CIPS暑期学校

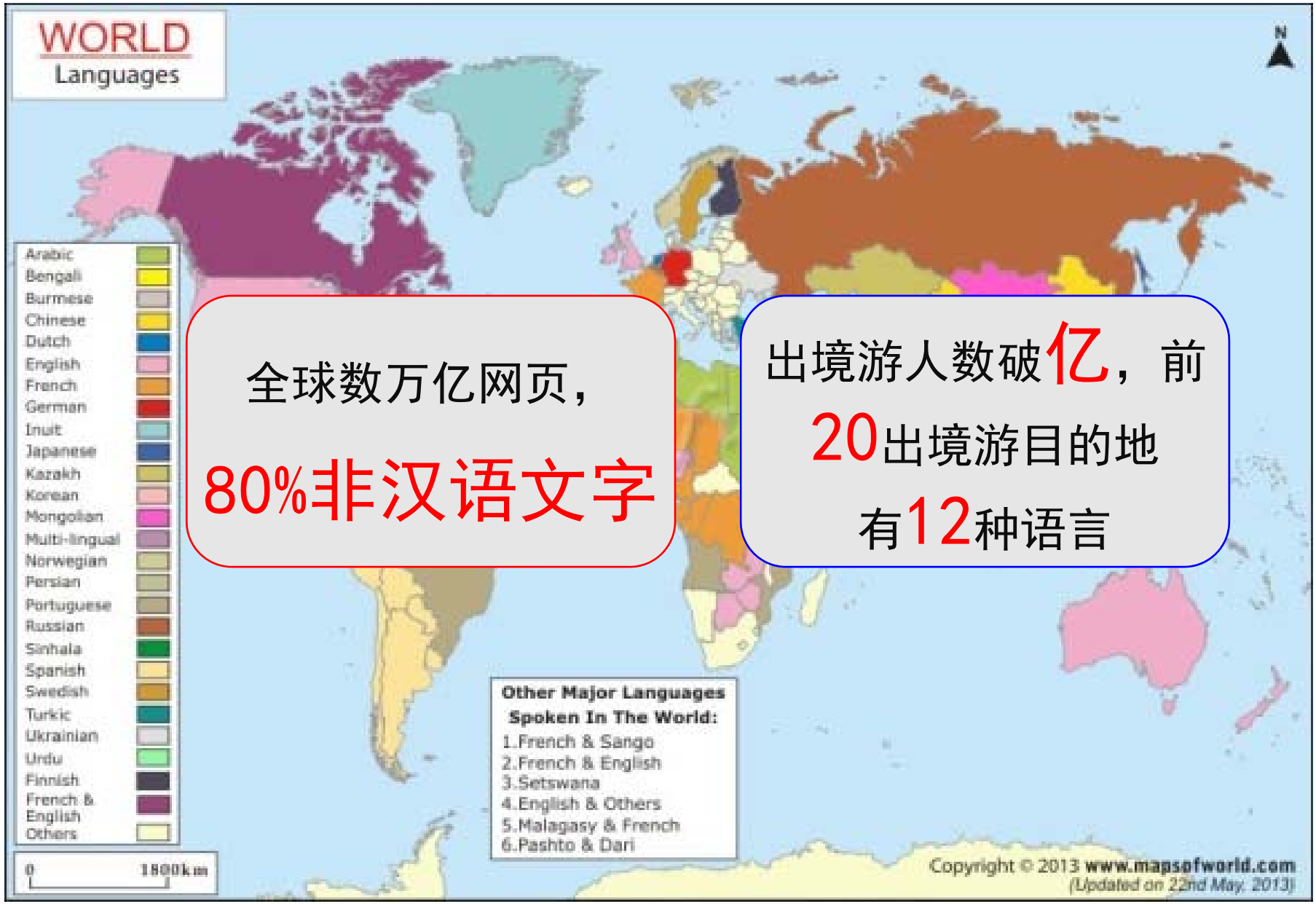
深度学习与机器翻译
(1/2)

张家俊
中国科学院自动化研究所
www.nlpr.ia.ac.cn/cip/jjzhang.htm
jjzhang@nlpr.ia.ac.cn

机器翻译

Tower of Babel (巴别塔)

世界语言地图



全球数万亿网页，

80%非汉语文字

出境游人数破亿，前20出境游目的地有12种语言

中国语言地图


Orogen
Evenki
Buriat
Daur
Mandarin
Manchu
Korean
Manchu
Korean
Manchu
Daur
Kazakh
Xibe
Mongolian
Kyrgyz
Kalmyk
Uygur
Mandarin
Wakhi
Yugur
Jinyu
Kalmyk
Tu
Salar
Amdo
Choni
Mandarin
Baima
Qiang
Panang
Tibetan
Jiarong
Wu
Huizhou
Sherpa
Lhomi
Groma
Luoba
Drung
Nuosu
Hmong
Miao
Xiang
Gan
Min-Bei
Min-Dong
Lisu
Naxi
Lipo
Miao
Pu-Xian
Nasu
Xishanba
Bai
Bouyei
Dong
Hakka
Lolopo
Min-Zhong
Axi
Tai-Nua
Nisu
Zhuang
Yue
(Cantonese)
Min-Nan
Riang
Parauk
Hani
Ping
Lu
Bugan
Vo-Wa
Lahu
Honi
Biyo
Min-Nan
Hlai

