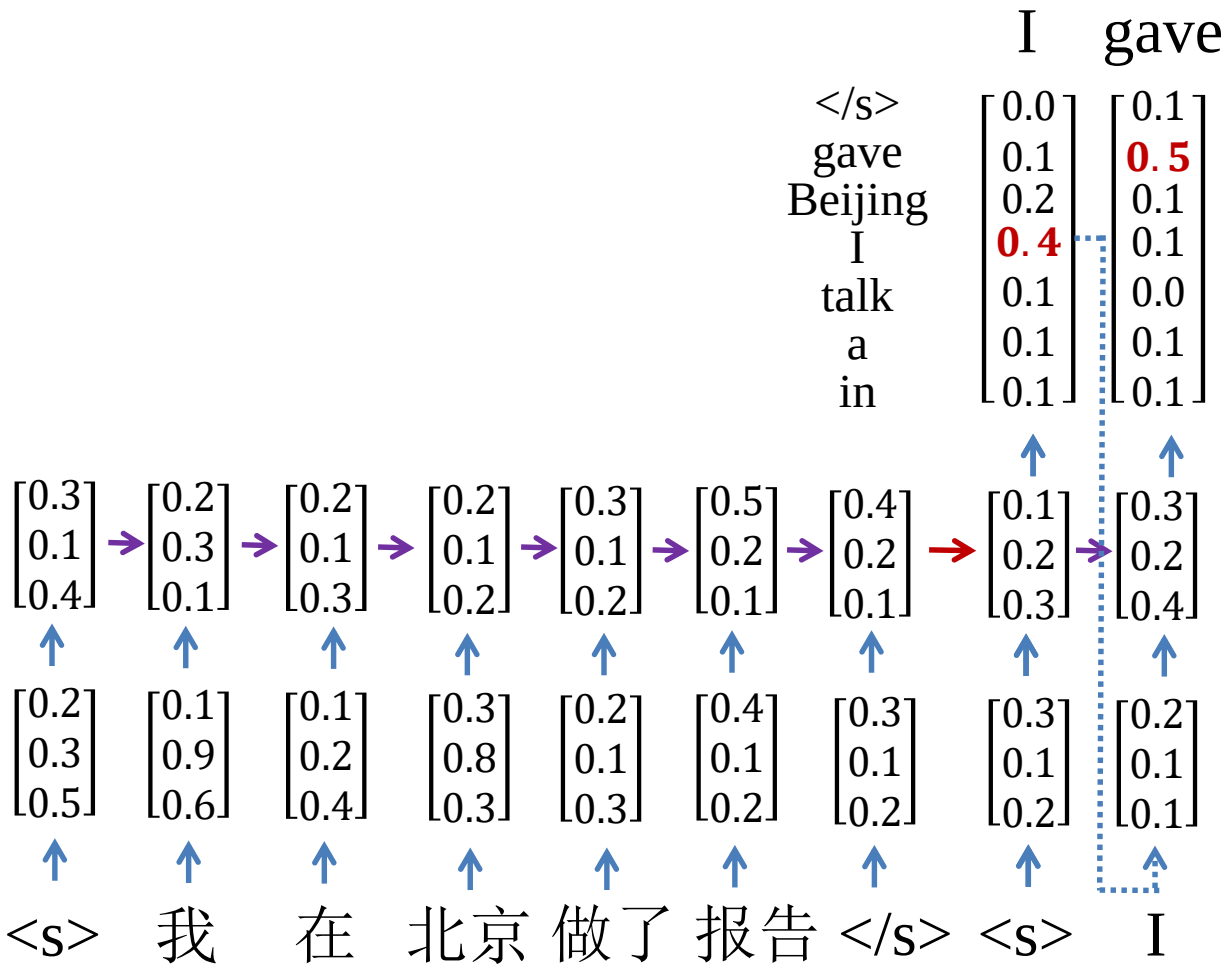

0.3 0.1 0.4
0.2 0.3 0.1
0.2 0.1 0.3
0.2 0.1 0.2
0.3 0.1 0.2
0.5 0.2 0.1
0.4 0.2 0.1
0.2 0.3 0.5
0.1 0.9 0.6
0.1 0.2 0.4
0.3 0.8 0.3
0.2 0.1 0.3
0.4 0.1 0.2
0.3 0.1 0.2
<s> 我 在 北京 做了 报告 </s>

神经机器翻译-测试

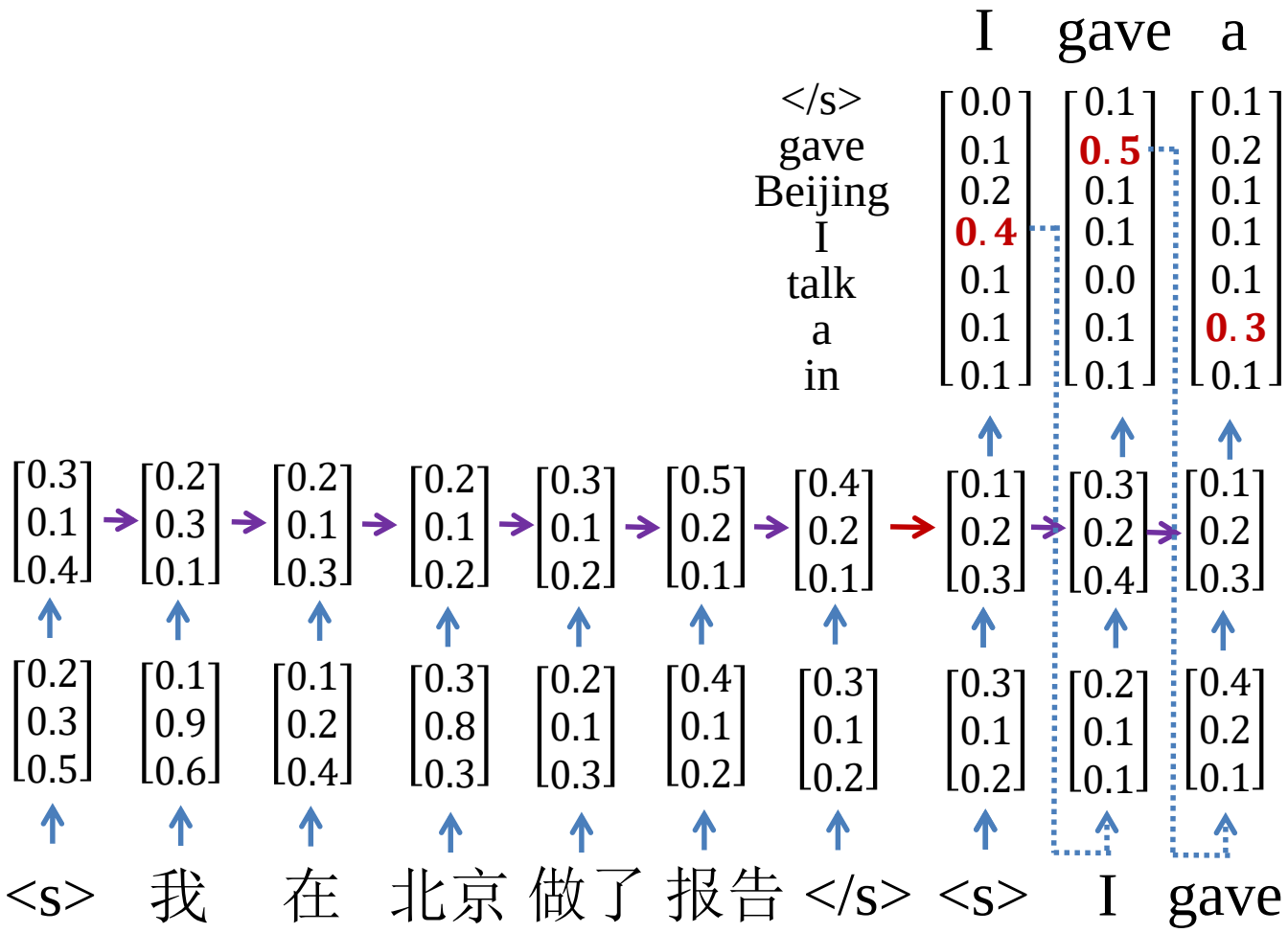
$$p(y_t) = \frac{e^{y_t h_t}}{\sum_i e^{y_i h_t}}$$



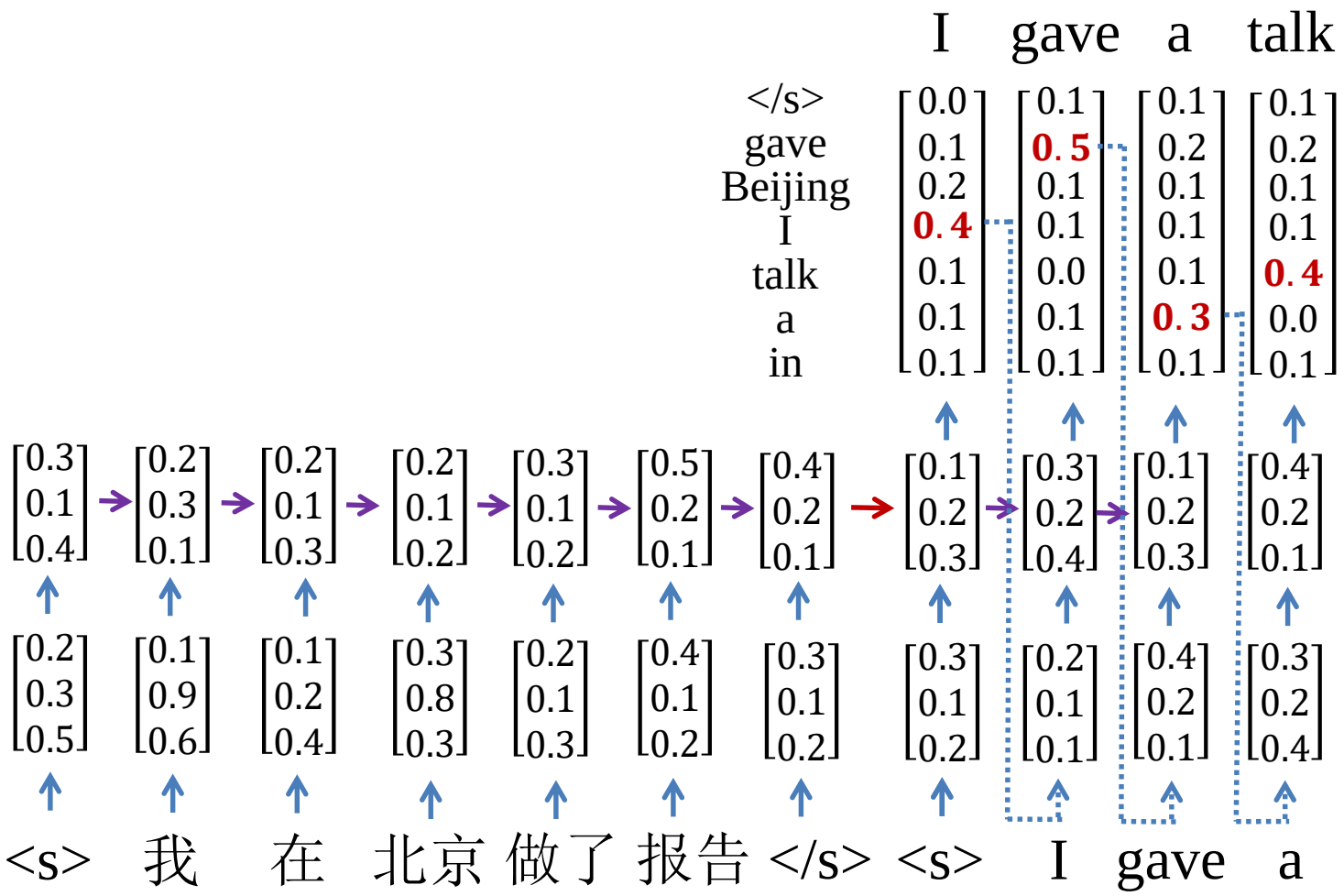
神经机器翻译-测试


I
gave
</s>
gave
Beijing
I
talk
a
in
0.0 0.1 0.2 0.4 0.1 0.1 0.1
0.1 0.5 0.1 0.1 0.0 0.1 0.1
0.3 0.1 0.4
0.2 0.3 0.1
0.2 0.1 0.3
0.2 0.1 0.2
0.3 0.1 0.2
0.5 0.2 0.1
0.4 0.2 0.1
0.1 0.2 0.3
0.3 0.2 0.4
0.2 0.3 0.5
0.1 0.9 0.6
0.1 0.2 0.4
0.3 0.8 0.3
0.2 0.1 0.3
0.4 0.1 0.2
0.3 0.1 0.2
0.3 0.1 0.2
0.2 0.1 0.1
<s>
我
在
北京
做了
报告
</s>
<s>
I

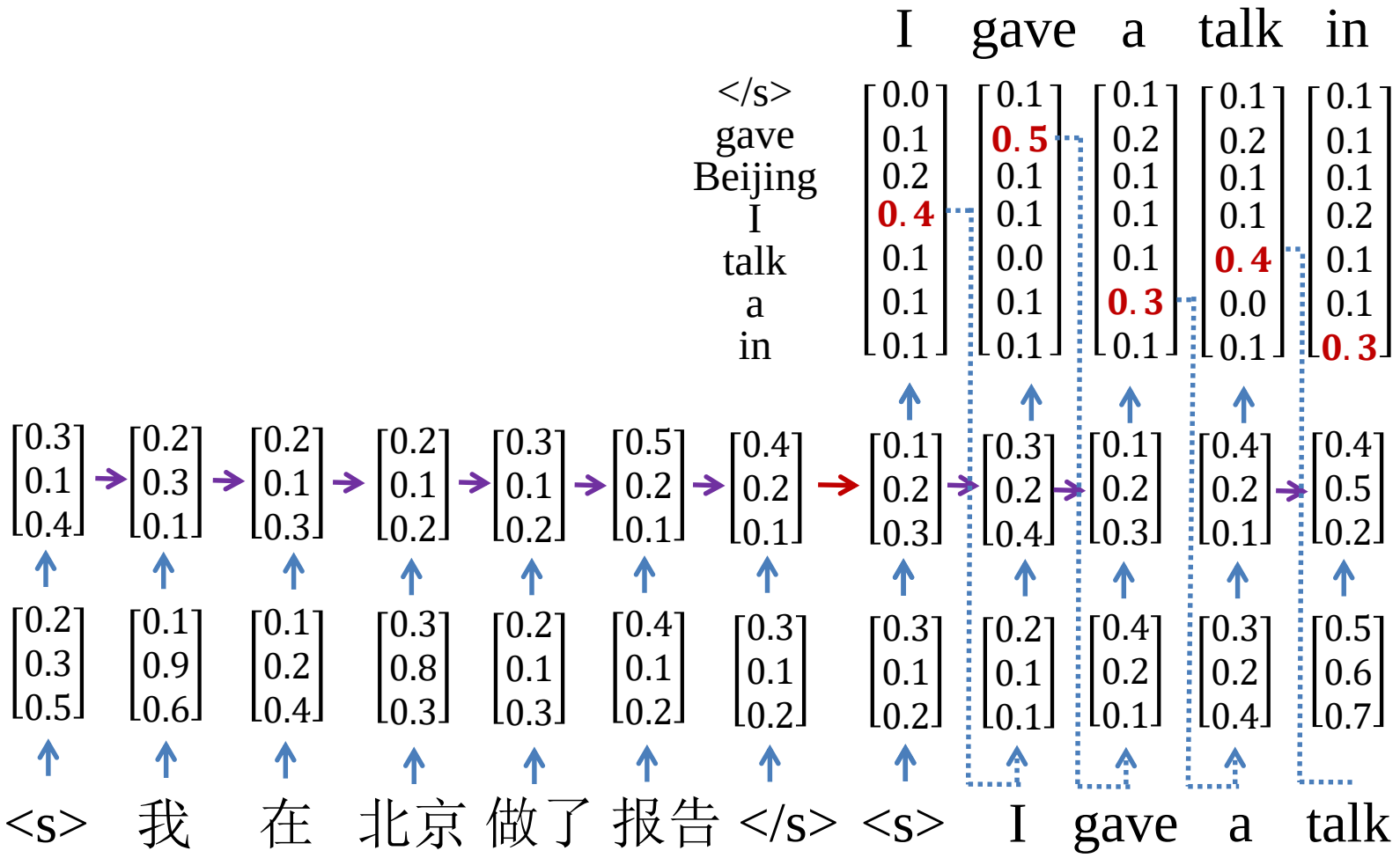
神经机器翻译-测试


I gave a
</s>
gave
Beijing
I
talk
a
in
0.0 0.1 0.1
0.1 0.5 0.2
0.2 0.1 0.1
0.4 0.1 0.1
0.1 0.0 0.1
0.1 0.1 0.3
0.1 0.1 0.1
0.3 0.2 0.1
0.2 0.3 0.1
0.2 0.1 0.3
0.2 0.1 0.2
0.3 0.1 0.2
0.5 0.2 0.1
0.4 0.2 0.1
0.2 0.3 0.5
0.1 0.9 0.6
0.1 0.2 0.4
0.3 0.8 0.3
0.2 0.1 0.3
0.4 0.1 0.2
0.3 0.1 0.2
0.3 0.1 0.2
0.2 0.1 0.1
0.4 0.2 0.1
<s> 我 在 北京 做了 报告 </s> <s> I gave

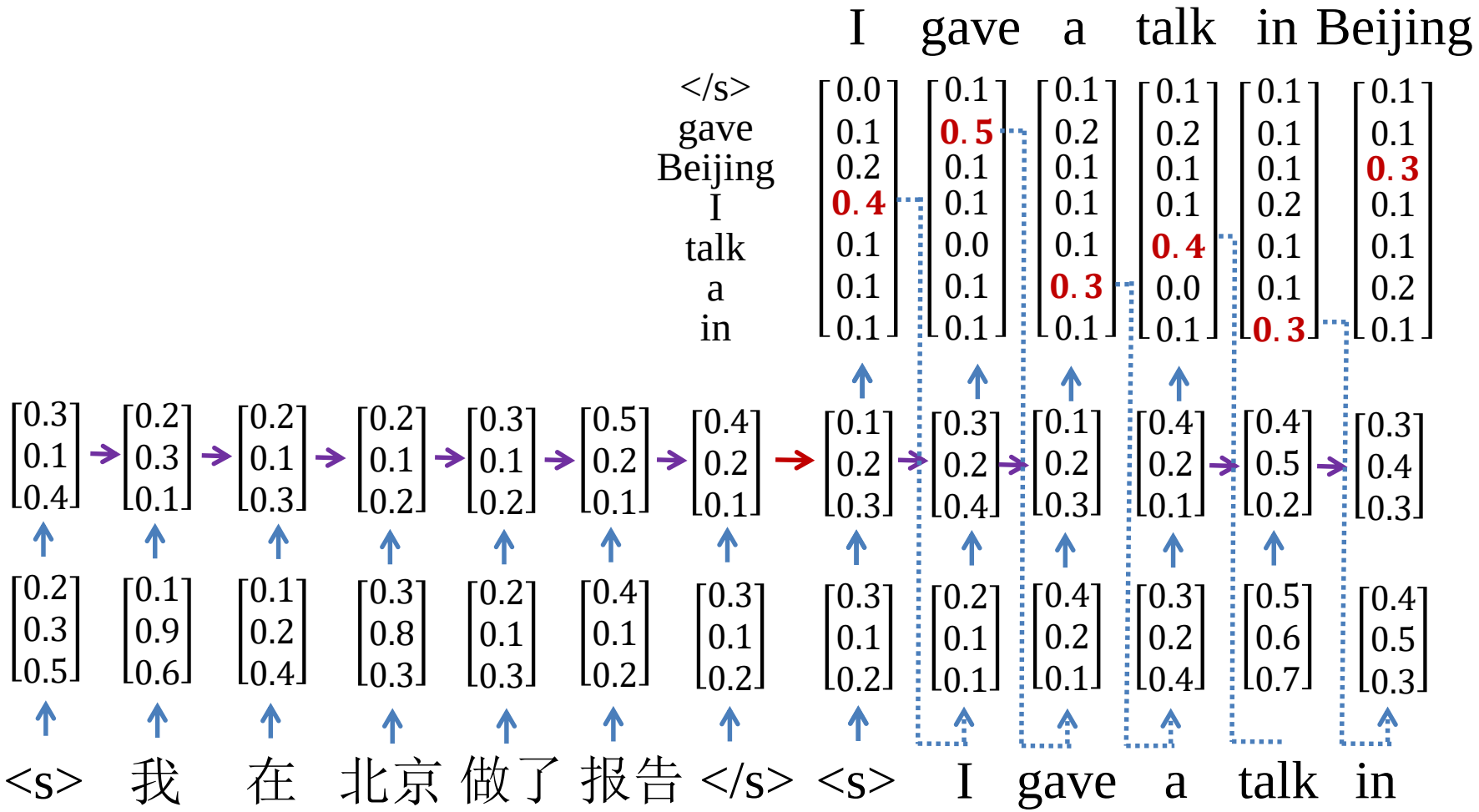
神经机器翻译-测试


I gave a talk
</s>
gave
Beijing
I
talk
a
in
0.0 0.1 0.2 0.4 0.1 0.1 0.1
0.1 0.5 0.1 0.1 0.0 0.1 0.1
0.1 0.2 0.1 0.1 0.1 0.3 0.1
0.1 0.2 0.1 0.1 0.4 0.0 0.1
0.3 0.1 0.4
0.2 0.3 0.1
0.2 0.1 0.3
0.2 0.1 0.2
0.3 0.1 0.2
0.5 0.2 0.1
0.4 0.2 0.1
0.1 0.2 0.3
0.3 0.2 0.4
0.1 0.2 0.3
0.4 0.2 0.1
0.2 0.3 0.5
0.1 0.9 0.6
0.1 0.2 0.4
0.3 0.8 0.3
0.2 0.1 0.3
0.4 0.1 0.2
0.3 0.1 0.2
0.3 0.1 0.2
0.2 0.1 0.1
0.4 0.2 0.1
0.3 0.2 0.4
<s> 我 在 北京 做了 报告 </s> <s> I gave a

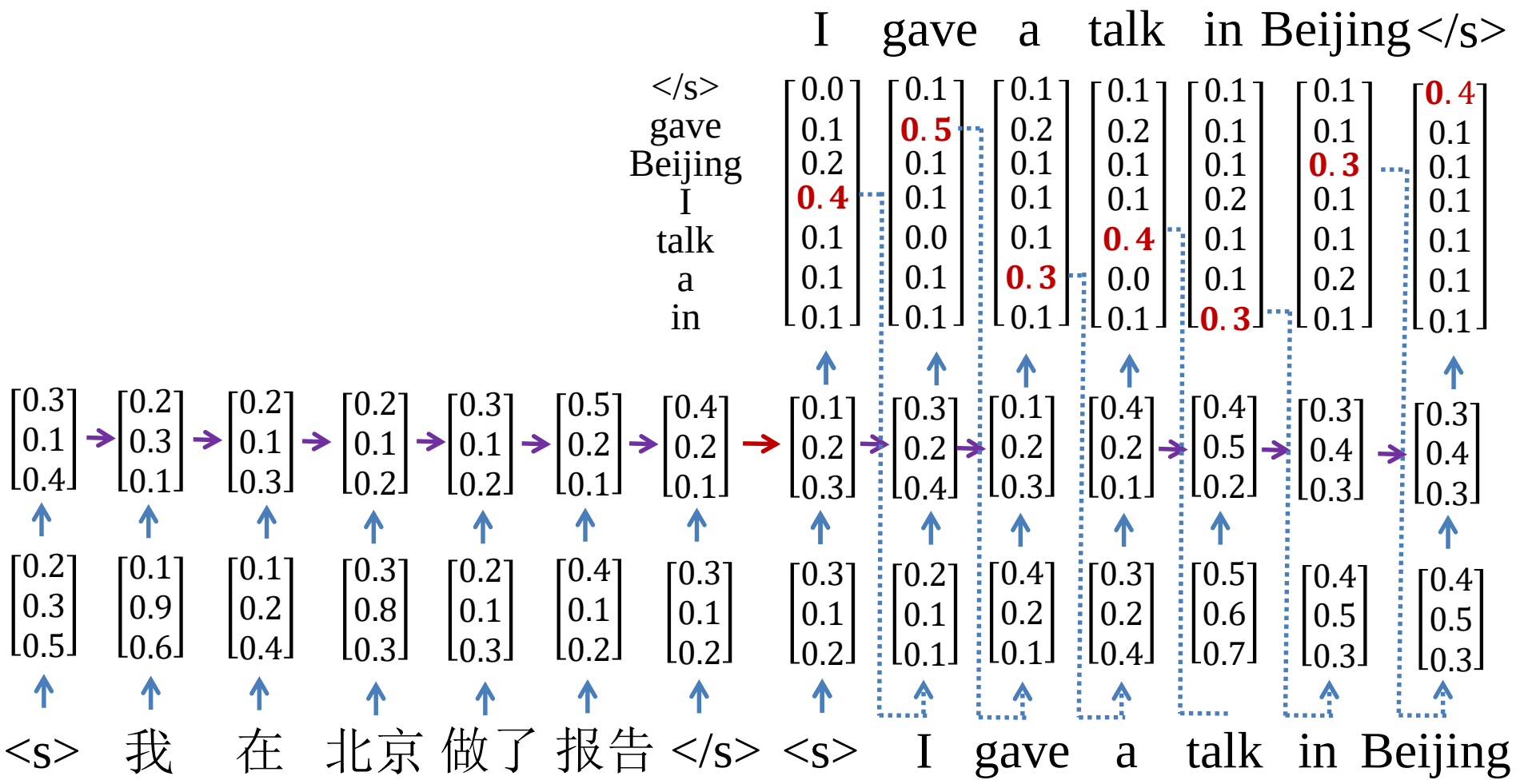
神经机器翻译-测试


I gave a talk in
</s>
gave
Beijing
I
talk
a
in
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk

神经机器翻译-测试

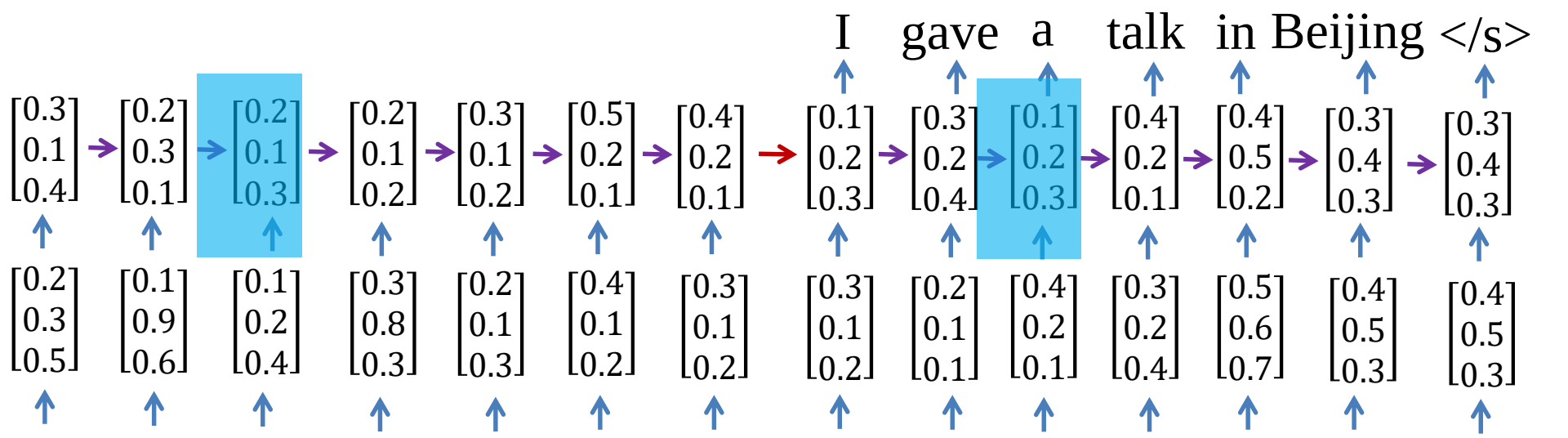

I gave a talk in Beijing
</s>
gave
Beijing
I
talk
a
in
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk in

神经机器翻译-测试

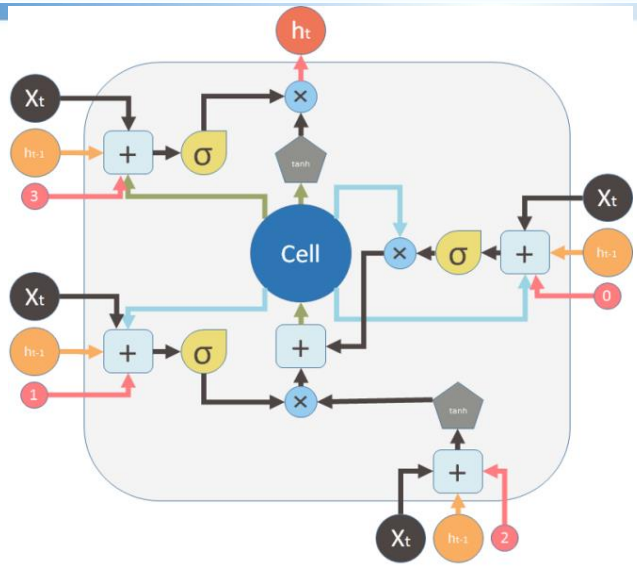

I gave a talk in Beijing </s>
</s>
gave
Beijing
I
talk
a
in
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk in Beijing