背景：一带一路

64个国家和地区

44亿人口

50多种语言

机器翻译

定义：机器翻译是利用计算机将一种自然语言（源语言）自动转换为另一种自然语言（目标语言）的技术。

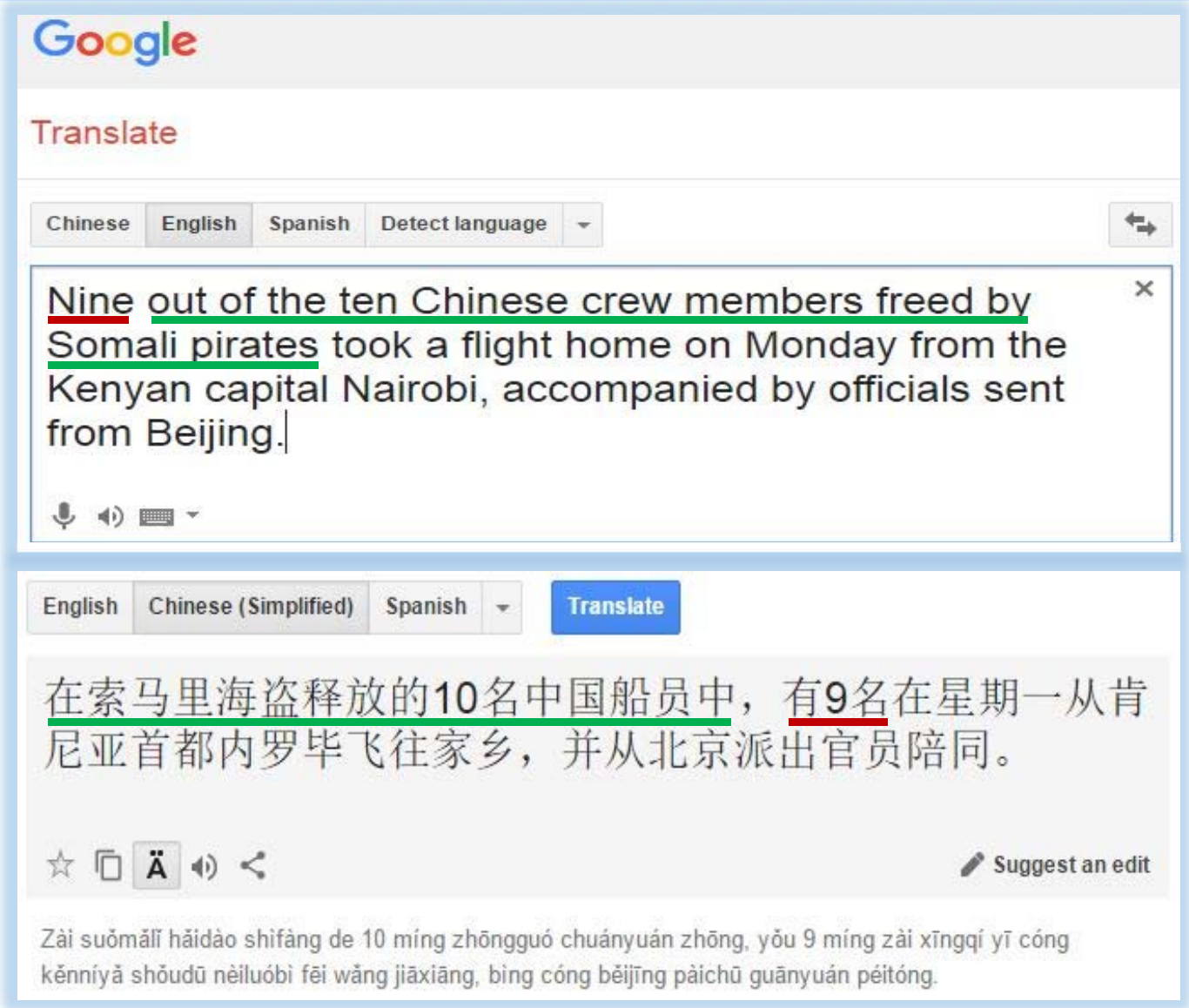
机器翻译

定义：机器翻译是利用计算机将一种自然语言（源语言）自动转换为另一种自然语言（目标语言）的技术。

很多人眼中的机器翻译

很多人眼中的机器翻译

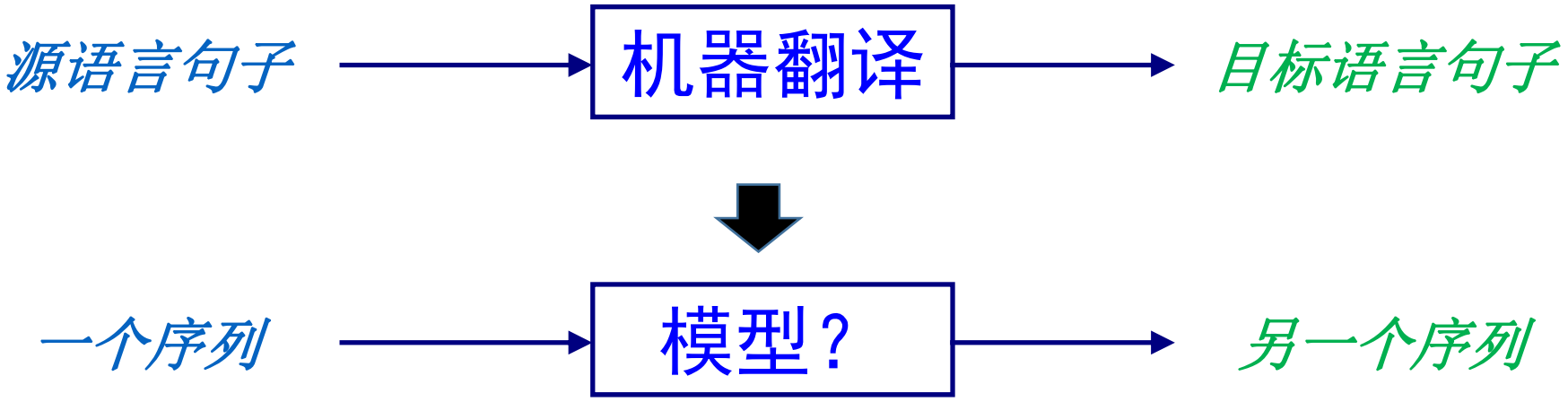
现在的机器翻译


Google
Translate
Chinese
English
Spanish
Detect language
Nine out of the ten Chinese crew members freed by Somali pirates took a flight home on Monday from the Kenyan capital Nairobi, accompanied by officials sent from Beijing.
English
Chinese (Simplified)
Spanish
Translate
在索马里海盗释放的10名中国船员中，有9名在星期一从肯尼亚首都内罗毕飞往家乡，并从北京派出官员陪同。
Suggest an edit
Zài suǒmǎlǐ hǎidào shìfàng de 10 míng zhōngguó chuányuán zhōng, yǒu 9 míng zài xīngqí yī cóng kěnníyǎ shǒudū nèiluóbì fēi wǎng jiāxiāng, bìng cóng běijīng pàichū guānyuán péitóng.

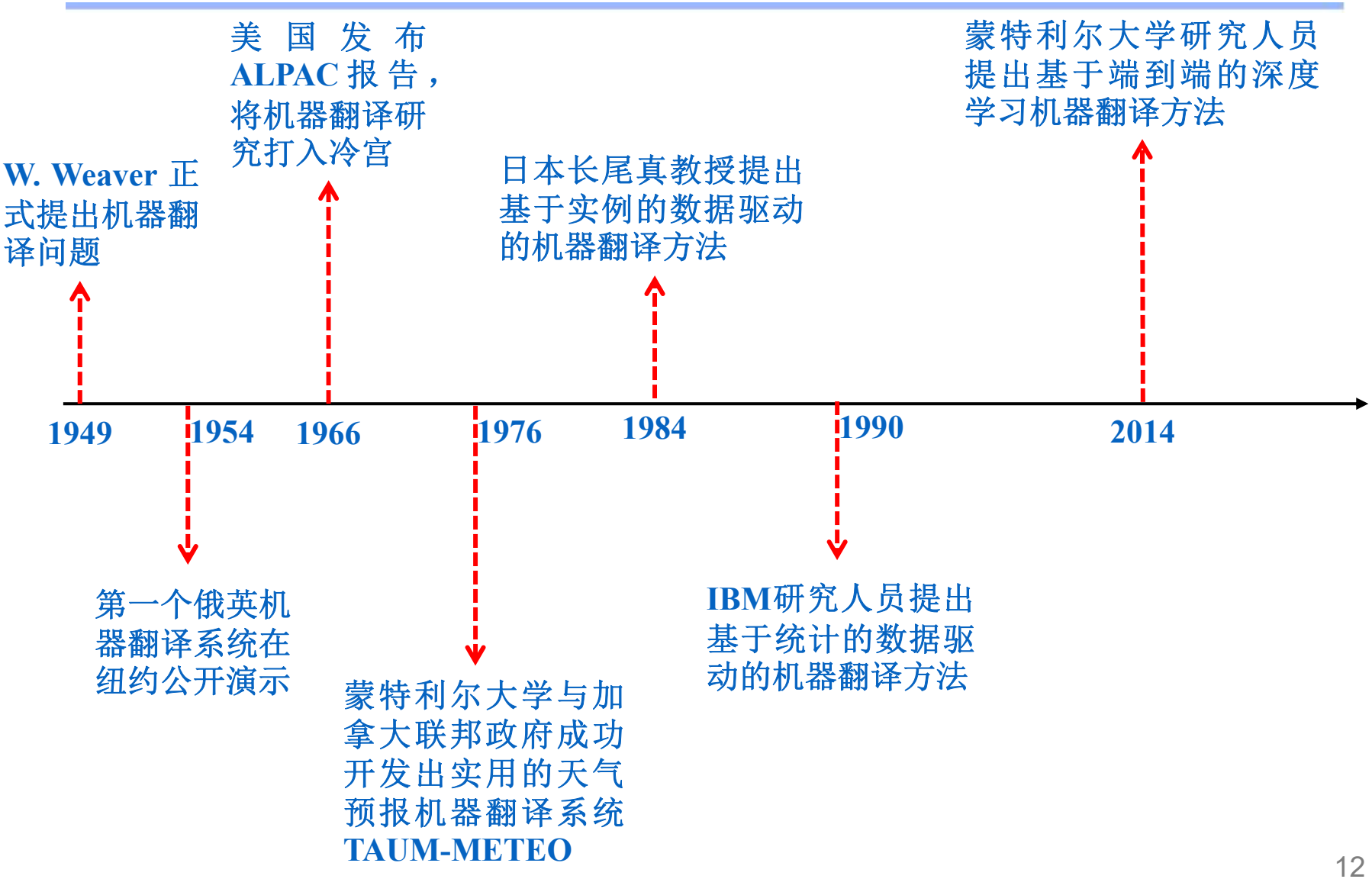
机器翻译形式化

今天北京天气不错

The weather is fine in Beijing today



机器翻译简史


W. Weaver 正式提出机器翻译问题
美国发布ALPAC报告，将机器翻译研究打入冷宫
日本长尾真教授提出基于实例的数据驱动的机器翻译方法
蒙特利尔大学研究人员提出基于端到端的深度学习机器翻译方法
1949
1954
1966
1976
1984
1990
2014
第一个俄英机器翻译系统在纽约公开演示
蒙特利尔大学与加拿大联邦政府成功开发出实用的天气预报机器翻译系统TAUM-METEO
IBM研究人员提出基于统计的数据驱动的机器翻译方法

机器翻译的困难

- **自然语言中普遍存在的歧义和未知现象**
  - 句法结构歧义/词汇歧义/语用歧义 …
  - 新的词汇、术语、结构、语义 …
- **机器翻译不仅仅是字符串的转换**
  - 不同语言之间文化的差异
  - 现有方法无法表示和利用世界知识和常识
- **机器翻译的解不唯一，而且始终存在的人为的标准**

几乎自然语言处理中的所有问题在机器翻译中都会遇到

机器翻译方法

- 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（数据驱动）的翻译方法
  - 基于实例的翻译方法
  - 统计机器翻译
  - 神经网络机器翻译

机器翻译方法

- 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（数据驱动）的翻译方法
  - 基于实例的翻译方法
  - 统计机器翻译
  - 神经网络机器翻译

双语对照数据

人类 共 有 二十三 对 染色体 。

humans have a total of 23 pairs of chromosomes .