神经机器翻译-计算单元

$$h_s = tanh(UL(w_s) + Wh_{s-1}) \quad h_t = tanh(UL(w_t) + Wh_{t-1})$$



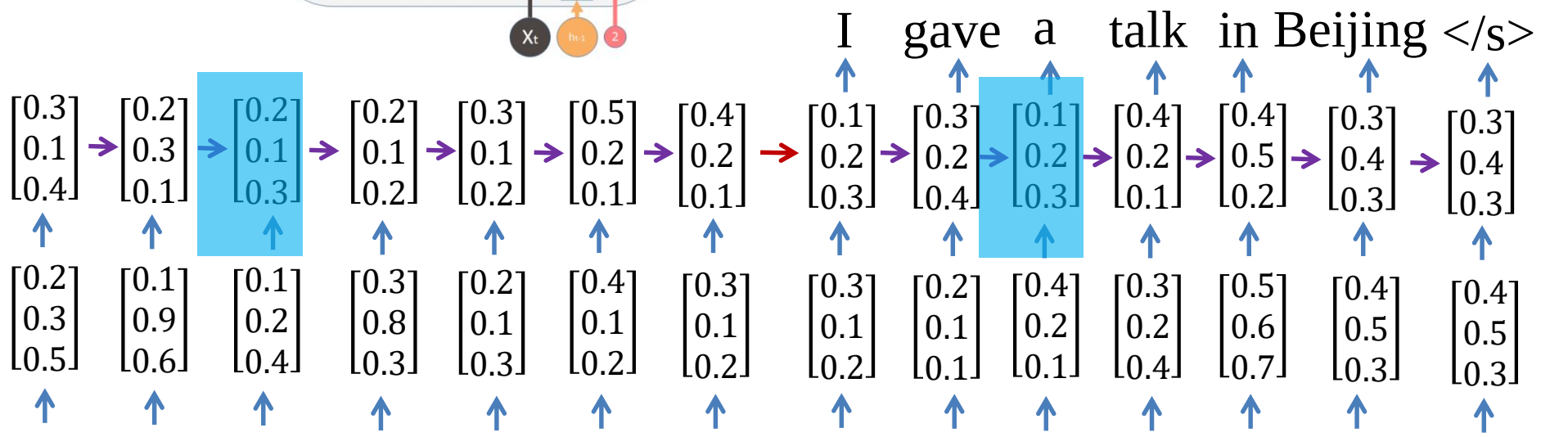
神经机器翻译-计算单元



LSTM计算单元

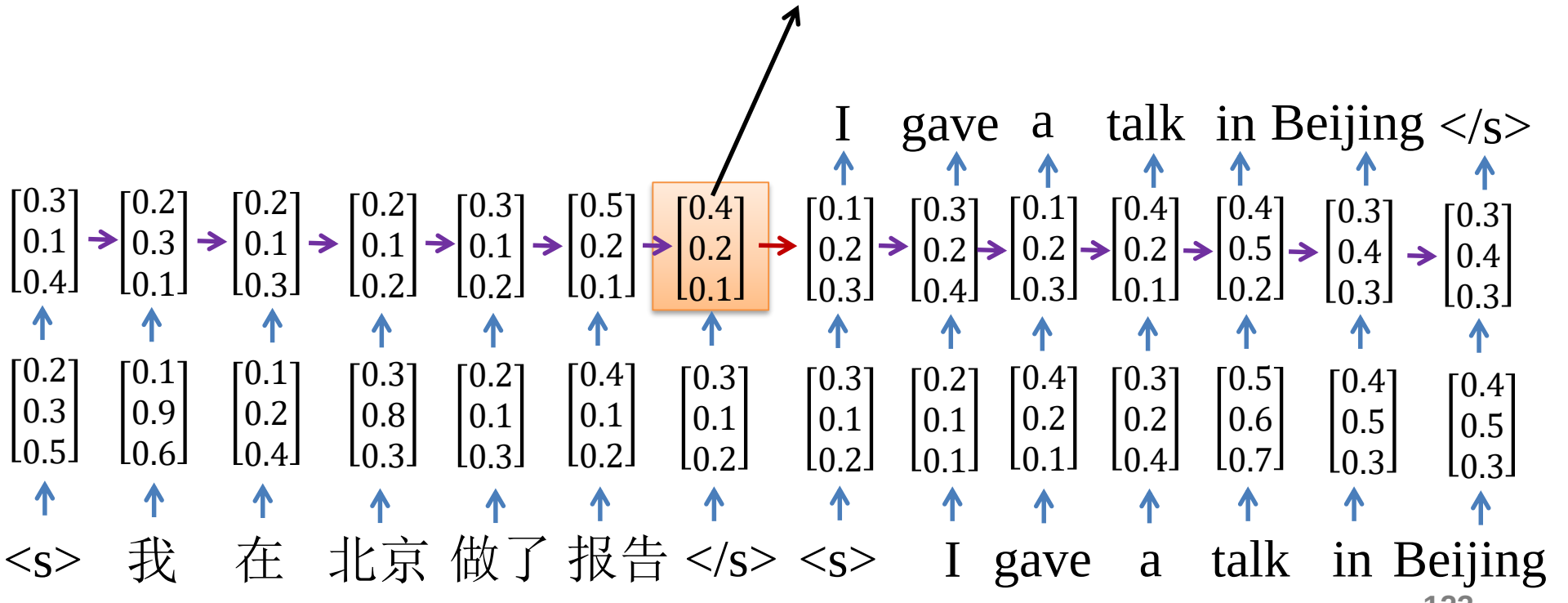
$$h_s = LSTM(w_s, h_{s-1})$$

$$h_t = LSTM(w_t, h_{t-1})$$

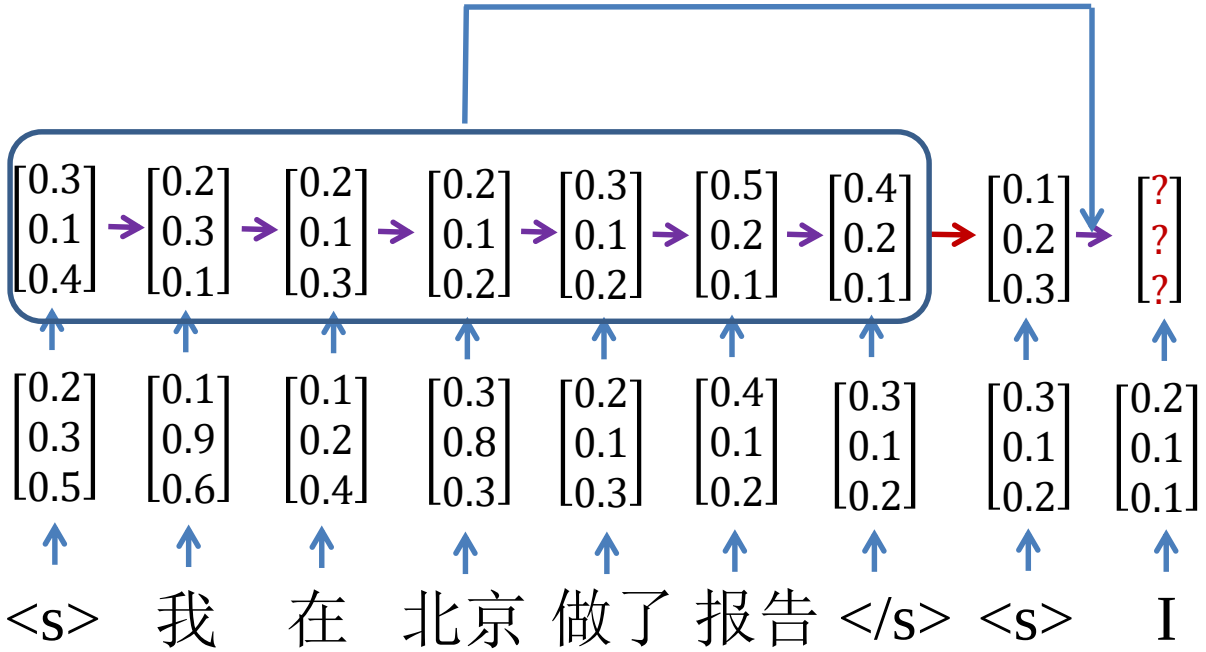


神经机器翻译-计算单元

一个实数向量无法表示
源语言句子的完整语义

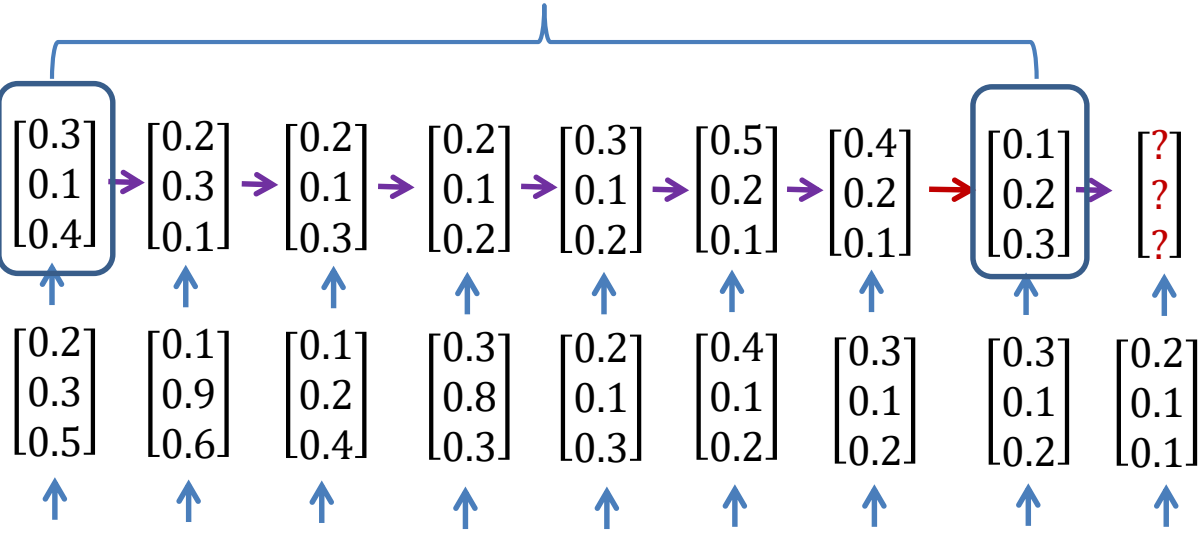


神经机器翻译-注意机制

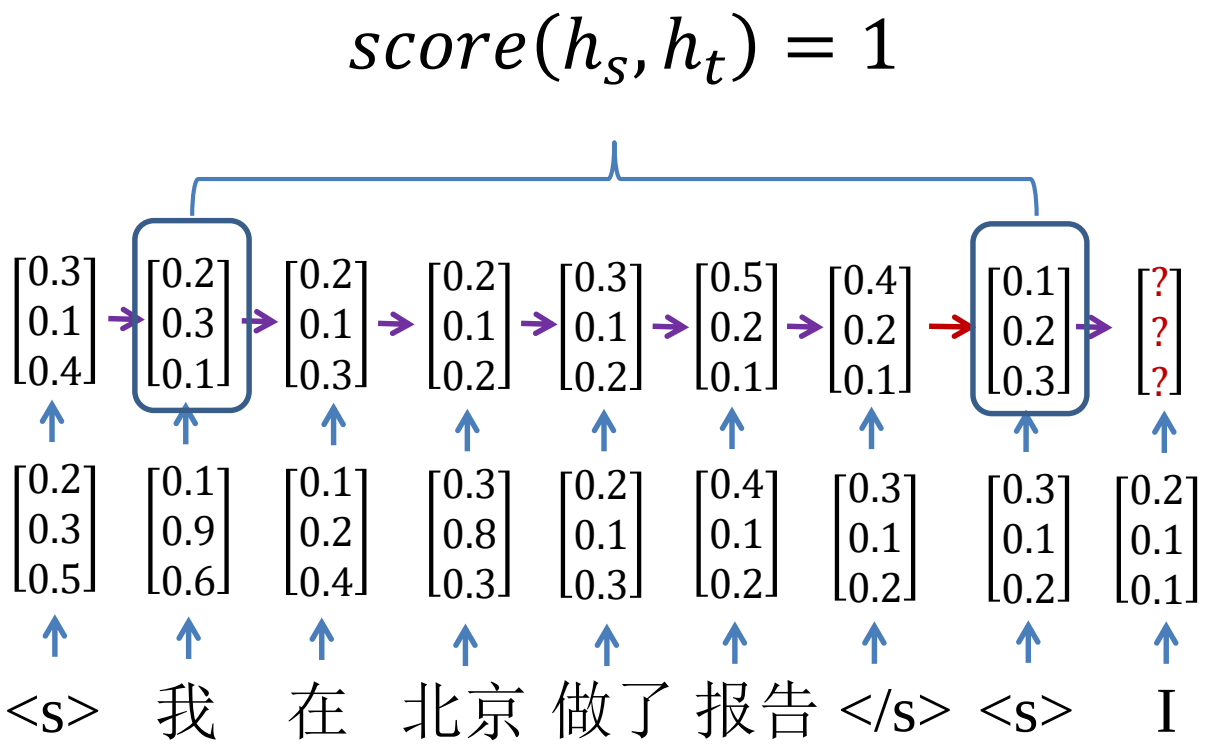

[0.3 0.1 0.4]
[0.2 0.3 0.1]
[0.2 0.1 0.3]
[0.2 0.1 0.2]
[0.3 0.1 0.2]
[0.5 0.2 0.1]
[0.4 0.2 0.1]
[0.1 0.2 0.3]
[? ? ?]
[0.2 0.3 0.5]
[0.1 0.9 0.6]
[0.1 0.2 0.4]
[0.3 0.8 0.3]
[0.2 0.1 0.3]
[0.4 0.1 0.2]
[0.3 0.1 0.2]
[0.3 0.1 0.2]
[0.2 0.1 0.1]
<s> 我 在 北京 做了 报告 </s> <s> I

神经机器翻译-注意机制

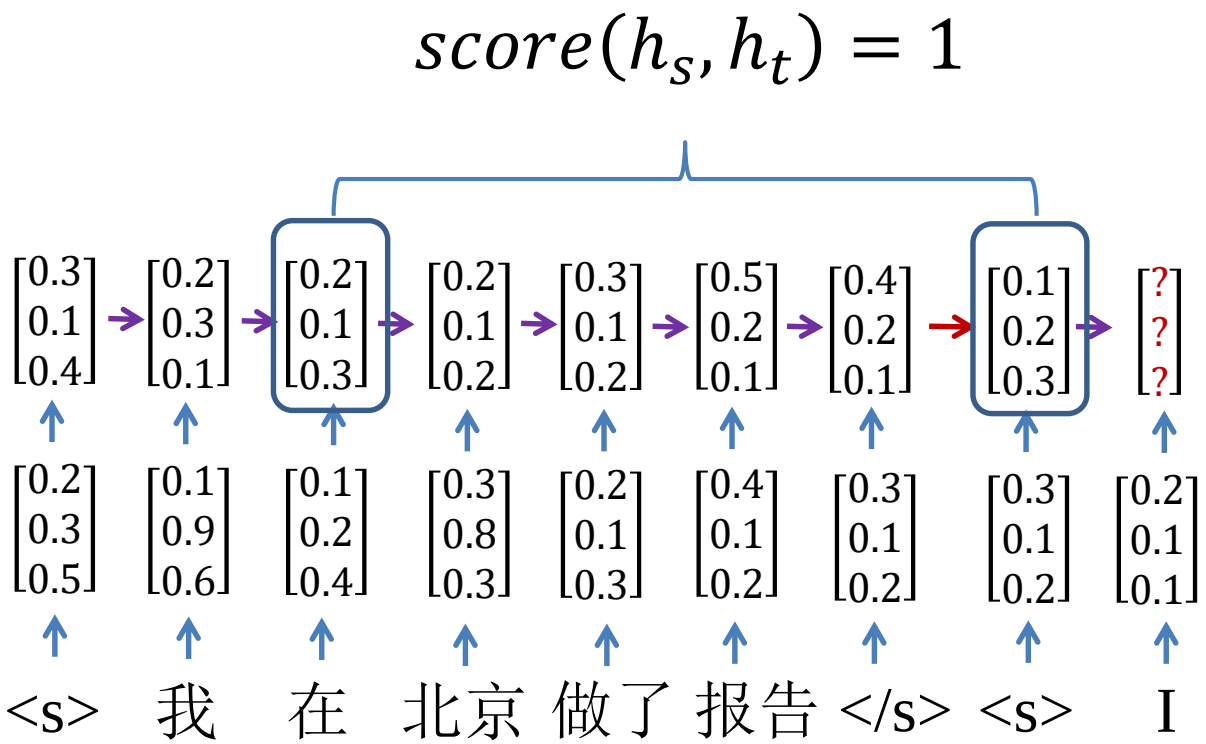
$$score(h_s, h_t) = 1$$



神经机器翻译-注意机制

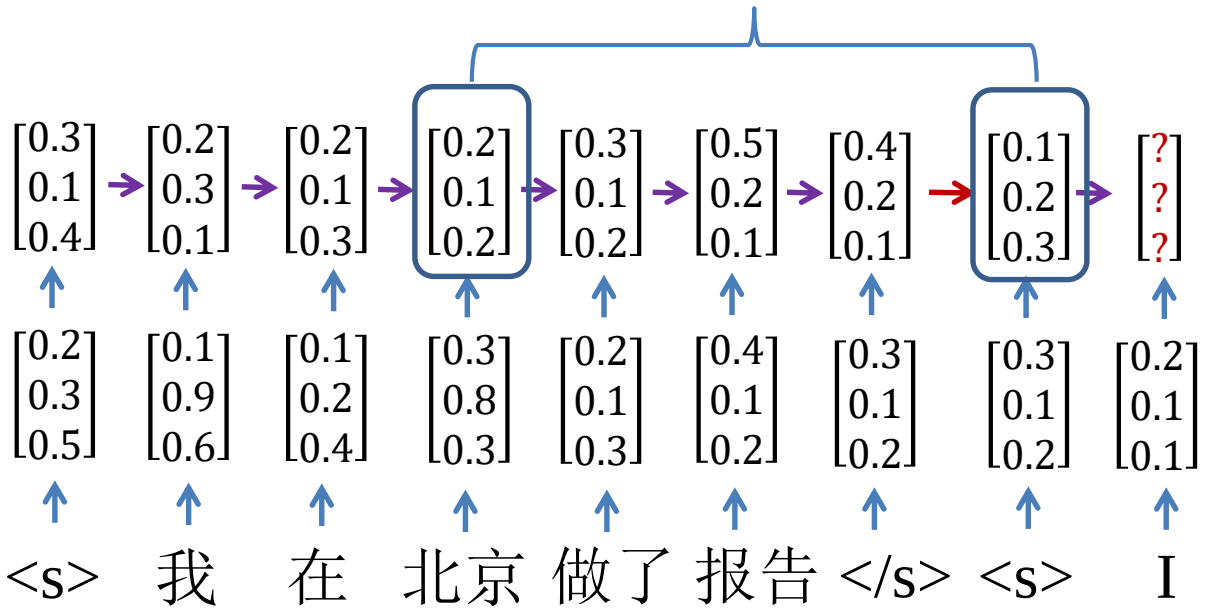


神经机器翻译-注意机制


$score(h_s, h_t) = 1$
[0.3 0.1 0.4] [0.2 0.3 0.1] [0.2 0.1 0.3] [0.2 0.1 0.2] [0.3 0.1 0.2] [0.5 0.2 0.1] [0.4 0.2 0.1] [0.1 0.2 0.3] [? ? ?]
[0.2 0.3 0.5] [0.1 0.9 0.6] [0.1 0.2 0.4] [0.3 0.8 0.3] [0.2 0.1 0.3] [0.4 0.1 0.2] [0.3 0.1 0.2] [0.3 0.1 0.2] [0.2 0.1 0.1]
<s> 我 在 北京 做了 报告 </s> <s> I