澳洲 重新 开放 驻 马尼拉 大使馆

australia reopens embassy in manila

中国 大陆 手机 用户 成长 将 减缓

growth of phone users in mainland china to slow

外交 人员 搭乘 第五 架 飞机 返国

diplomatic staff take the fifth plane home

驻 南韩 美军 三千人 奉命 冻结 调防

us freezes transfer of 3,000 troops in south korea

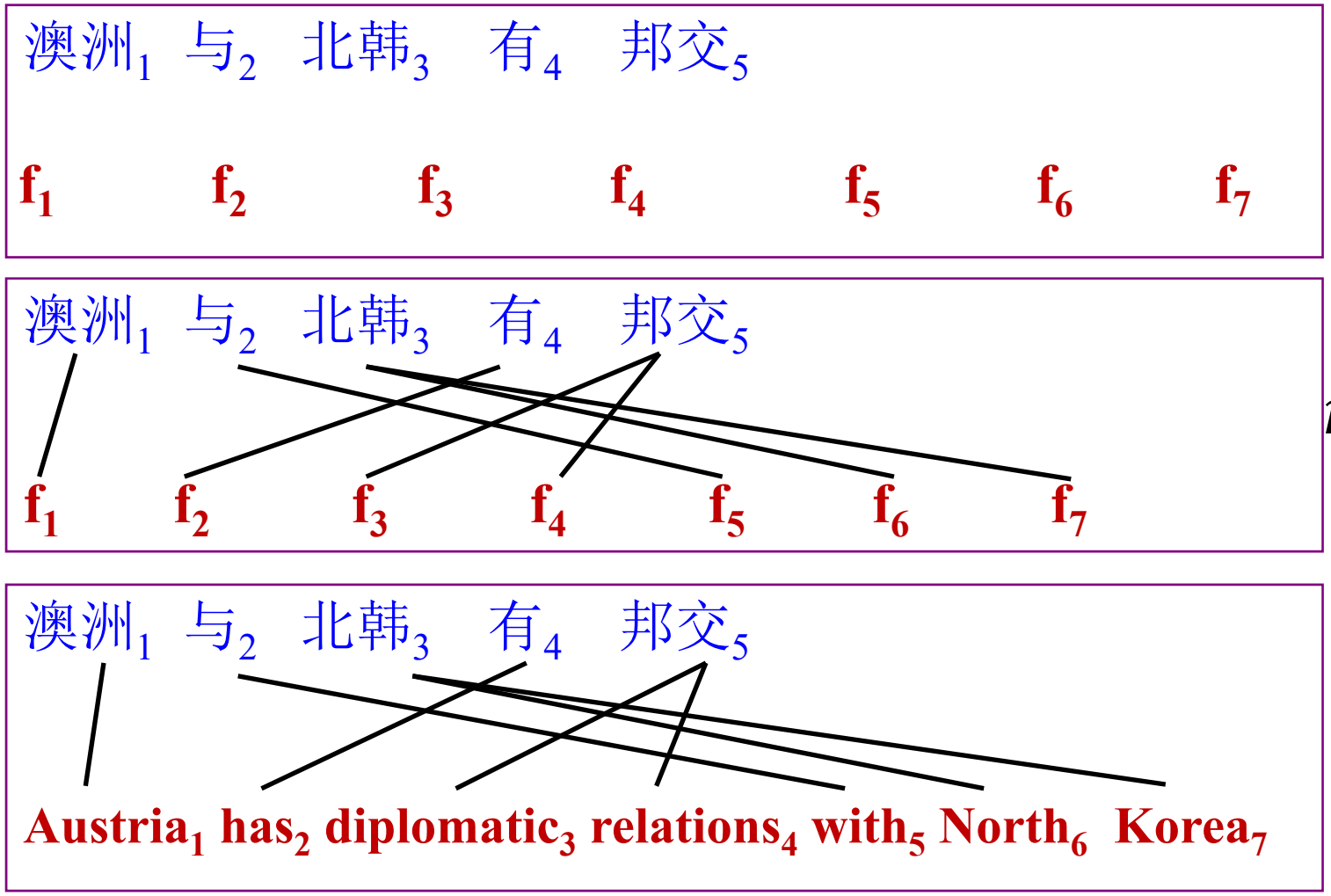
姚明 感慨 NBA 的 偶像 来 得 太 快

yao ming feels nba stardom comes too fast

… …

… …

统计机器翻译-基于词的模型

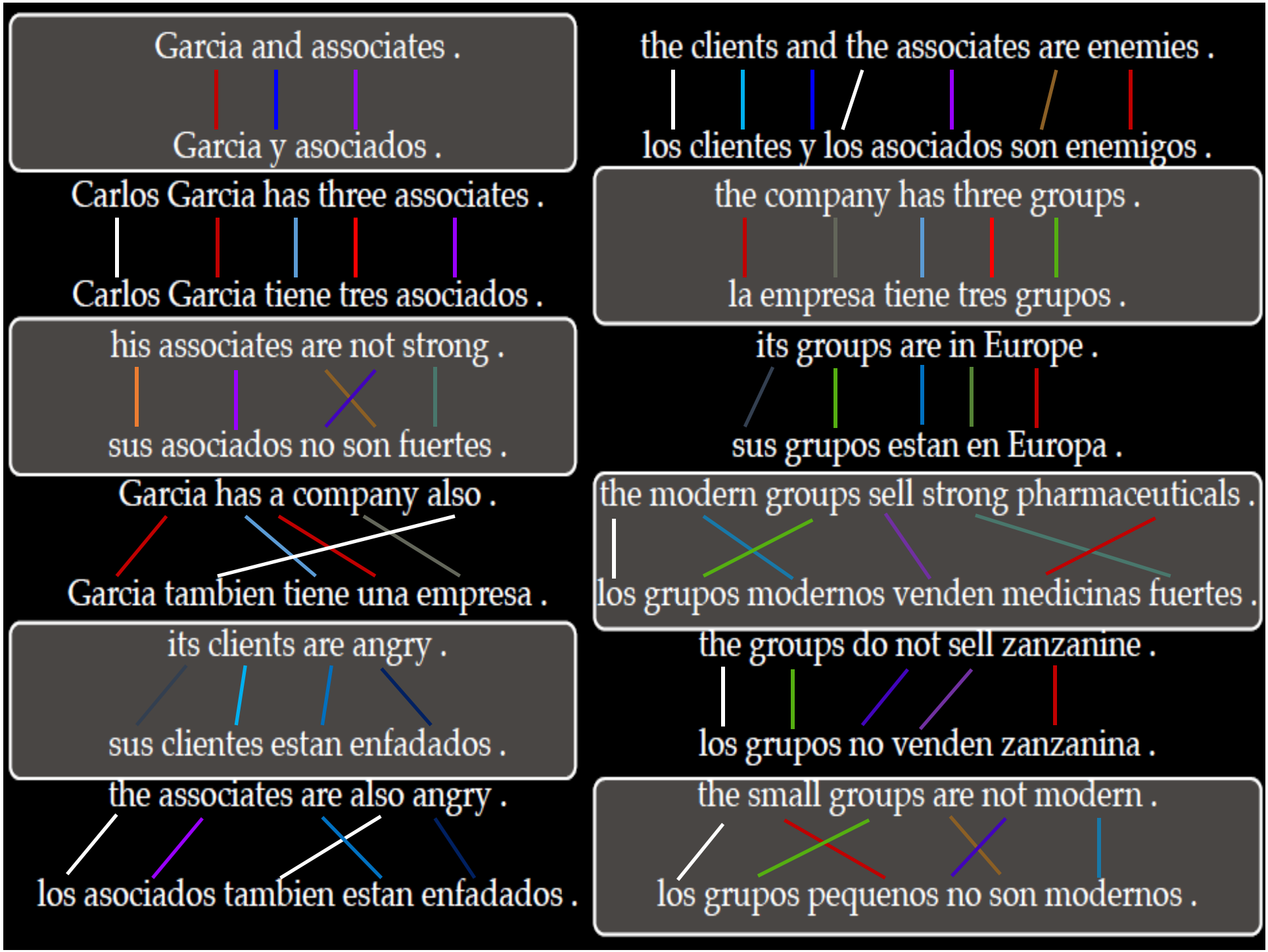


$\varepsilon \equiv p(m|T)$

$p(a_j|j,m,l)$

$p\left(s_j|t_{a_j}\right)$

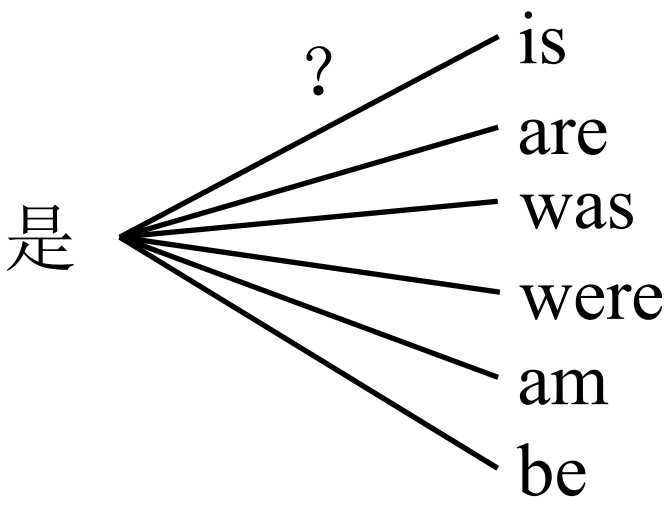
IBM Model-2


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 .

基于短语的统计机器翻译

- 基于词的翻译模型的问题：
  - 很难处理词义消岐问题
  - 很难处理一对多、多对一和多对多的翻译问题

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一



基于短语的统计机器翻译

- 基于词的翻译模型的问题：
  - 很难处理词义消岐问题
  - 很难处理一对多、多对一和多对多的翻译问题

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

北韩 $\overset{?}{\rightarrow}$ North Korea

邦交 $\overset{?}{\rightarrow}$ the diplomatic relations

基于短语的统计机器翻译

- 基于短语的统计机器翻译：

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

(澳洲 是, Australia is)

(与 北韩, with North Korea)

(有 邦交, have the diplomatic relations)

(的 少数 国家 之一, one of the few countries that)

基于短语的统计机器翻译

[Koehn, 2003]

短语：连续的词串 （非句法意义）

$$\boldsymbol{T}' = \underset{T}{\operatorname{argmax}} \boldsymbol{P}(\boldsymbol{T}|\boldsymbol{S})$$

$$= \underset{T,S_1^K}{\operatorname{argmax}} \boldsymbol{P}(\boldsymbol{T}, S_1^K|\boldsymbol{S})$$

$$= \underset{T,S_1^K,T_1^K,T_1^{K'}}{\operatorname{argmax}} \boldsymbol{P}(S_1^K|\boldsymbol{S}) \times \boldsymbol{P}(T_1^K|S_1^K,\boldsymbol{S}) \times \boldsymbol{P}\left(T_1^{K'}|T_1^K,S_1^K,\boldsymbol{S}\right) \times \boldsymbol{P}\left(\boldsymbol{T}|T_1^{K'},T_1^K,S_1^K,\boldsymbol{S}\right)$$