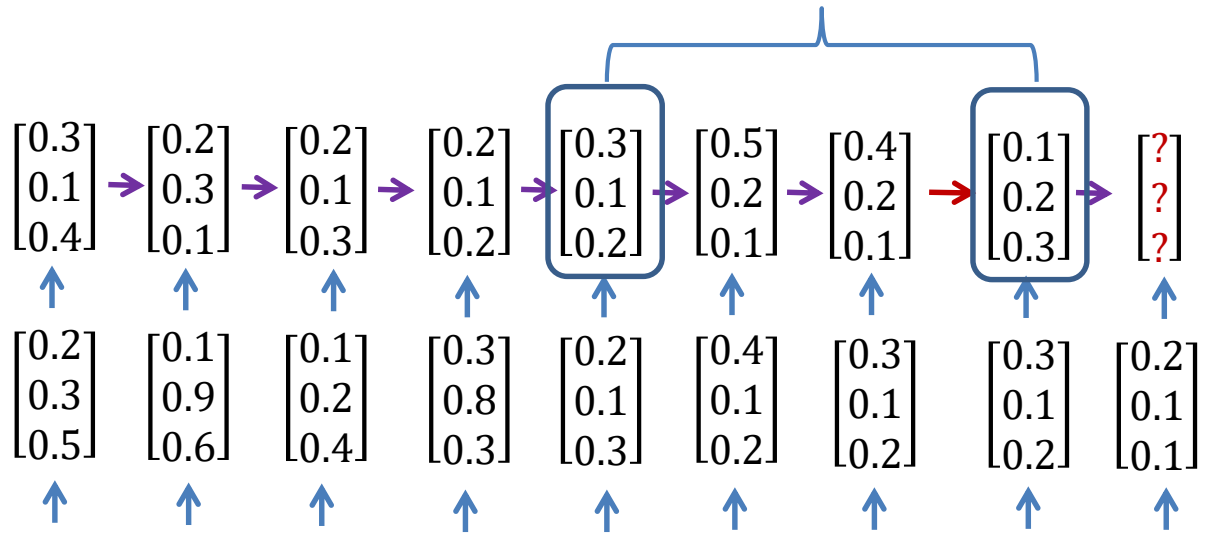
神经机器翻译-注意机制

$$score(h_s, h_t) = 1$$



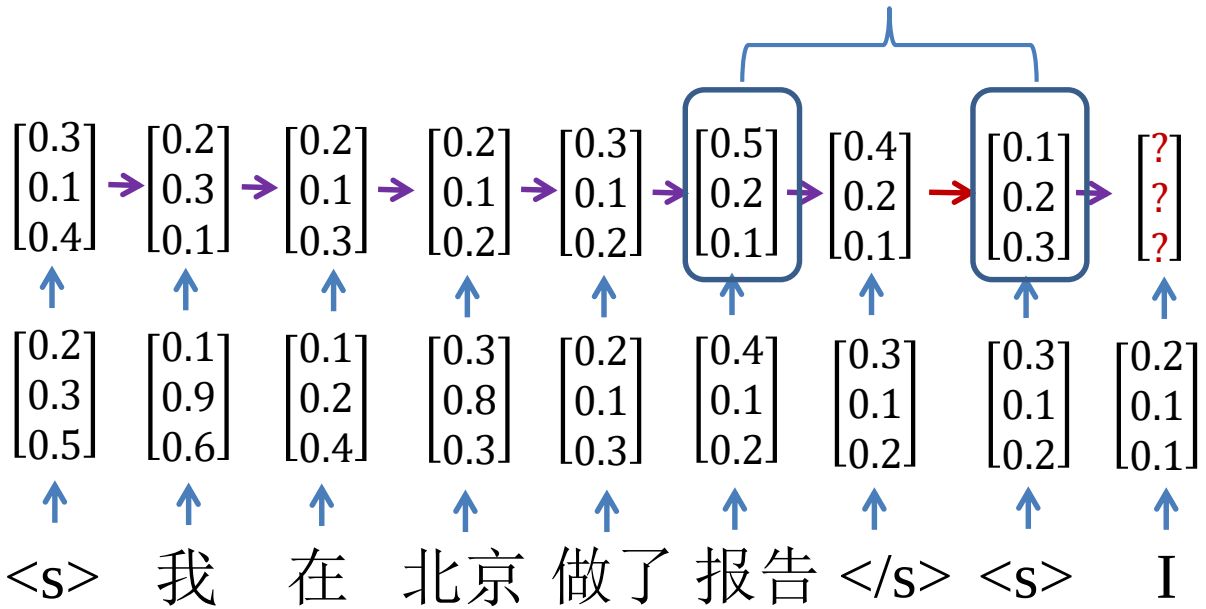
神经机器翻译-注意机制

$$score(h_s, h_t) = \mathbf{4}$$



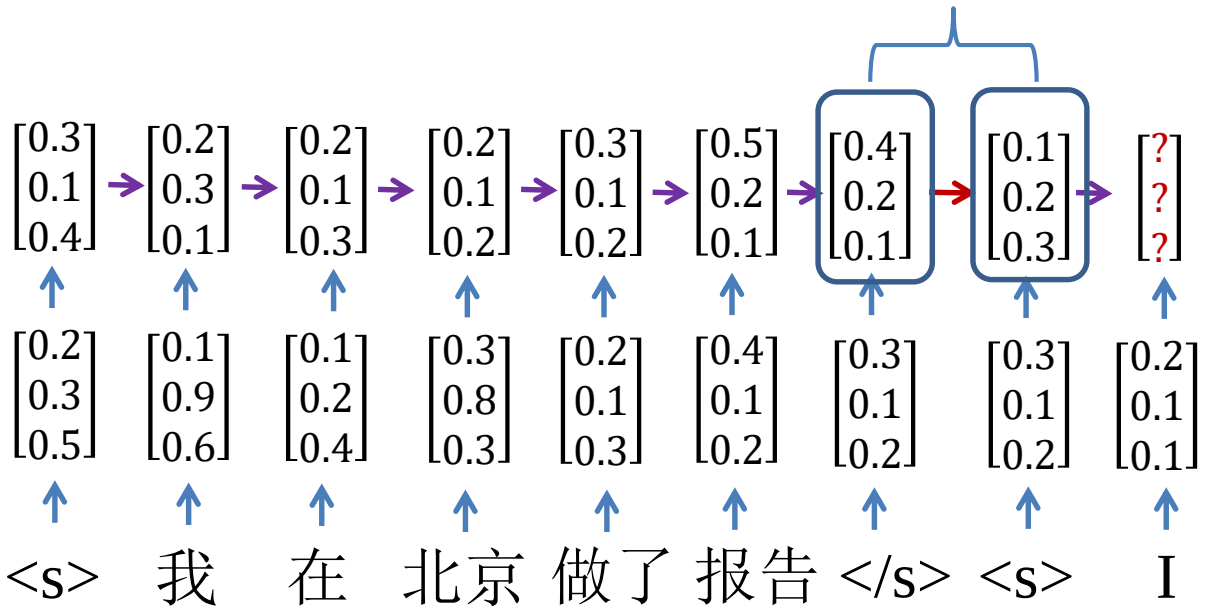
神经机器翻译-注意机制

$$score(h_s, h_t) = 2$$



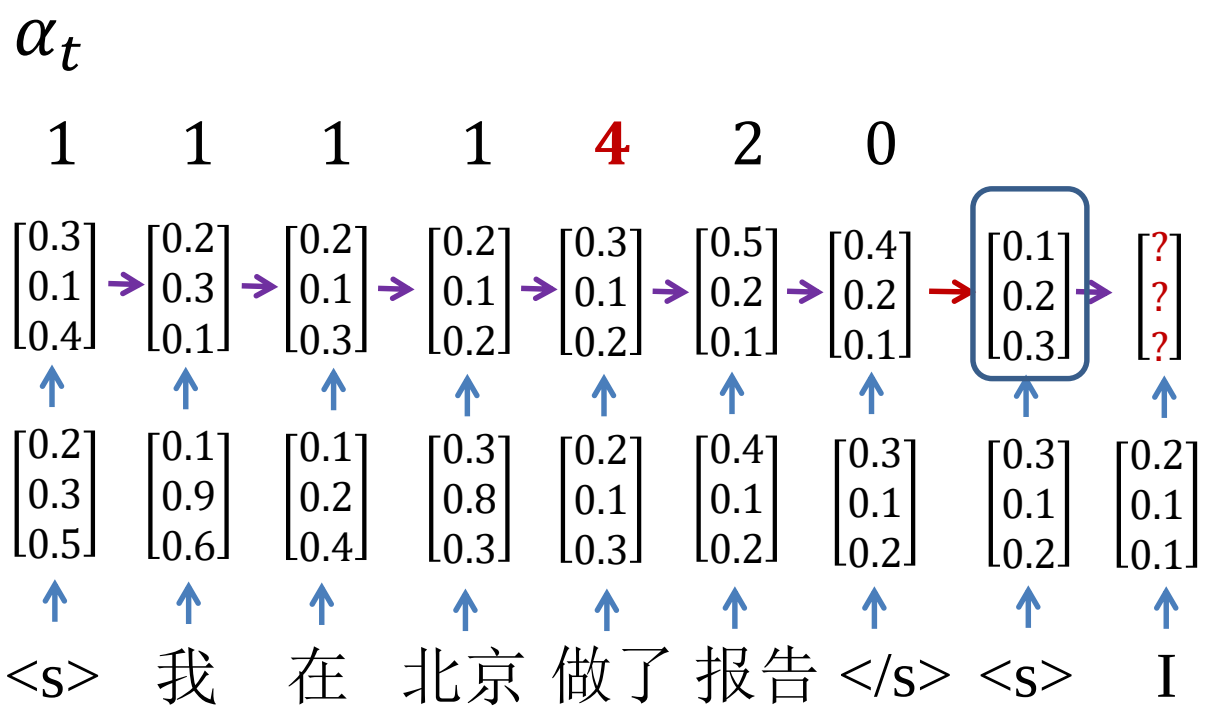
神经机器翻译-注意机制

$$score(h_s, h_t) = 0$$

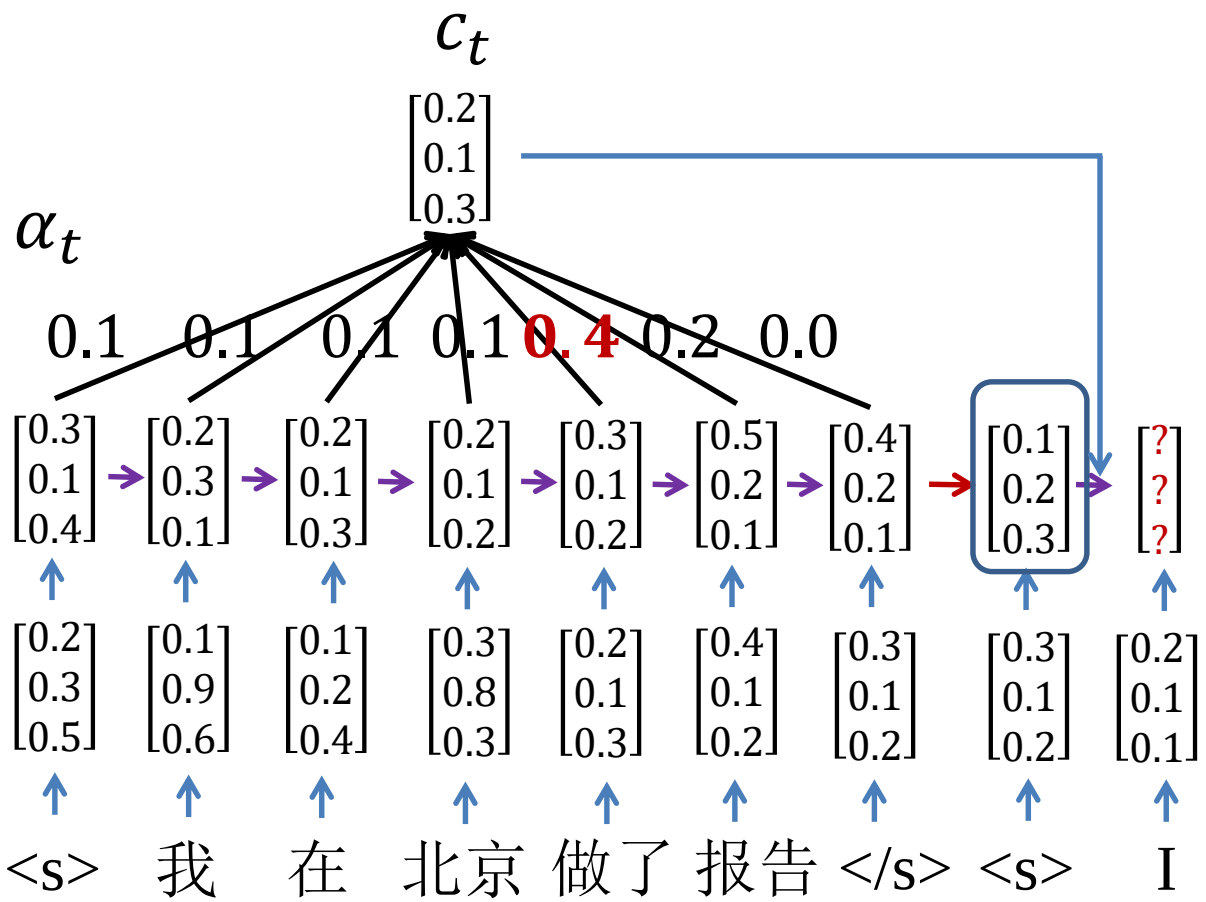


神经机器翻译-注意机制

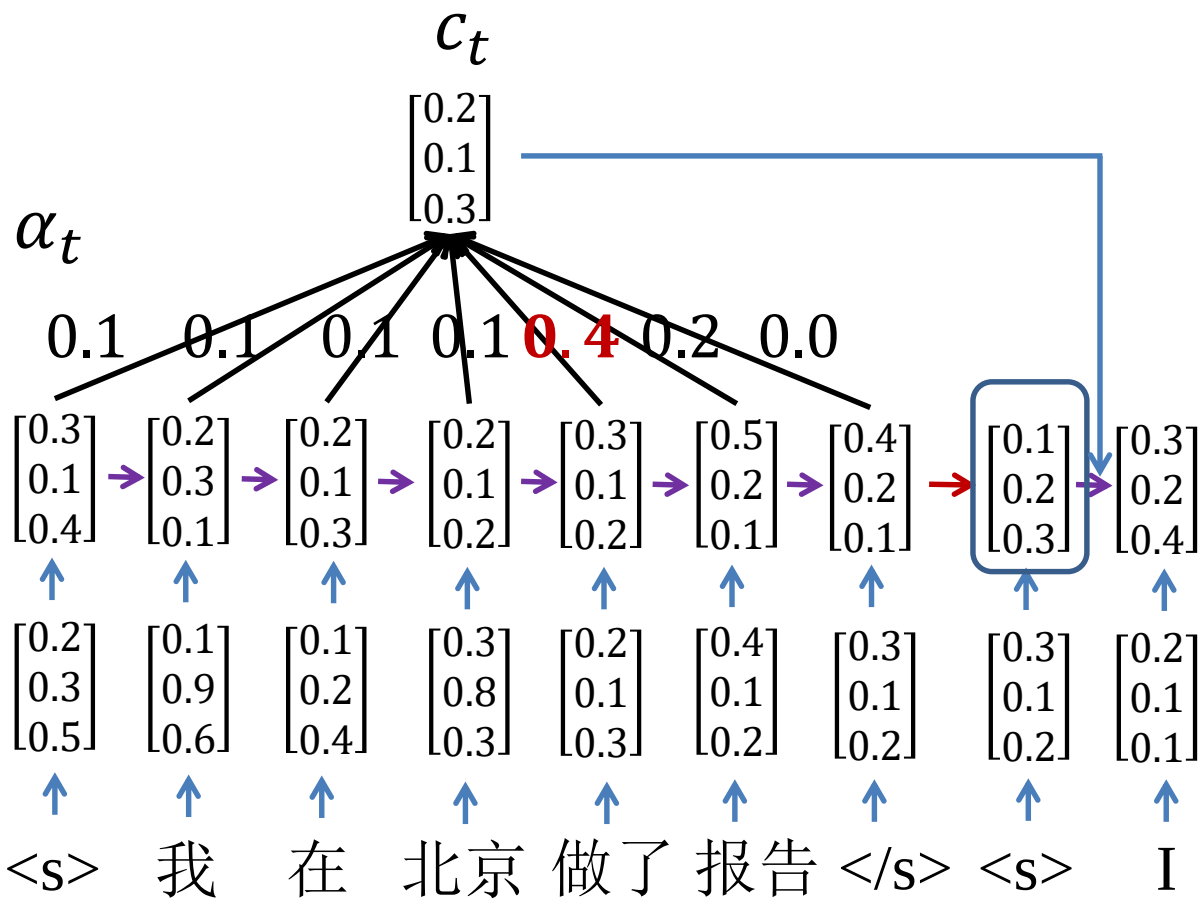
$$score(h_s, h_t)$$



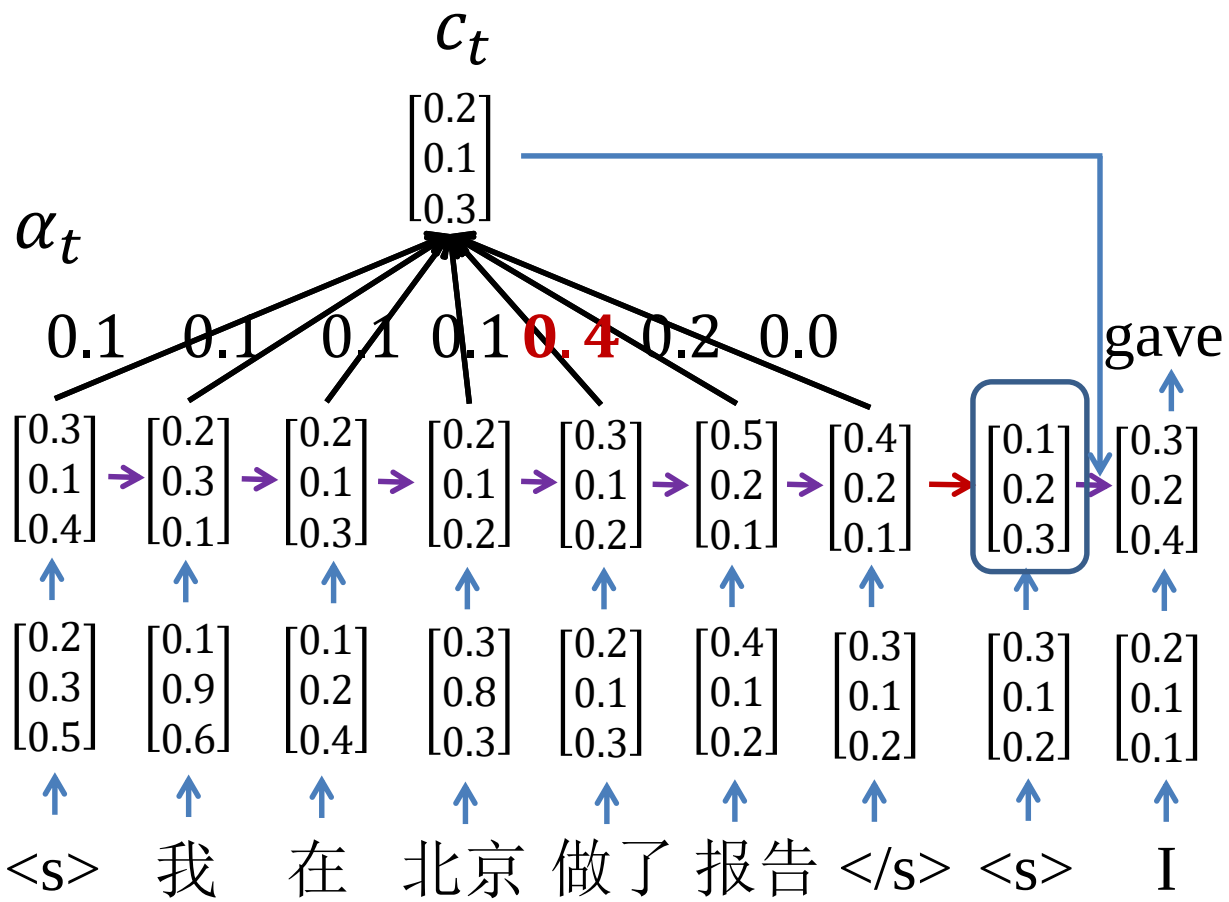
神经机器翻译-注意机制



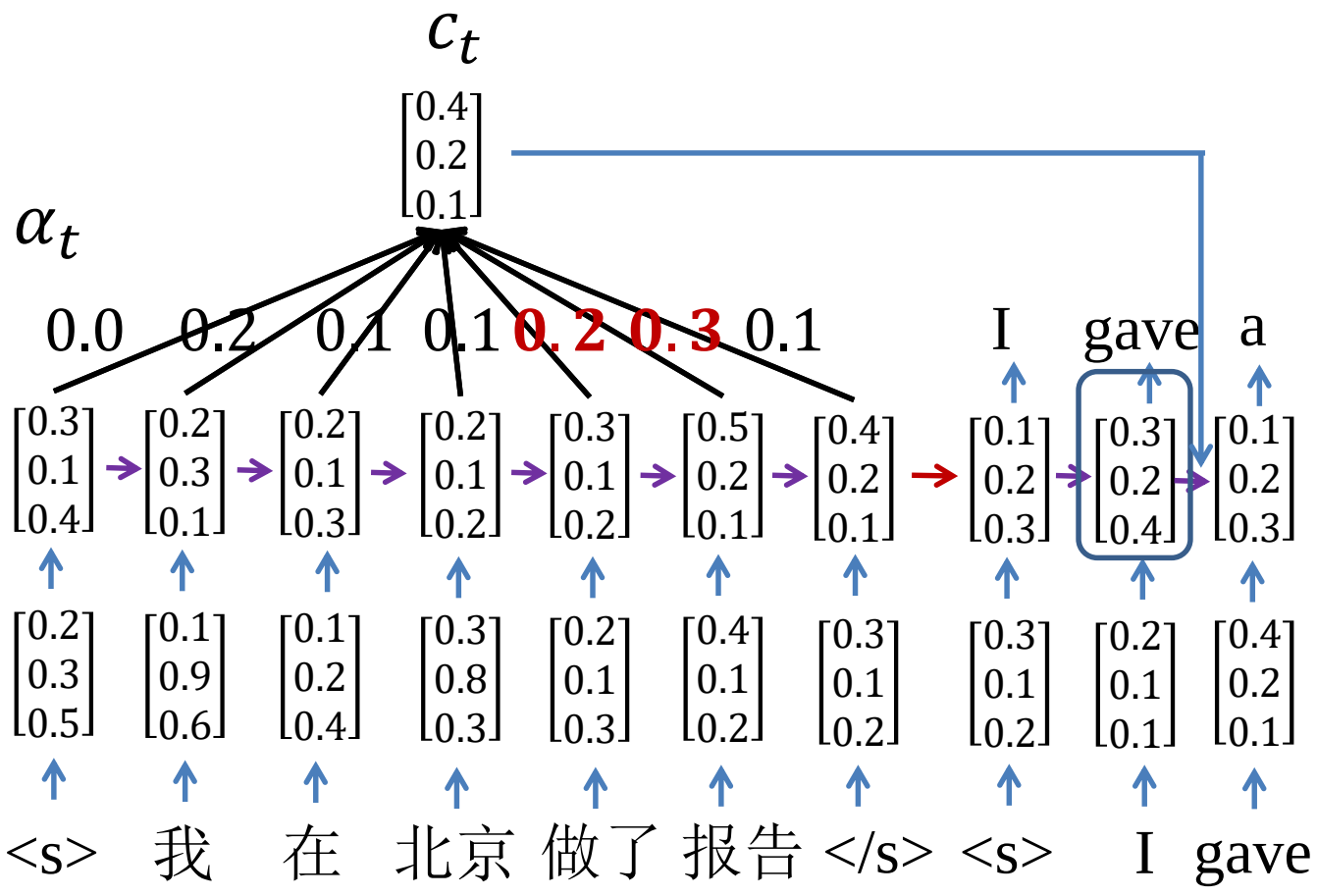
神经机器翻译-注意机制



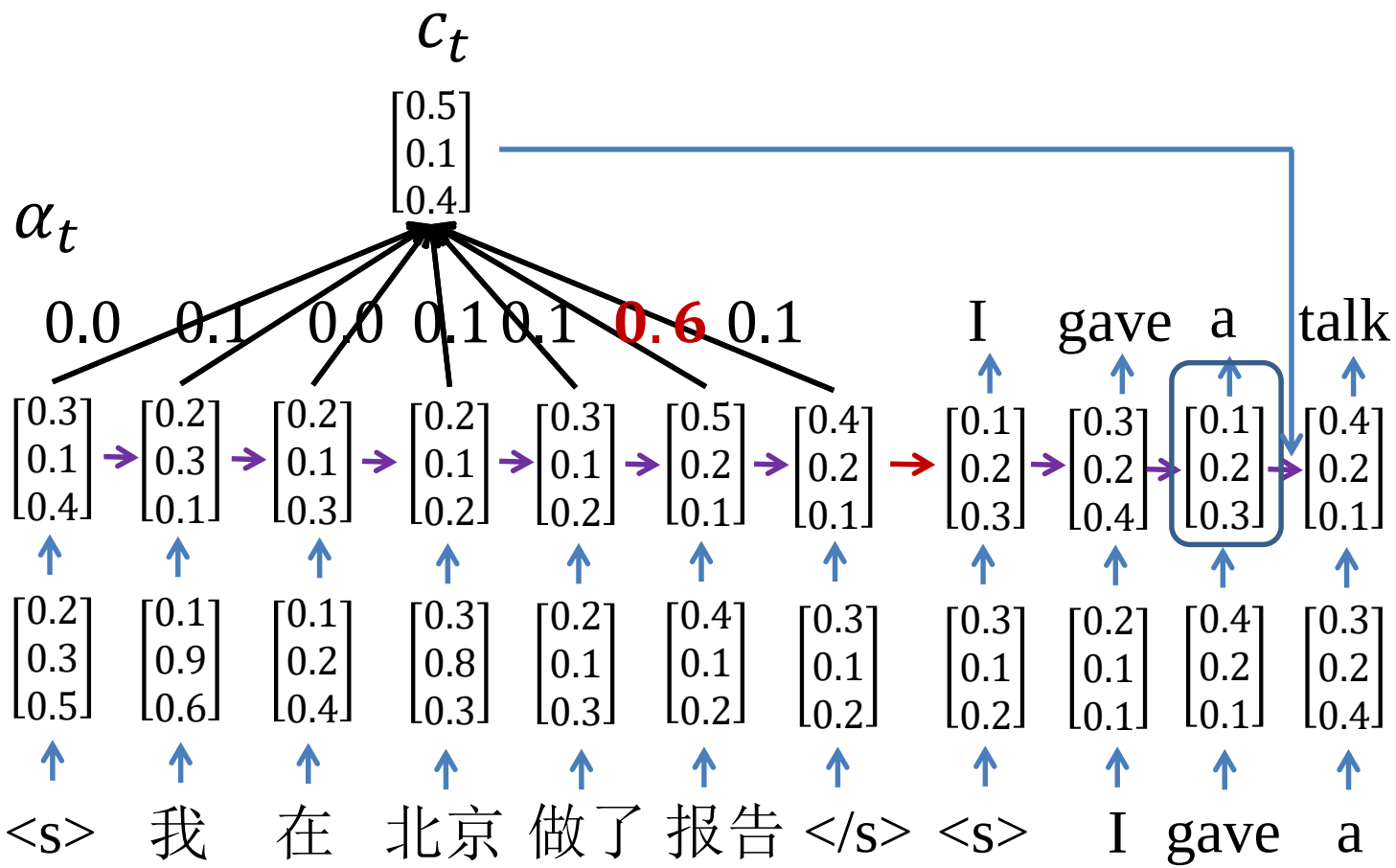
神经机器翻译-注意机制



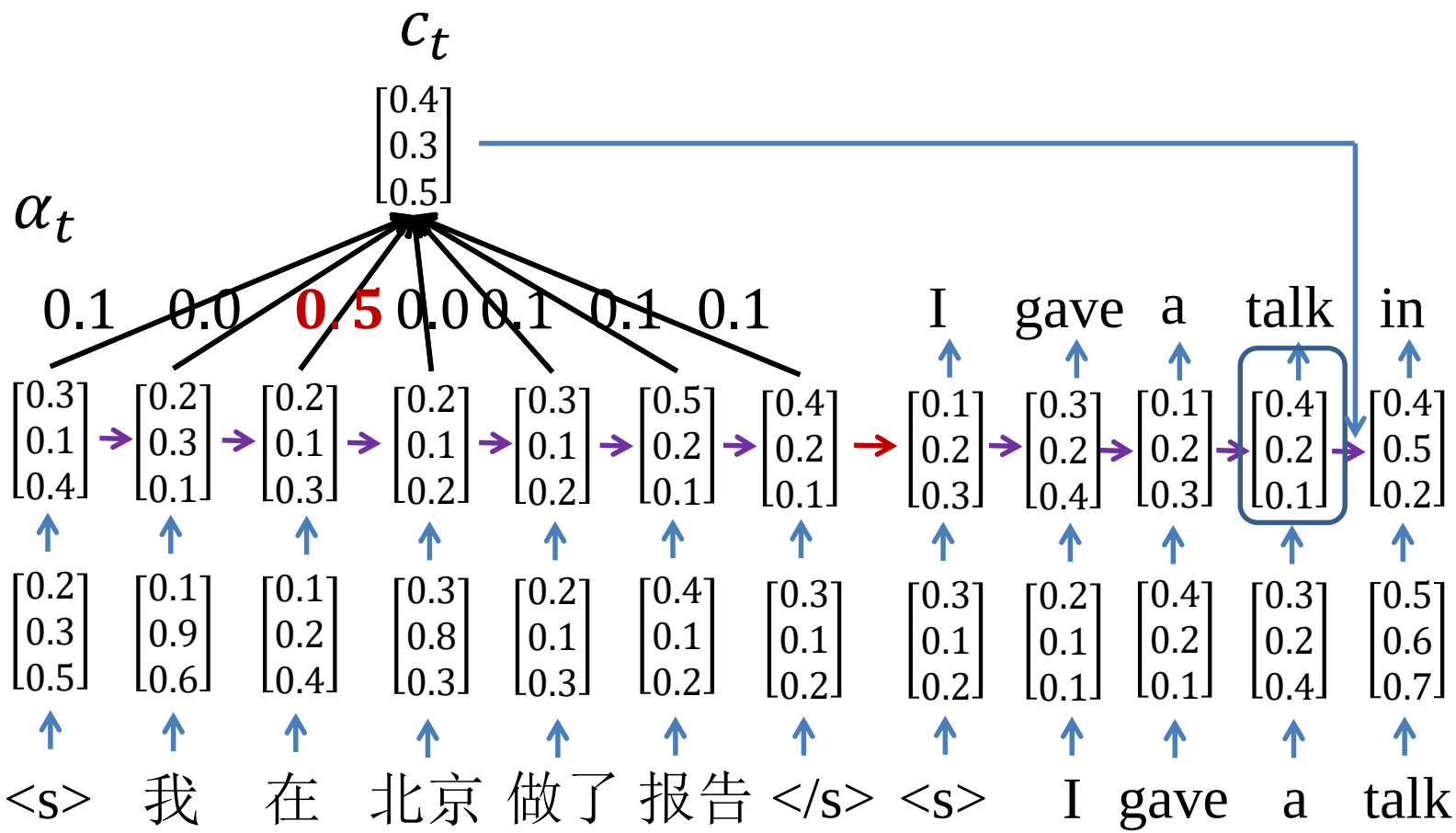
神经机器翻译-注意机制



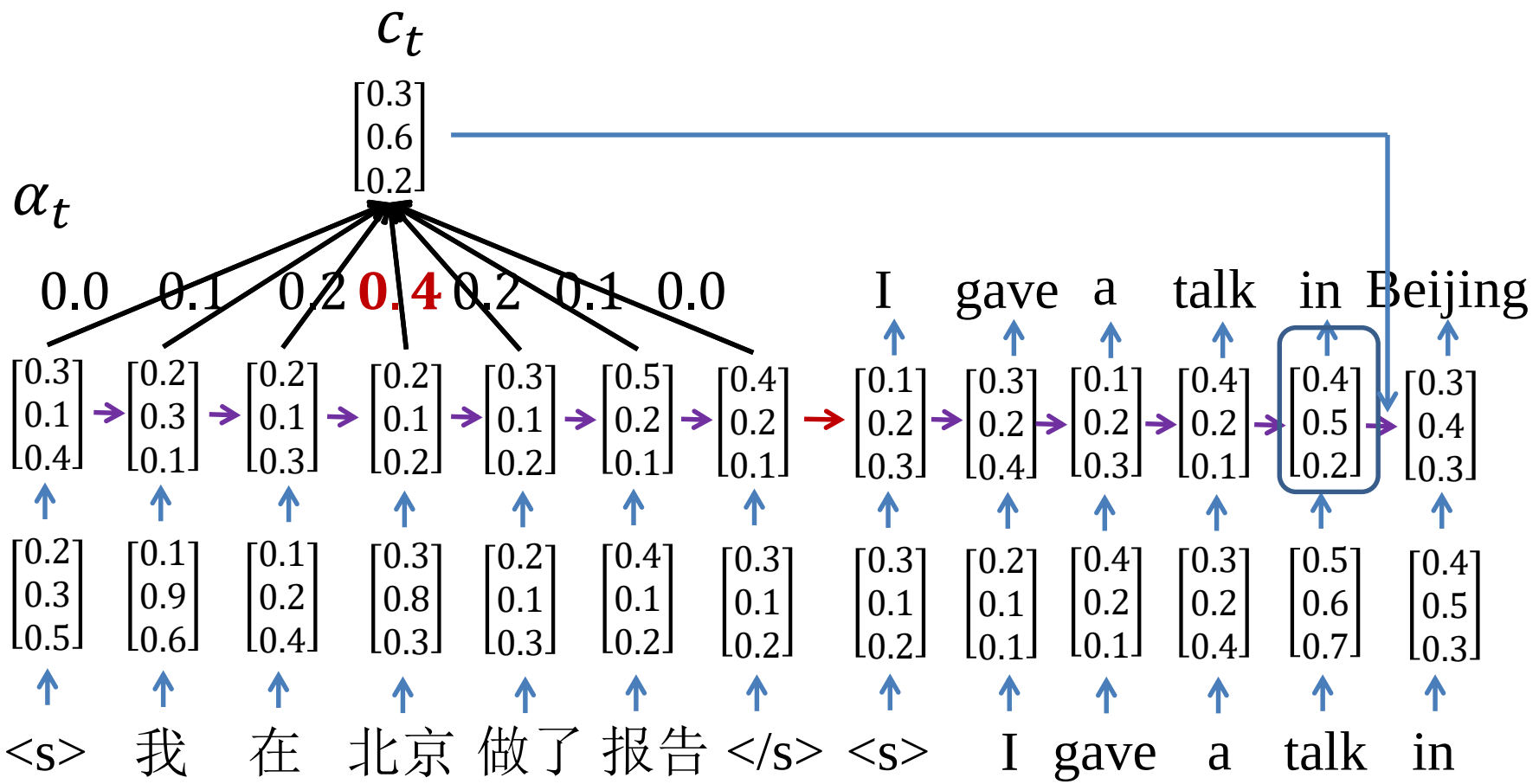
神经机器翻译-注意机制



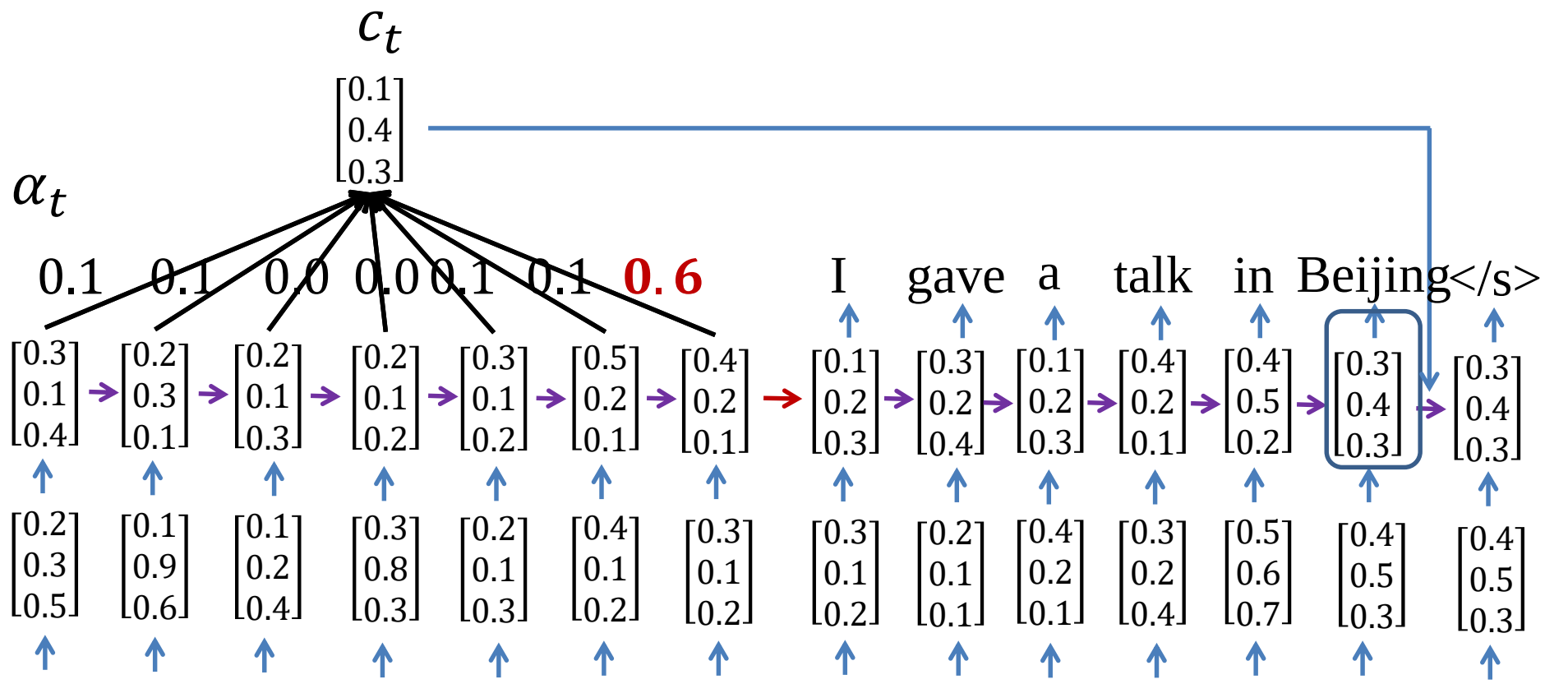
神经机器翻译-注意机制