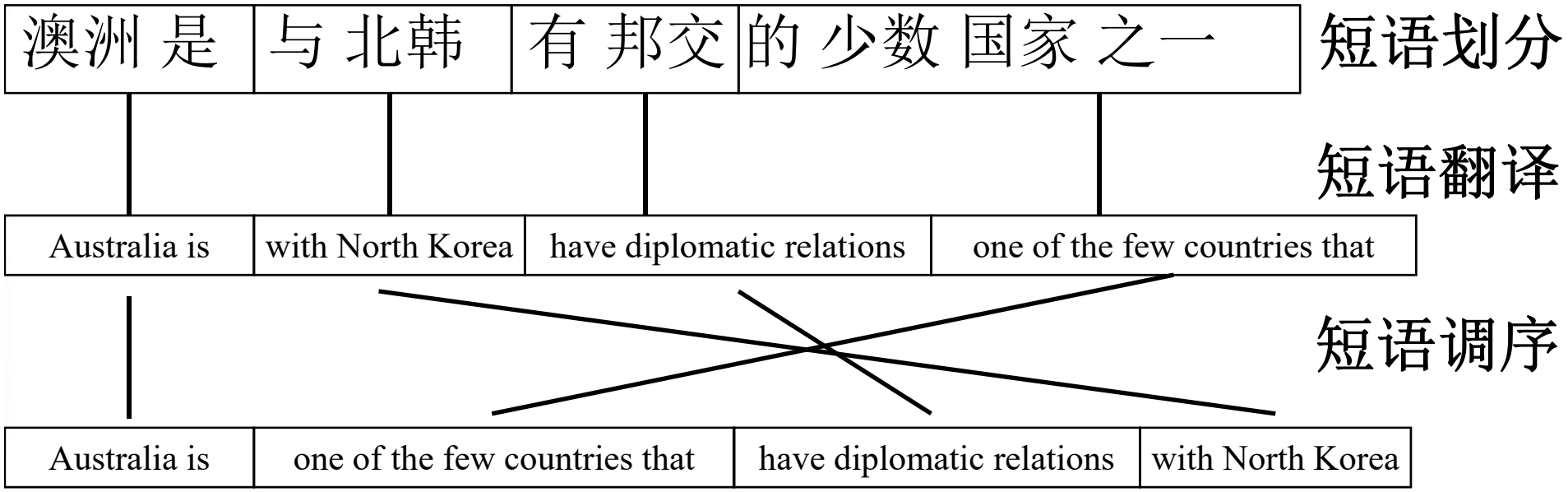
基于短语的统计机器翻译

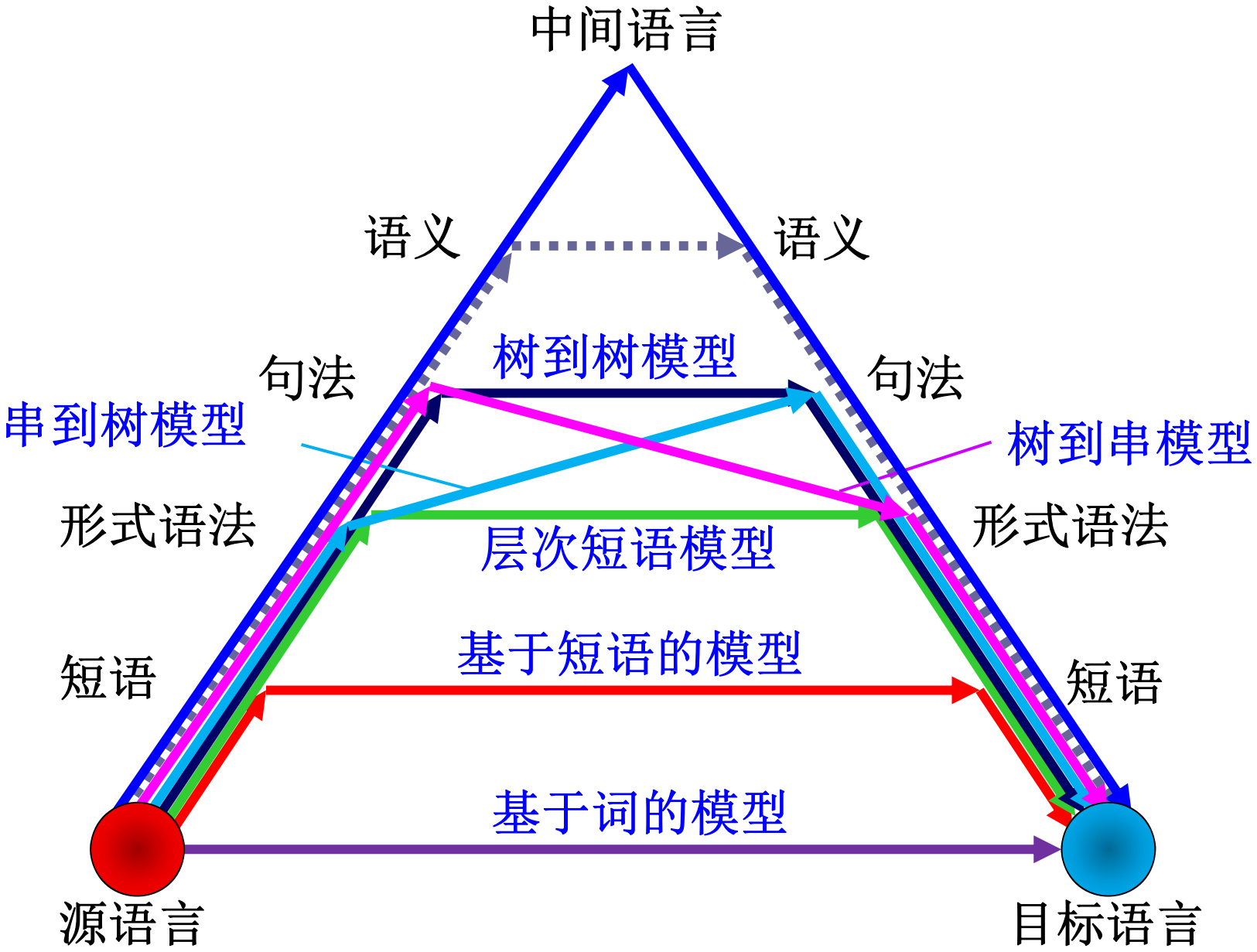
$$\boldsymbol{T}' = \underset{T}{\operatorname{argmax}} \boldsymbol{P}(\boldsymbol{T}|\boldsymbol{S})$$

$$= \underset{T,S_1^K}{\operatorname{argmax}} \boldsymbol{P}(\boldsymbol{T}, S_1^K|\boldsymbol{S})$$

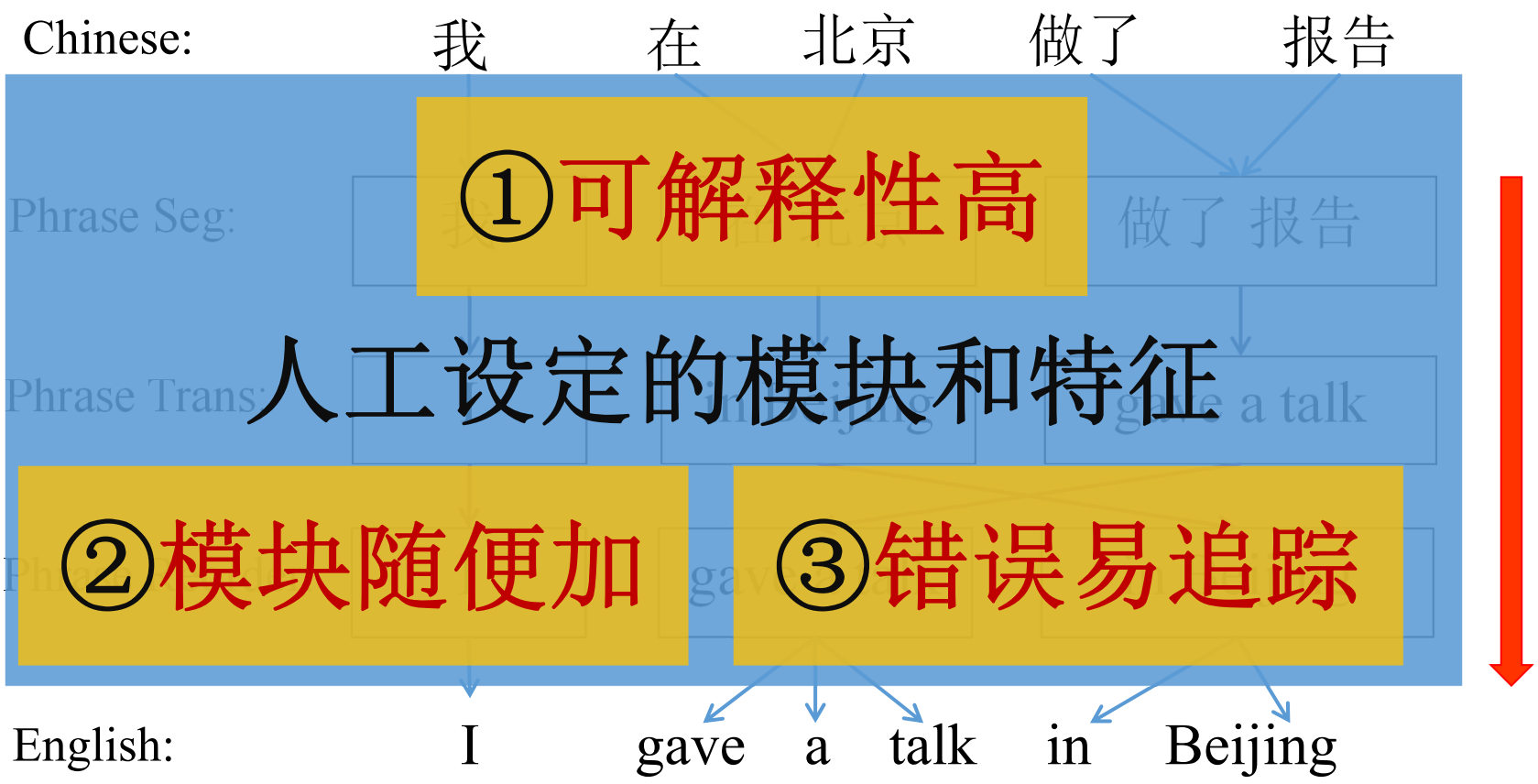
$$= \underset{T,S_1^K,T_1^K,T_1^{K'}}{\operatorname{argmax}} \underbrace{\boldsymbol{P}(S_1^K|\boldsymbol{S})}_{\text{短语划分模型}} \underbrace{\boldsymbol{P}(T_1^K|S_1^K,\boldsymbol{S})}_{\text{短语翻译模型}} \underbrace{\boldsymbol{P}\left(T_1^{K'}|T_1^K,S_1^K,\boldsymbol{S}\right)}_{\text{短语调序模型}} \underbrace{\boldsymbol{P}\left(\boldsymbol{T}|T_1^{K'},T_1^K,S_1^K,\boldsymbol{S}\right)}_{\text{目标语言模型}}$$

基于短语的统计机器翻译

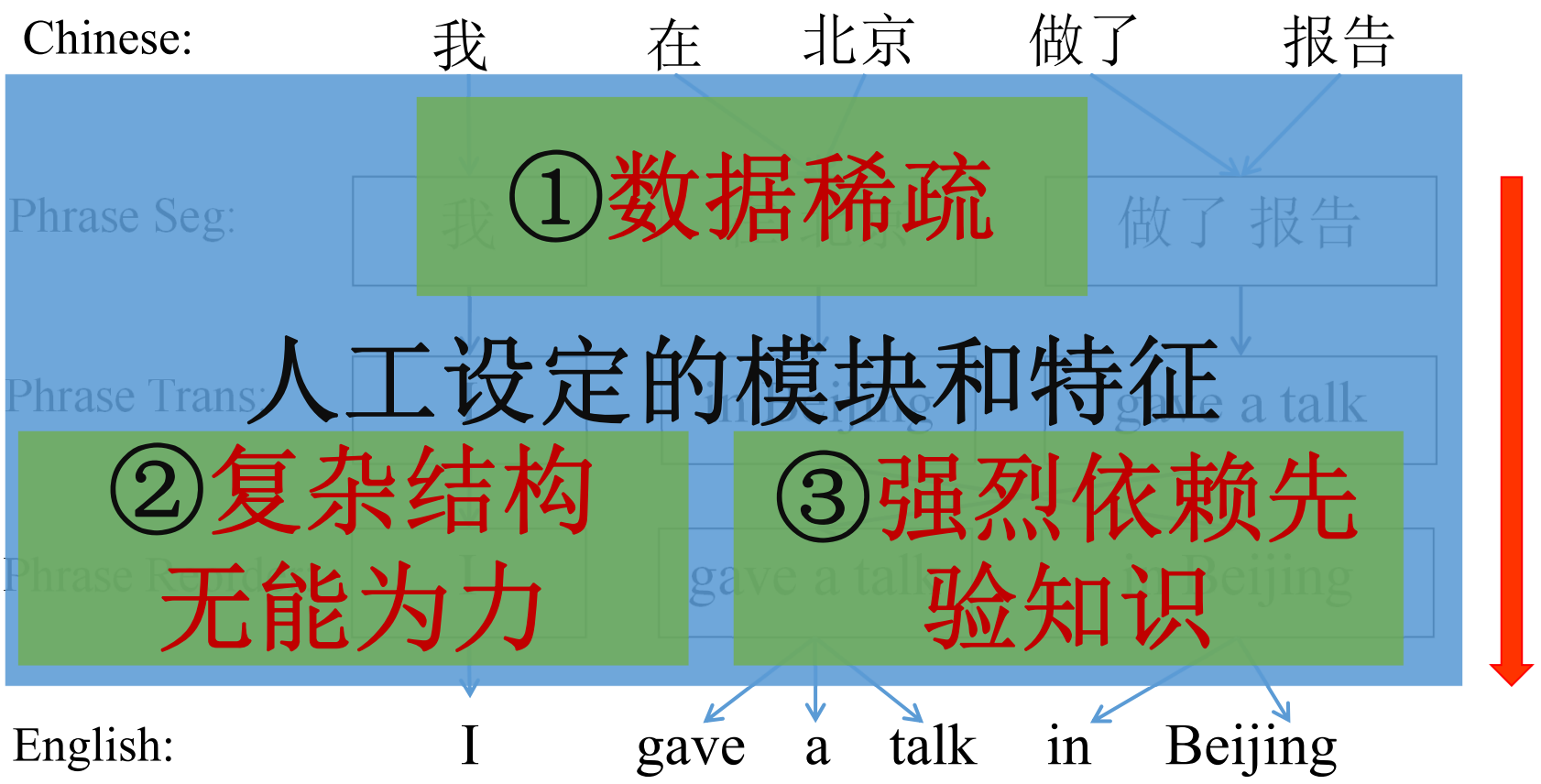



中间语言
语义
语义
句法
句法
树到树模型
串到树模型
树到串模型
形式语法
形式语法
层次短语模型
短语
短语
基于短语的模型
基于词的模型
源语言
目标语言