c_t
α_t
0.1 0.0 0.5 0.0 0.1 0.1 0.1
I gave a talk in
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk

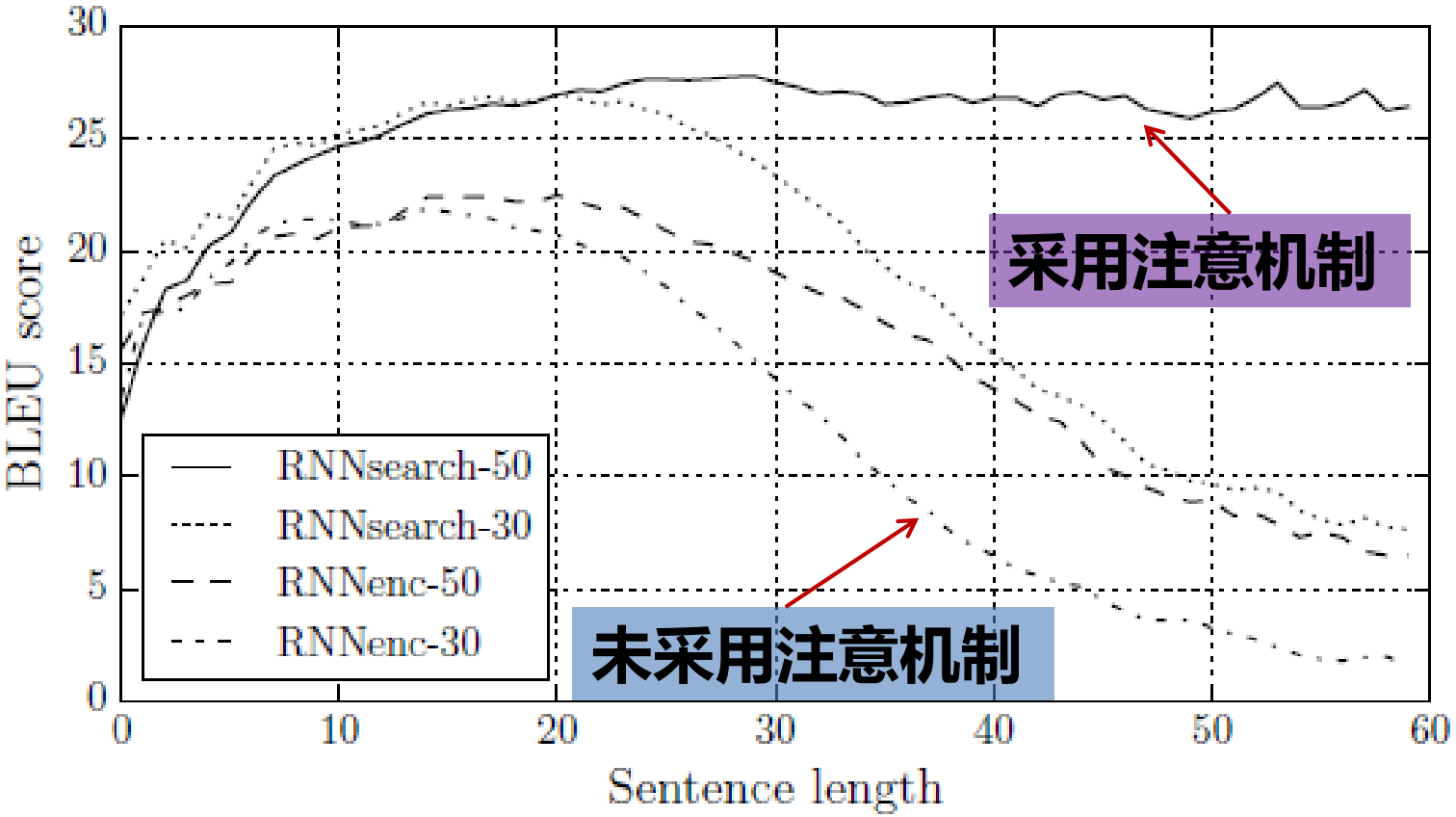
神经机器翻译-注意机制


c_t
α_t
0.0 0.1 0.2 0.4 0.2 0.1 0.0
I gave a talk in Beijing
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk in

神经机器翻译-注意机制

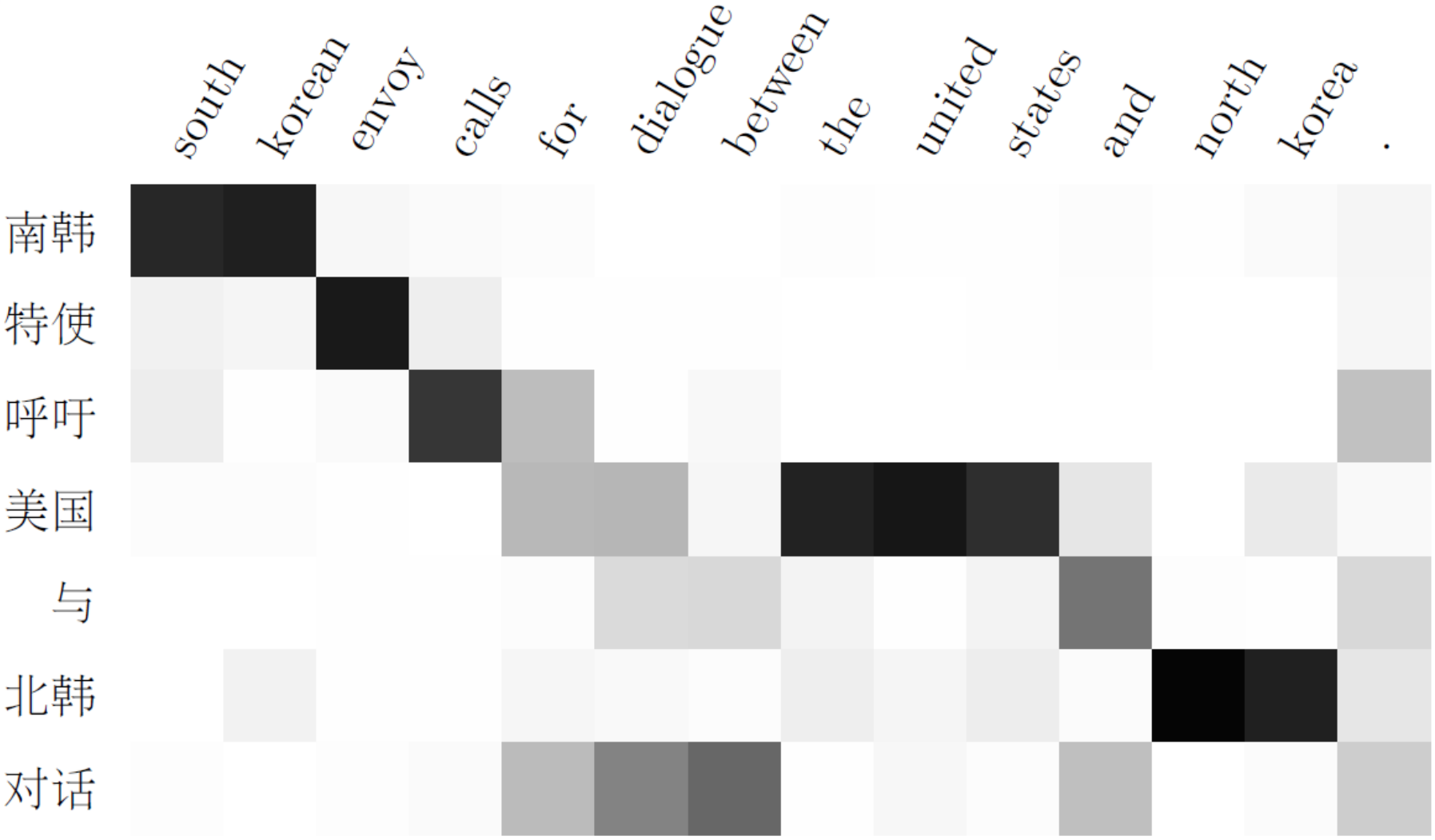

c_t
α_t
0.1 0.1 0.0 0.0 0.1 0.1 0.6
I gave a talk in Beijing</s>
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk in Beijing

神经机器翻译-注意机制



RNNenc: 无注意机制，**RNNsearch**: 采用注意机制

翻译实例


south
korean
envoy
calls
for
dialogue
between
the
united
states
and
north
korea
.
南韩
特使
呼吁
美国
与
北韩
对话

工业界研究机构

- 国外：
  - Google
  - Microsoft
  - IBM
  - Facebook
  - …
- 国内：
  - 百度
  - 华为
  - 阿里巴巴
  - 腾讯
  - 搜狗
  - 有道
  - …

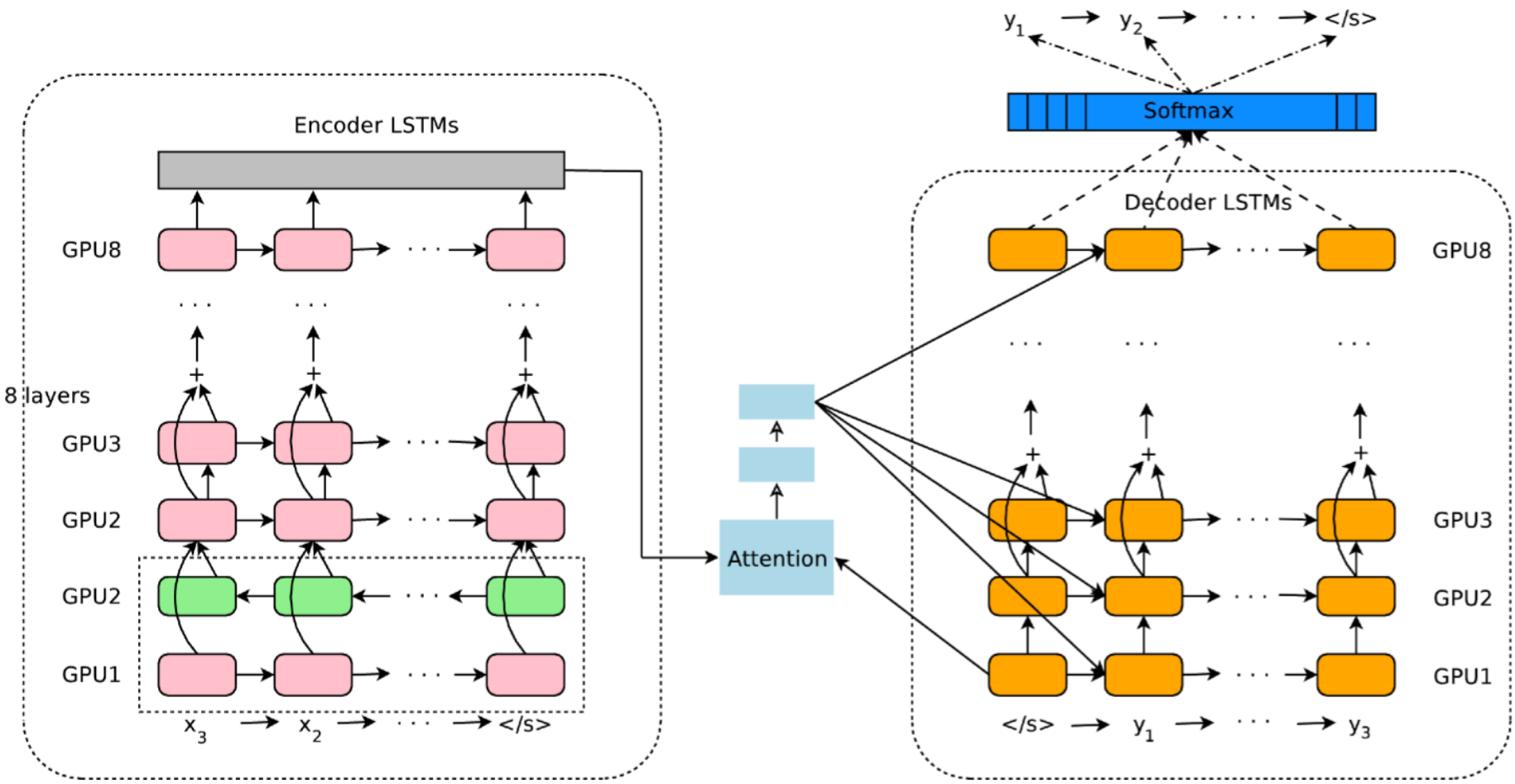
工业界线上产品


Bai du 翻译
检测到英语
中文
翻 译
State Councilor Yang Jiechi, speaking at a reception to celebrate the 45th anniversary of the epoch-making event, said China will continue to work with the UN and all other countries to advance peace and development for mankind.
2015年上线，系统描述未知
国务委员杨洁篪，在一个酒会庆祝的划时代事件第四十五周年，表示中国将继续与联合国和其他国家努力推进人类和平与发展。
拼音
双语对照

工业界线上产品


Google
Translate
2016年9月上线，有详细系统描述
Chinese
English
Spanish
Detect language
State Councilor Yang Jiechi, speaking at a reception to celebrate the 45th anniversary of the epoch-making event, said China will continue to work with the UN and all other countries to advance peace and development for mankind.
English
Chinese (Simplified)
Spanish
Translate
国务委员杨洁chi在庆祝成立45周年的招待会上说，中国将继续与联合国和所有其他国家合作，推动人类的和平与发展。
Suggest an edit

工业界线上产品



GNMT：谷歌神经翻译系统

工业界线上产品

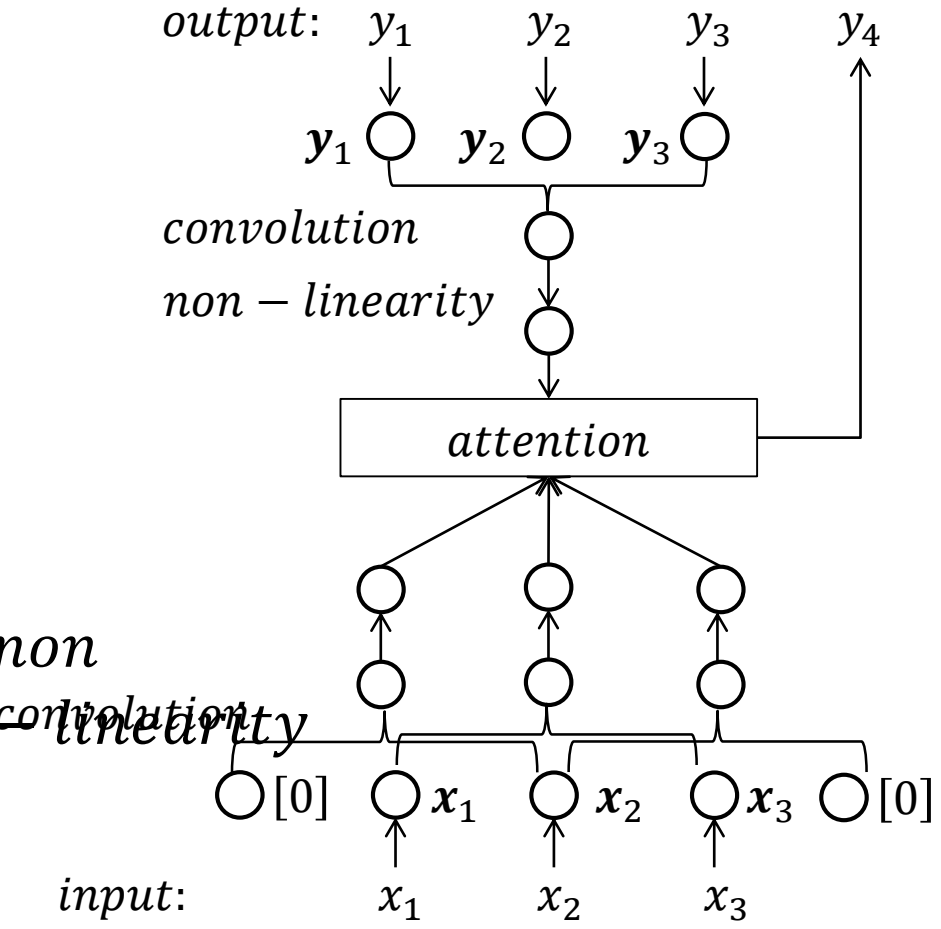
	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.550	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.035	4.594	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

人工评测提升显著！

GNMT：谷歌神经翻译系统

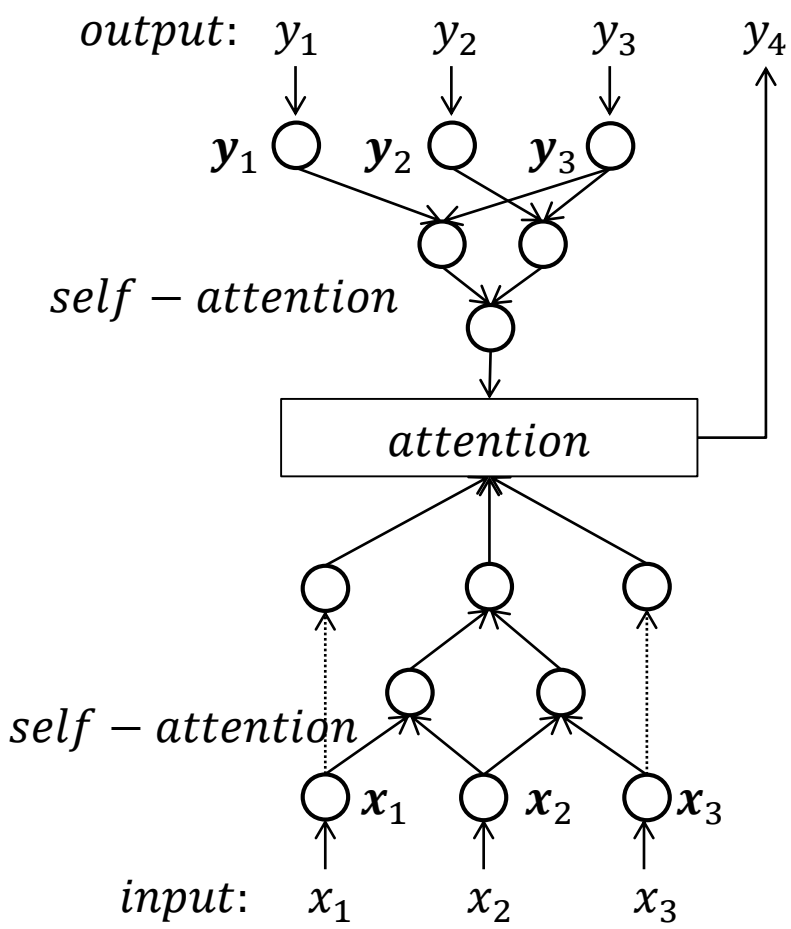
[Wu et al, 2016]

模型变革



(a)基于卷积神经网络的翻译模型

CNMT：Facebook神经翻译系统



(b)基于纯注意机制的翻译模型

Transformer：Google第二代

开源工具

1. TensorFlow (Transformer): 谷歌，python, C++/GPU
2. ConvolutionalNMT: Facebook，Torch/GPU
3. OpenNMT: Systran+哈佛，Torch/GPU
4. GroundHog: 加拿大蒙特利尔大学，python/GPU
5. dl4mt: 美国纽约大学，python/GPU
6. Paddle: 百度，C++/GPU
7. Zoph_RNN: 美国南加州大学，C++/GPU
8. EUREKA-MangoNMT: 中科院自动化所，C++/CPU
9. Nematus: 爱丁堡大学，C++/GPU

……

机器翻译技术落地

- 在线翻译（谷歌、微软、百度、有道、搜狗等）
- 翻译机（科大讯飞、准儿、百度、搜狗等）
- 同传机器翻译（微软、讯飞、腾讯、搜狗等）
  - 基于PowerPoint的语音同传（微软，TAUS 3.22-23 ）
  - 面向自由说话人的语音同传（腾讯，博鳌亚洲论坛 4.8-11）

未来展望

- 神经机器翻译采用编码解码网络，简单有效，已逐渐取代统计机器翻译，成为主流研究范式
- 神经机器翻译仍面临诸多问题
  - 缺乏可解释性
  - 难利用先验知识、语言相关知识
  - 训练、测试复杂度高（需GPU、甚至TPU）
  - 领域、场景迁移性能差

未来展望

- 未来发展
  - 神经机器翻译的可解释性研究
  - 与专家知识、常识知识的融合研究
  - 场景、领域的迁移和定制化研究
  - 面向资源稀缺语言的机器翻译建模
  - 多模态机器翻译（语音和文本的一体化）研究
  - 与硬件的一体化研究

译文评估方法

译文评估方法

◆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。

流畅性	
5	完美的英语表达 (Flawless English)
4	较好的英语表达 (Good English)
3	非母语的英语表达 (Non-native English)
2	不流畅的英语表达 (Disfluent English)
1	无法理解的英语表达 (Incomprehensible)

译文评估方法

充分性	
5	全部信息都 已充分表达了出来 (All information)
4	绝大部分信息已经表达了出来 (Most information)
3	很多信息被表达了出来 (Much information)
2	表达了少量信息 (Little information)
1	没有表达任何信息 (None)

* 2004年日本ATR组织口语翻译评测(IWSLT)评测时使用了“流畅性”和“充分性”这两个标准。

译文评估方法

语义保持性 (meaning maintenance)	
0	意思完全相反 (Total different meaning)
1	部分语义相同，但引入了误导信息 (Partially the same meaning but misleading information is introduced)
2	部分语义相同，没有引入新的信息 (Partially the same meaning and no new information)
3	意思几乎相同 (Almost the same meaning)
4	意思完全相同 (Exactly the same meaning)

*** 2005年CMU组织IWSLT评测时除了使用“流畅性”和“充分性”以外，还使用了这一标准。**

译文评估方法

源语言句子：This boy is very lovely.

译文1：这个小孩很可爱。

译文2：这个桌子很可爱。

译文3：这个小孩不可爱。

译文评估方法

客观评测

(1) **句子错误率**：**译文与参考答案不完全相同的句子为错误句子。错误句子占全部译文的比率。**

(2) **单词错误率(Multiple Word Error Rate on Multiple Reference, 记作 mWER)**：分别计算译文与每个参考译文的编辑距离，以最短的为评分依据，进行归一化处理

译文评估方法

(3)与位置无关的单词错误率 (Position independent mWER, **记作mPER)**：不考虑单词在句子中的顺序

(4) METEOR 评测方法

对候选译文与参考译文进行词对齐，计算词汇完全匹配、词干匹配、同义词匹配等各种情况的准确率(*P*)、召回率(*R*)和*F*平均值

译文评估方法

$$F = \frac{10PR}{R+9P} \qquad Score = F \times (1 - Penalty)$$

$$Penalty = 0.5 \times \left(\frac{\#chunks}{\#unigrams\_matched} \right)^3$$

#chunks 指系统译文中所有被映射到参考译文中一元文法可能构成的语块的个数；

#unigrams_matched 为所有匹配的一元文法的个数。不区分大小写。

分数取值为0~1，0为译文最差，1为最好。

译文评估方法

(5) 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*n*元词数目，得到评测结果。

译文评估方法

例如：

- 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，该候选译文可以得到 7/7 的打分，但显然这种翻译结果几乎没有任何意义。

译文评估方法

修正的计算一元语法精确度的方法：针对某个待评测的系统译文句子，首先统计每个单词在所有参考译文中出现次数的最大值 ***Max_Ref_Count***，然后，统计该单词在系统译文中出现的总次数 ***Count***，取***Count*** 和 ***Max_Ref_Count*** 两者中小的一个，即

$$Count_{clip}=\min(Count, Max\_Ref\_Count)$$

这样保证了每个系统译文中的单词计数不会超过该词在某个参考译文中出现次数的最大值。

译文评估方法

把系统译文中所有单词的$Count_{clip}$值累加起来，得到$Total_Count_{clip}$，即待评测的系统译文中出现在参考译文中的单词个数，
最后，用$Total_Count_{clip}$ 除以系统译文中全部单词的个数。

在上面的例子中，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。

译文评估方法

候选译文：It_1 is_2 a_3 $guide_4$ to_5 $action_6$ $which_7$ $ensures_8$ $that_9$ the_{10} $military_{11}$ $always_{12}$ $obeys_{13}$ the_{14} $commands_{15}$ of_{16} the_{17} $party_{18}$.

参考译文1：It is a guide to action that ensures that the military will forever heed Party commands.

参考译文2：It is the guiding principle which guarantees the military forces always being under the command of the Party.

参考译文3：It is the practical guide for the army always to heed the directions of the party.

单词1~6, 8~11, 14~15, 17~18 均出现在参考译文1中，第7个单词出现在参考译文2中，第12、16个单词均出现在译文2和3中。只有第13个词没有出现在任何译文中。

因此，一元语法精确度为：17/18。

译文评估方法

如果考虑 bi-gram {(it is), (is to), (to insure) …}，那么，修正的2元语法的精度为：10/17。

对于含多个句子的测试文本，以句子*C*为单位分别计算 ***n*-gram** 的匹配情况，然后，累计所有翻译句子修正后的 ***n*-gram** 计数，及测试集***Candidates***的所有 ***n*-gram** 计数，二者相除，得到修正后的精确度记分：

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

译文评估方法

考虑到在修正的 *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。

译文评估方法

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

最大语法的阶数，实际取4。

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。

$w_n=1/N$

长度过短句子的惩罚因子

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

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。

情感分析

相关概念

- 情感分析研究观点挖掘、倾向性分析等
- 什么是观点挖掘与倾向性分析？
- 为什么需要观点挖掘与倾向性分析？

相关定义

- ***观点：*** 人们对事物的看法，具有明显的主观性，不同人对同一事物的看法存在差异
- ***倾向性***：观点中所包含的情感倾向性

- ***观点挖掘与倾向性分析：*** 从海量数据中挖掘观点信息，并分析观点信息的倾向性
  - 非结构化→结构化

*情感分析或观点挖掘(in Wikipedia) 是自然语言处理、计算语言学与文本挖掘中的一个研究领域。*它的目标在于确定一个说话者或作者对于相关话题的 情感、观点或态度。

例子

“我今年入手诺基亚5800，把玩不到24小时，目前感觉5800屏幕很好，操作也很方便，通话质量也不错，但是外形有些偏女性化，不适合男生。这些都是小问题，最主要的问题是电池不耐用，只能坚持一天，反正我觉得对不起这个价格。”

- 外形
- 电池

- 屏幕
- 操作
- 通话质量

为什么需要？

- 文本信息主要包含两类
  - 客观性事实(Facts)
  - 主观性观点(Opinions)
- 随着Web2.0的飞速发展以及Web3.0的兴起，互联网中出现大量的UGC数据，其中包含了大量的观点信息
  - 博客、微博、商品评论、论坛....
- 已有文本分析方法主要侧重于客观性文本内容(factual information)的分析和挖掘

有什么用？

- **企业对观点挖掘和倾向性分析的需求**
  - 自动发现用户情感与观点 (市场智能化)
  - 感知社会发展趋势
  - 获取商业机会
  - 在线名誉管理
  - 目标导向地广告
- **普通用户对观点挖掘和倾向性分析的需求**
  - 有助于购买产品
  - 有利于发现针对政治话题的观点
- **政府对观点挖掘和倾向性分析的需求**
  - 控制公众整体情绪
  - 检测公共事件

情感分析与人工智能

Minsky
“人工智能之父”

The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.

我认为，智能不光是 IQ，更重要的还要有情感

沈向洋
微软全球执行副总裁

李飞飞
斯坦福人工智能
实验室主任

情绪、情感，是人工智能未来的方向

情感分析是人工智能中的意识形态！

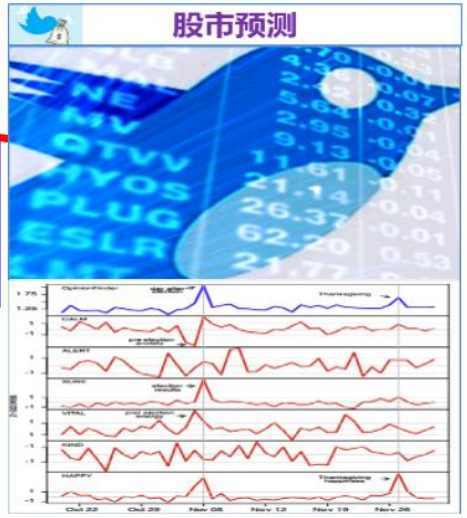
情感分析有广泛的应用场景和巨大的应用价值

社会舆情

电子商务

传统行业

金融领域



情感分析发展七项关键技术

- 情感分类
- 情感元素抽取
- 跨领域情感分析
- 个性化情感分析
- 隐式情感分析
- 情感原因发现
- 情感生成

情感分类

- 情感倾向性分类任务
  - 输入：句子/篇章
  - 输出：句子/篇章的褒贬极性及其强度

新闻哥吐槽 V

中国维和警察喊话吴京，《战狼2》再一次点燃了国人的爱国情怀。满满正能量！战狼2票房刷新国内纪录 电影热台词...

- 情绪分类任务
  - 输入：句子/篇章
  - 输出：句子/篇章情绪类别（如：喜怒悲恐惊等）

A-蕊儿啊 V

下午终于看了战狼2

真的是全程都感动 激动 满含泪水

情感分类

- 基于传统机器学习方法的情感分类
- 基于深度学习方法的情感分类
- 面向评价对象的情感分类

情感元素抽取

- 情感词表示学习
- 评价对象抽取
- 评价搭配抽取

情感元素抽取

- 情感词抽取/表示/词典
  - 褒义：好，贬义：坏
- 评价对象抽取
  - 评价的实体
- 评价搭配抽取
  - <评价对象，情感词>

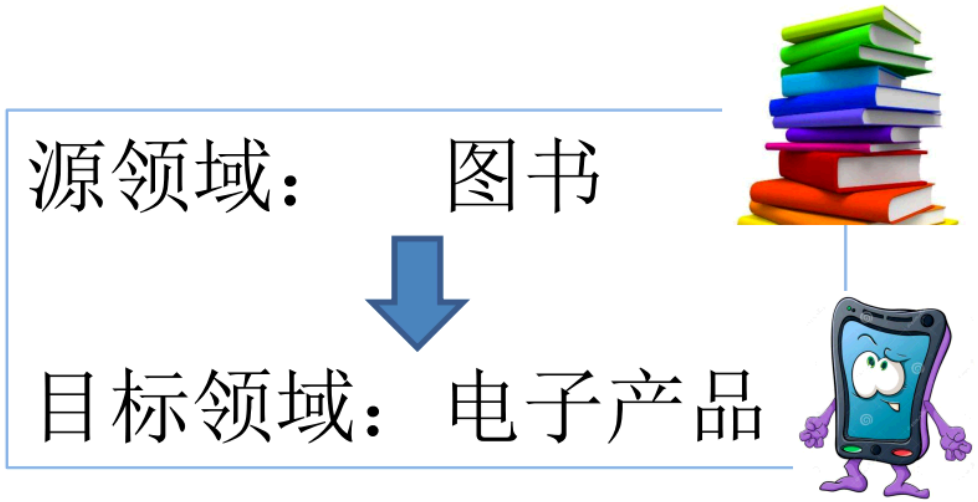
Absolute_Zer0_

今天 00:08 来自 iPhone 6s

iphone x很不错，除了贵，买不起,新的i watch 可以买一个，跑步就不要带手机了，看联通有什么优惠力度

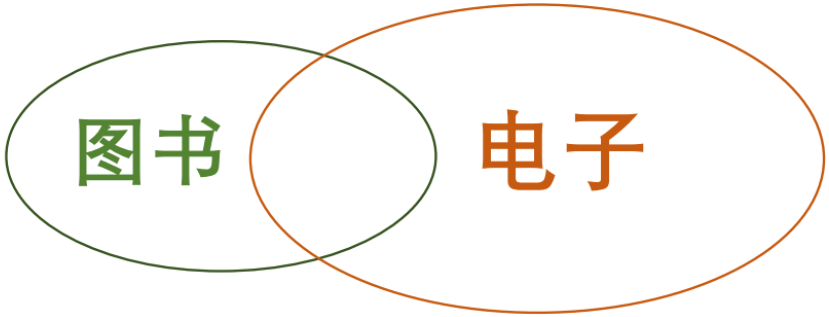
跨领域情感分析

- 从源领域到目标领域进行模型的迁移
- 目的
  - 标注少量或不标注目标领域的语料，利用源领域的语料在目标领域达到较好的性能



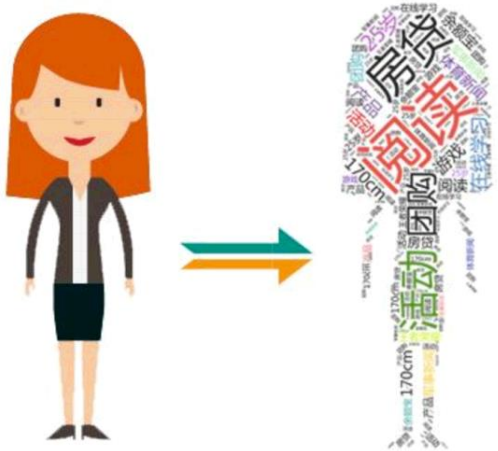
跨领域情感分析

- 情感分析任务的特点
  - 不同领域的评价对象不尽相同
  - 印刷、情节、 快递… V.S. 电池、音质、快递…
- 不同领域的评价表达千差万别
  - 精美、引人入胜、实惠… V.S. 耐用、高清、模糊、实惠…
- 不同领域中的同一情感表达的极性不同
  - 情节简单… V.S. 上手简单…