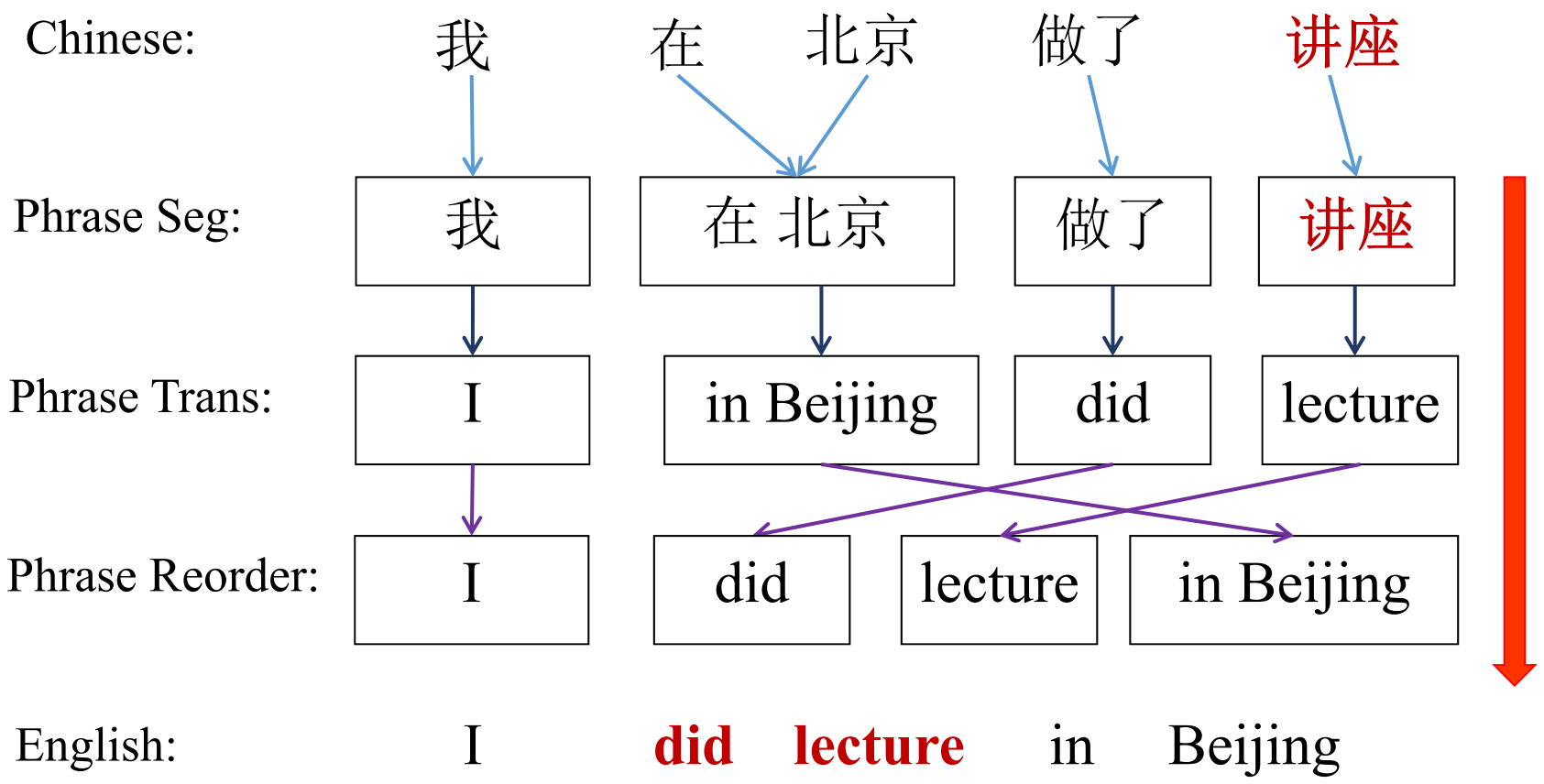
统计机器翻译


Chinese:
我 在 北京 做了 报告
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
English:
I gave a talk in Beijing

统计机器翻译


Chinese:
我
在
北京
做了
报告
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
English:
I
gave
a
talk
in
Beijing