个性化情感分析

- 在情感分析中加入个性化的元素
- 情感分析的展示变得独特、另类、拥有自己特质的需要，独具一格

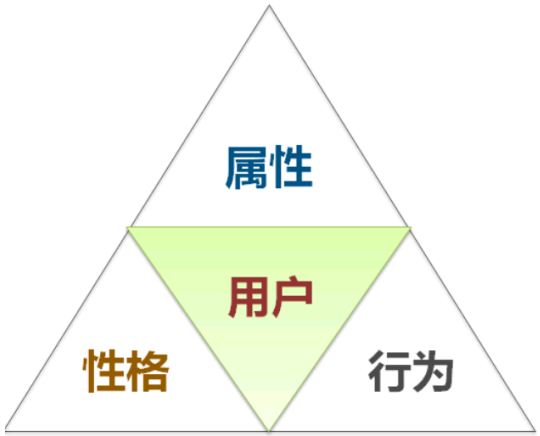


个性化情感分析

- 基于用户用词习惯的方法
  - 不同用户和群体情感倾向具有差异性
    - 由于用户群体立场存在差异，不同的用户群体往往对同一话题的情感倾向不同
    - 不同用户群体表达相同情感时，用词风格不尽相同
- 基于认知理论的方法
- 基于网络结构的方法

基于认知理论的方法

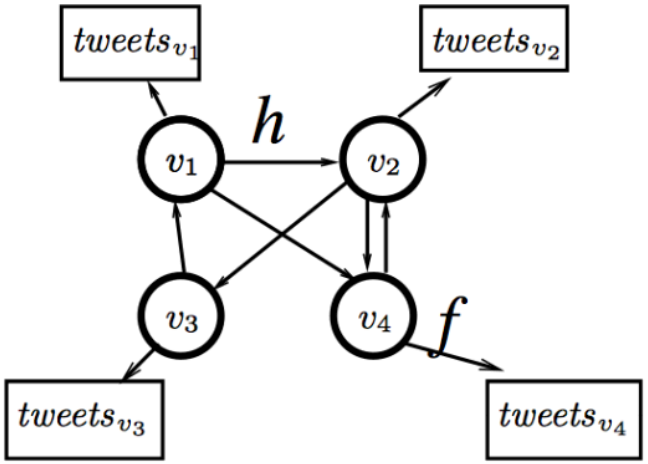
- 用户画像
  - 属性维度：自然欣喜
  - 性格维度：大五人格
  - 行为维度：用户偏好



- 结合用户信息进行更深入的情感分析与展示
  - 不同的用户群往往对同一话题的情感倾向不同
  - 用户群可按性别、年龄、职业等进行区分

基于网络结构的方法

- 传统的情感分类算法仅关注文本（句子、段落）特征
  - 单条文本的情感可能存在歧义
- 社交网络上用户之间的连接关系（关注、赞同、@等），这种连接关系表征了相同的情感倾向性
  - 在用户级别进行情感分析



*h*表示用户和用户之间的关系
*f*表示用户和文本之间的关系

隐式情感分析

- 社会媒体中文本情感表达方式复杂
  - 多数没有显式情感词
  - 多使用语言修辞表达或事实性陈述

隐式情感分析

- 事实型隐式情感分析
- 修辞型隐式情感分析

事实型隐式情感分析

- 事实性描述
  - 贬义描述

小子媛的爸

@如家酒店集团 @和颐酒店 看看你们的酒店多干净!一进房间上个房客的拖鞋，卫生纸放在桌上。桌子一层灰，而且总台客服电话统统打不通。投诉都不行!? 和颐酒店(...

如***叔

金牌会员

京东的速度超级的快！从下单到收到货，不到20小时。太令人惊叹了！小米六陶瓷板，运行内存6g，内存128g的手机，真是的一款高性价比的手机。这部手机，拿在手上看，着就像一件艺术品。小米这款手机真是不错。我喜欢5.7寸的手机，如果小米6这要是5.7寸的，那该有多好啊。小米六的启动速度，比前几款手机，明显要快了很多。前几天刚买了一部note2。小米六，屏幕要是再大一点儿，那就更完美了。在京东购物，感觉就是好。为京东赞！为小米的产品赞！

陶瓷黑 6GB 128GB 裸机 2017-08-12 08:51 举报 2

修辞型隐式情感分析

- 修辞手法
  - 反讽、 隐喻、 夸张、对比、排比

你还要我怎样qwq

#明日之子黑幕##明日之子黑幕#

一个真人秀，一个选秀节目，拿机器人和人比...你咋这么聪明呢！你咋这么厉害呢！！你咋这么棒呢！！！辣鸡！！！！

8月28日23:13 来自 红米2

拟人的拟

你就是一杯奶茶，你这么有气场干嘛

情感原因发现

- 基于文本的情感原因发现
- 基于个体立场的情感原因发现
- 基于群体立场的情感原因发现

- 情感原因通常是由个体、所在群体及内容共同作用产生的

情感生成

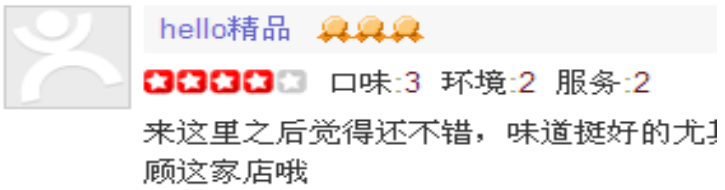
- 评论文本生成
- 情绪对话生成

观点挖掘与倾向性分析相关任务

- 观点及倾向性识别
  - 情感识别（Sentiment Identification）
- 观点要素抽取
  - 观点属性抽取（Opinion Attribute Extraction）
  - 观点摘要（Opinion Summarization）
- 观点检索

情感识别

- 观点识别 (subjective/Objective)
  - 中美两方的代表就朝鲜核问题进行了磋商。(Objective)
  - 中方发言人就美国近期对阿富汗的行动进行了强烈的谴责 (Subjective)
- 极性分类 (Positive/Negative/Neutral)
  - 这家餐厅总体来说还可以。(Neutral)
  - 但是价格偏贵，人均消费100块。(Negative)
  - 抛开价格的因素还是很不错的，值得推荐。(Positive)
- 强度识别（情感强度识别）
  - iPhone X的价格太贵了，两个肾都没了。(强烈反对)
  - iPhone X的价格有点贵。(有点差)



情感识别

- 词级别
  - 识别一个词的倾向性
- 特征级别(Aspect Level)
  - 识别一个Aspect的倾向性
    - “这家餐厅价格偏贵，人均消费100块”→ 价格
- 句子级别
  - 识别一个句子的观点倾向性
- 文档级别
  - 识别一篇文本（包含多个句子）整体的倾向性

观点属性抽取

- 观点持有者抽取
  - “中方发言人”就美国近期对阿富汗的行动进行了强烈的谴责”
    - 在新闻语料中大量出现，通常为命名实体、名词性短语或者术语
    - 在商品评论文本中很少出现

- 观点目标抽取
  - “中方发言人就美国近期对阿富汗的行动进行了强烈的谴责”
  - “这款手机的屏幕太小，分辨率不足”
  - 术语、事件、实体等

观点摘要

"I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …"

观点摘要:

特征 1: Touch screen

Positive: 212

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

…

Negative: 6

- **The screen is easily scratched.**
- **I have a lot of difficulty in removing finger marks from the touch screen.**

…

特征 2: battery life

…

观点检索

- 根据用户的查询从文档中找出对于主题信息发表了观点的文档
  - 主题相关并且具有主观倾向性
  - 博客、微博、论坛……

华为 HUAWEI P10 全网通 4GB+64GB 钻雕金 移动联通电信4G手机 双卡双待

麒麟960芯片！wifi双天线设计！徕卡人像摄影！白条12期免息！华为更多优惠详情请见！

京东价 ¥3788.00 降价通知

好评度 96%

支持国产(95) 系统流畅(82) 照相不错(77) 反应快(67) 外观漂亮(62) 指纹识别(62) 金属机身(54) 通话质量好(51) 分辨率高(50) 功能齐全(49)

全部评价(2.2万+) 晒图(500) 追评(700+) 好评(2.1万+) 中评(500+) 差评(500+) 只看当前商品评价 推荐排序

jd_死胖子
金牌会员

外观很美系统流畅，同一个路由器p10比红米4下载要快一倍。安装软件特别快。亮屏3个小时了才用不到20%，虽然没玩游戏，但是这期间我在不停的下软件，导入旧手机数据看了一会贴吧，续航很强悍。一分钱一分货，红米白白了。（垃圾红米拍照真差）

C***e
金牌会员

手机买来快半个月了，特意用一段时间再来评论的，当初决定买这个手机就是图它电池容量，相对的屏幕大小，双卡双待还有质感。还有支持国产。首先手机屏幕和大小单手操作的话还是有点勉强，电池的话个人有点失望，勉强能维持一天时间，我每天电话比较多，其次系统，平时操作起来确实挺快的，没毛病，但有偶尔的卡机，这个试用体验真的很差，比较国产*起的手机也是有点贵

小军啊剩点花钱
金牌会员

第一，手机玩游戏发热，第二，这个电池太不耐用，正常打电话一天都用不上，就别说要游戏了！第三，刚用一天就升级，，第四，这手机信号也太差了吧，没信号！大家都看看！买了就后悔了！

曜石黑 64GB 2017-05-04 00:45 举报 182 129

Apple iPhone 7 Unlocked Phone 128 GB - US Version (Black)

4.0 ▾ 301 customer reviews | 763 answered questions

Available from these sellers.

Size: **128 GB**

4.5 CNET Editors' Rating

3.0 18 user reviews

The Good / Improved front and rear cameras -- now with optical image stabilization -- deliver much improved photos, especially in low light. Water resistant. A faster processor, plus slightly better battery life. More onboard storage than last year's models for the same price.

The Bad / No headphone jack (but there's a dongle and compatible wired headphones in the box). Click-free home button takes getting used to. Only the larger 7 Plus has the cool dual camera. Shiny jet-black version scratches easily.

The Bottom Line / The iPhone 7's notable camera, battery and water resistance improvements are worthwhile upgrades to a familiar phone design. But ask yourself if you really need an upgrade... and if the Plus might be a better choice.

8.7
OVERALL

Design	9.0
Features	8.0
Performance	9.0
Camera	9.0
Battery	7.0

TRACKING OPINIONS ON TWITTER

twitrratr

SEARCH

SEARCHED TERM	POSITIVE TWEETS	NEUTRAL TWEETS	NEGATIVE TWEETS	TOTAL TWEETS
iphone	2775	19720	846	23341

11.89% POSITIVE

- @schwa now there's a blast from the past. but it occurs to me that gliderpro would make a great iphone app. (view)
- alas fair iphone, you served me well and will be missed. (view)
- @mikediliberto @downtownrob @mitchwagner funny that i ended up following smoke signals as

84.49% NEUTRAL

- view from the iPhone: http://www.floodgap.com/iv/197 (view)
- That's "Memphis" Taproom. Goddamn iPhone. (view)
- @mothermusings This is the iPhone thingie, huh? Sooooo sorry! (view)

3.62% NEGATIVE

- @mikef1182 as bad as exchange on the iphone? (view)
- http://twitpic.com/i0se - iphone typing auto-correct changes 'just sayin' to 'just satin' - wrong msg indeed! (view)
- iphone applications don't whine about being left outside or going hungry or manual labor or using

主要内容

- 文本分类
- 文本聚类
- 情感分析
  - 相关概念
  - 典型方法
  - 问题与挑战

典型方法

- 情感识别
- 观点挖掘
- 观点检索

情感识别

- 词级别
- 句子级别
- 文档级别
- 其他

词级别情感识别

- 任务：
  - 识别词语的情感倾向性，构建词典资源
- 方法：
  - 基本思路：利用词之间的相似度进行扩展
  - 基于词典的方法
  - 基于语料库的方法

情感识别

- 词级别
- 句子级别
- 文档级别
- 其他

句子级情感识别

- 任务：识别句子的情感倾向性
  - “这部电影看得想吐，看了5分钟就看不下去了。”
- 关键问题
  - 如何进行特征表示
- 分类：
  - 基于语料库的方法
  - 基于词典的方法
  - 融合方法

与传统方法的区别

- 基于话题的文本分类
  - 侧重于主题词特征
    - “这款手机的屏幕太大了”(科技、手机)
- 情感识别
  - 表示倾向性的词语更加重要.
    - “这款手机的屏幕好大”(主观、褒义)

基于语料库的方法-特征选择

- 利用传统文本分类方法处理情感分类任务 (Pang EMNLP 2002)
  - 比较多种特征的效果
    - Unigram、bigram、POS、Adj.、Position
  - 比较多个分类器性能
    - SVM、Naïve Bayes、Maximum Entropy

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

情感识别

- 词级别
- 句子级别
- 文档级别
- 其他

文档级情感识别

- 任务：识别篇章整体观点倾向性

> 诺基亚5800屏幕很好，操作也很方便，通话质量也不错，但是外形偏女性化，而且电池不耐用，只能坚持一天，价格也偏贵，反正我觉得不值。

- 绝大多数方法与句子级别方法类似
  - 特征+分类器
- 关键问题
  - 多观点倾向性：一篇商品评论中可能包含对于商品多方面的观点，每个观点的倾向性也可能不同，如何识别篇章整体的观点倾向性
    - 按照句子划分
    - 按照主题划分

小结

- 篇章级观点倾向性识别仍然可以看做是一个文本分类任务
  - 如果仅仅是用词袋子模型，那么文档级别与句子级别在处理方法上没有区别
- 主要问题在多观点混合问题
  - 篇章中局部观点与整体观点不一致

情感识别

- 词级别
- 句子级别
- 文档级别
- 其他
  - 跨语言观点识别与分析
  - 领域适应性

典型方法

- 情感识别
- 观点挖掘
- 观点检索

观点对象抽取

- 任务：抽取观点评价的对象
  - 中方发言人就美国近期对阿富汗的行动进行了强烈的谴责。（新闻）
  - iphone7的屏幕简直太酷了！（商品评论）
    - 产品特征: 商品、商品属性、商品的部件、商品部件的属性 (Popescu EMNLP 2005)

Explicit Features	Examples	% Total
Properties	ScannerSize	7%
Parts	ScannerCover	52%
Features of Parts	BatteryLife	24%
Related Concepts	ScannerImage	9%
Related Concepts' Features	ScannerImageSize	8%

    - 不是所有的商品属性都是评价的对象
      - 诺基亚C1的屏幕尺寸有1.8寸。
      - iphone的价格太贵了

观点持有者抽取

- 基本思路(Kim AAAI 2005)
  - 命名实体识别
    - 人名、机构名
  - 句法结构特征
    - Convolution Kernel
  - 分类或者序列标注
    - SVM, Naïve Bayes, CRFs
  - 需要指代消解

典型方法

- 情感识别
- 观点挖掘
- 观点检索
- 资源和评测

观点检索

- 任务：
  - 从海量文本中根据查询找到观点信息
  - 根据主题相关度(topic relevance)与观点倾向性(opinion relevance)对结果进行重排序
    - 主题相关度：传统检索
    - 观点倾向性：观点识别
- 关键问题
  - 找到主题相关度得分与观点倾向性得分的折中

情感分析六大趋势

一、从粗粒度到细粒度

二、从单领域到跨领域

三、从文本到社会媒体

四、从显式情感到隐式情感

五、从情感分类到情感原因

六、从情感分析到情感生成