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

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

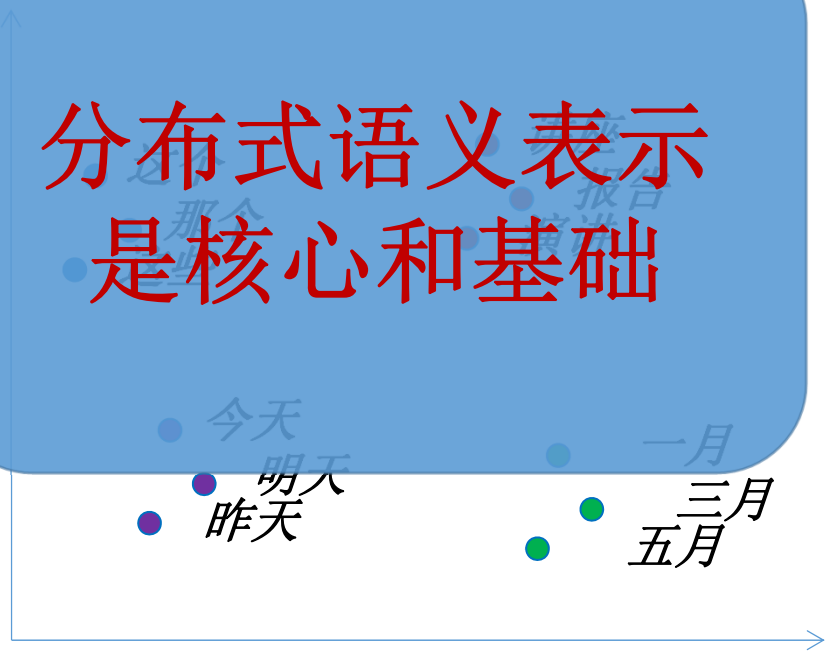
离散符号表示方法 ⇨ 连续分布式表示方法

讲座 ⊗ 报告 = 0

讲座 报告

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

分布式语义表示
是核心和基础



低维、稠密的连续实数空间

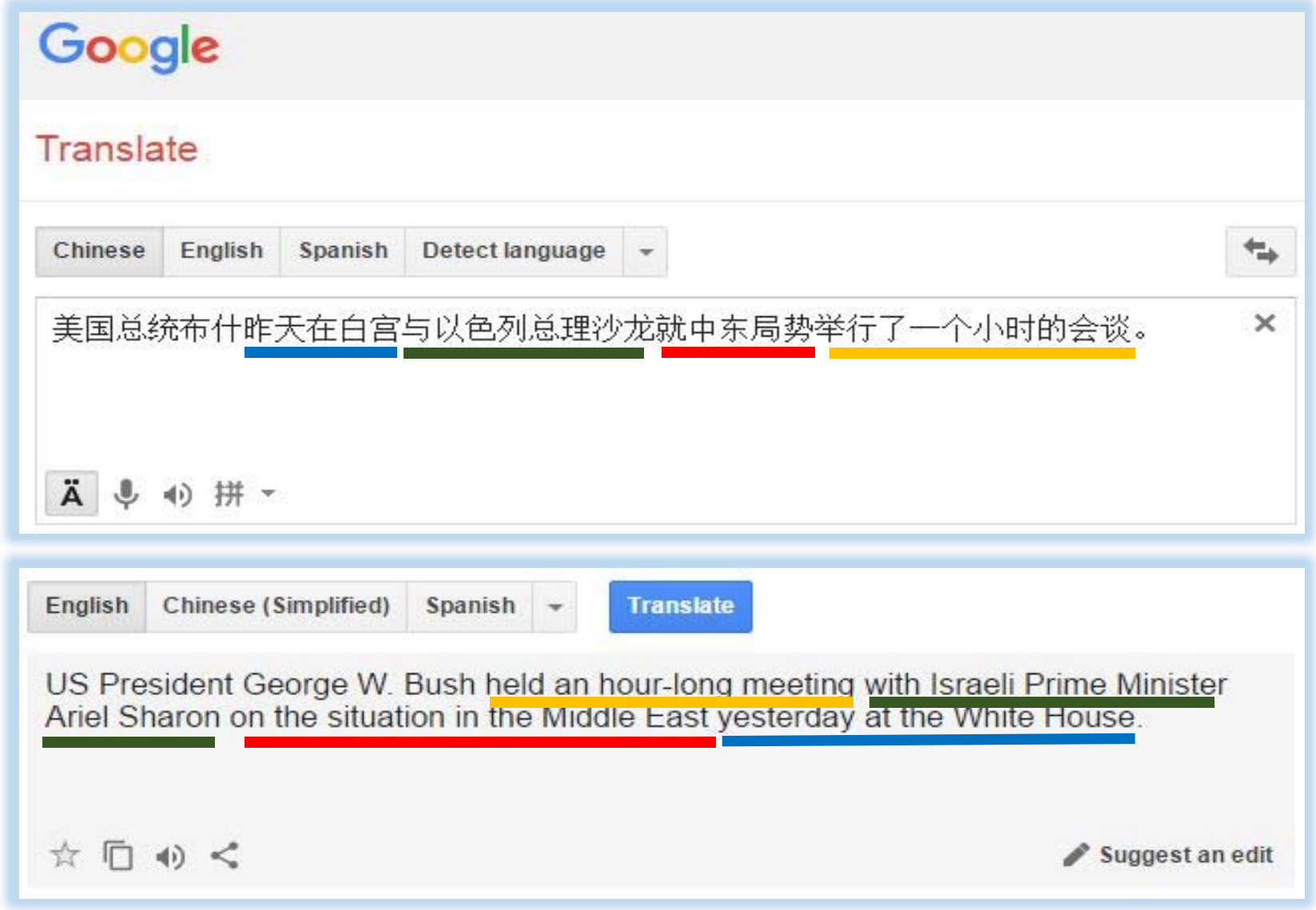
神经机器翻译

Chinese: 我 在 北京 做了 报告

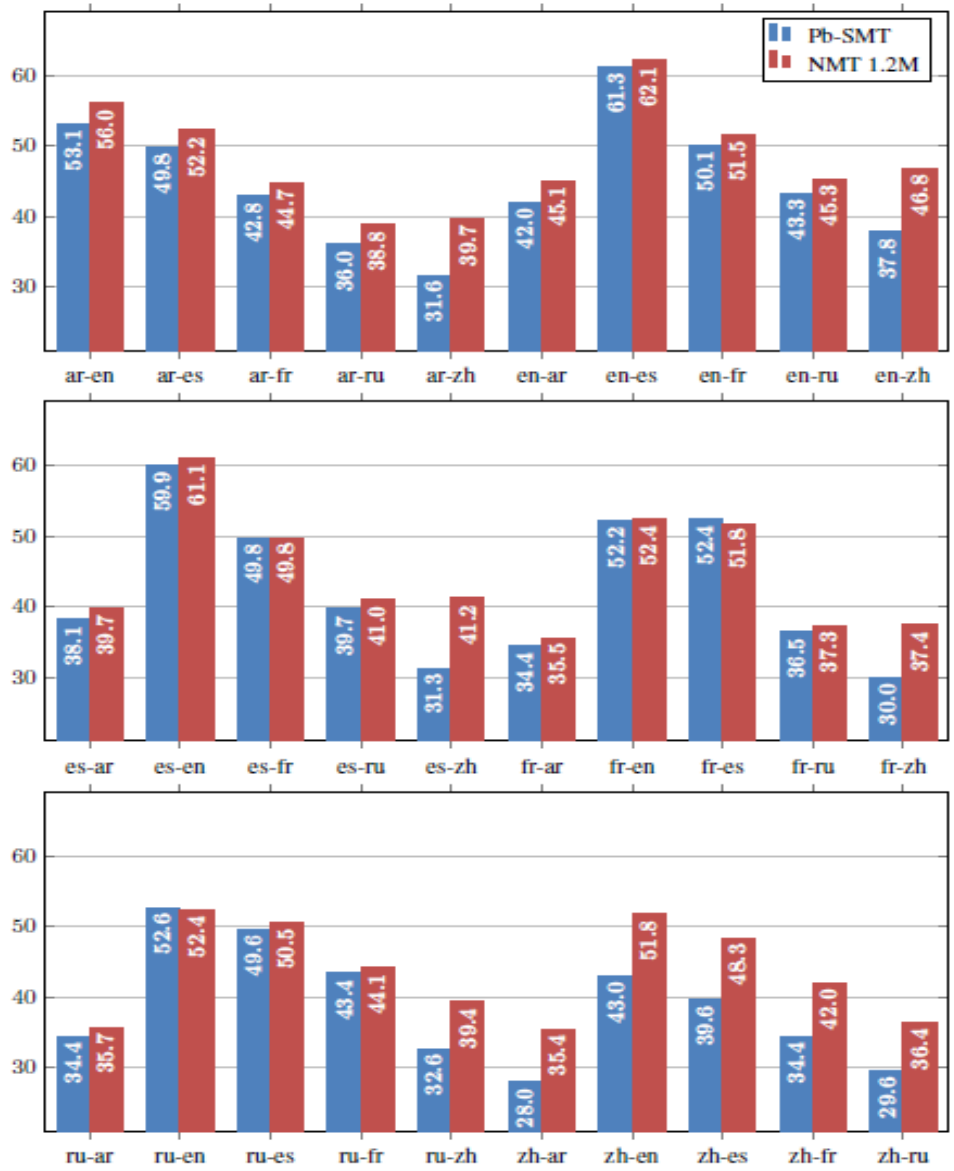
仅需要两个神经网络

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。
拼
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



神经机器翻译压倒性胜出！

[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

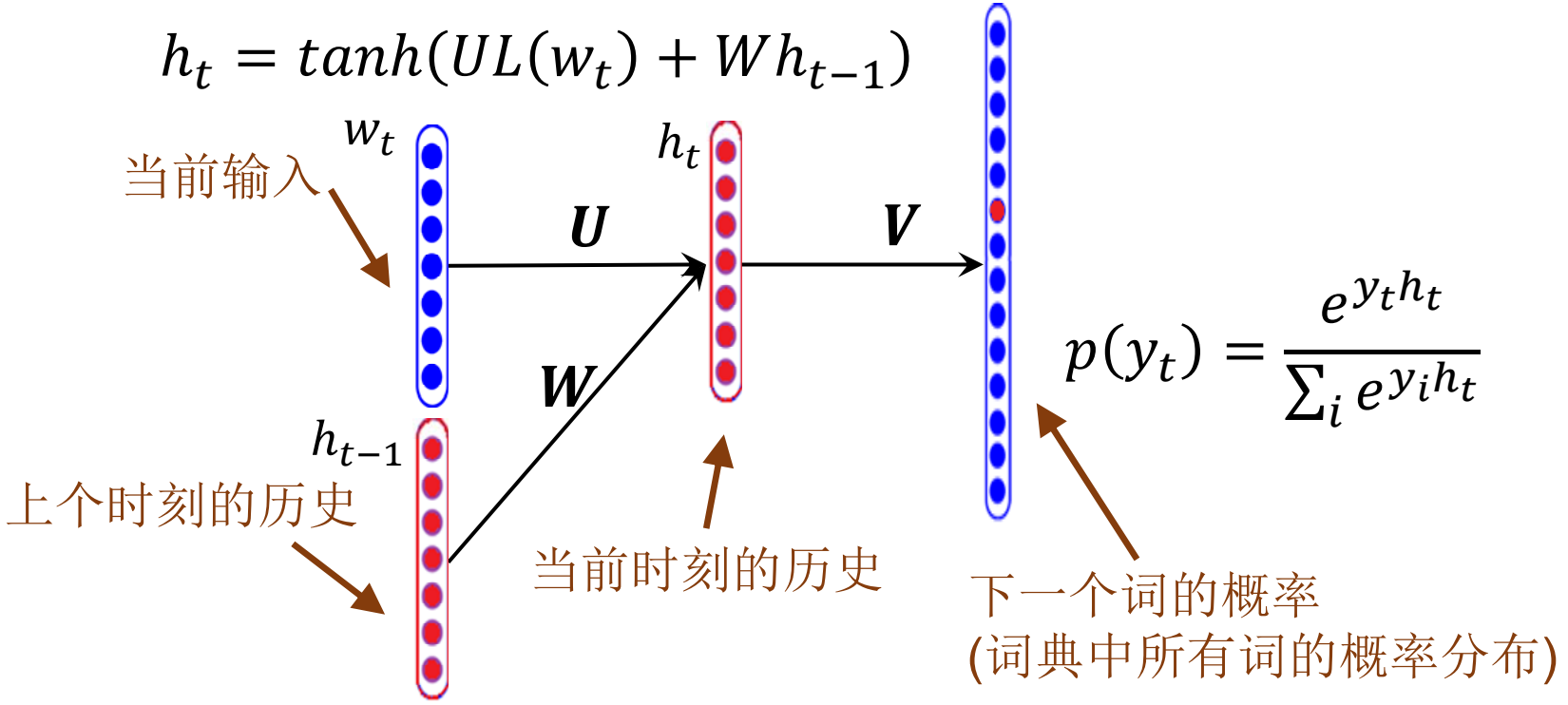
$$\begin{bmatrix}0.48\\0.46\\0.26\end{bmatrix} \otimes \begin{bmatrix}0.42\\0.51\\0.21\end{bmatrix} \approx 1$$

表示是核心
运算是关键

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 $t+1$时刻所有词的概率分布

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



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

源语言编码

$$h_t = tanh(U_S L(w_t) + W_S h_{t-1})$$

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

<s> 我 在 北京 做了 报告 </s>

$$\begin{bmatrix} 0.2 & 0.1 & 0.1 & 0.3 & 0.2 & 0.4 & 0.3 \\ 0.3 & 0.9 & 0.2 & 0.8 & 0.1 & 0.1 & 0.1 \\ 0.5 & 0.6 & 0.4 & 0.3 & 0.3 & 0.2 & 0.2 \end{bmatrix}$$

源语言编码

$$h_t = tanh(U_S L(w_t) + W_S h_{t-1})$$

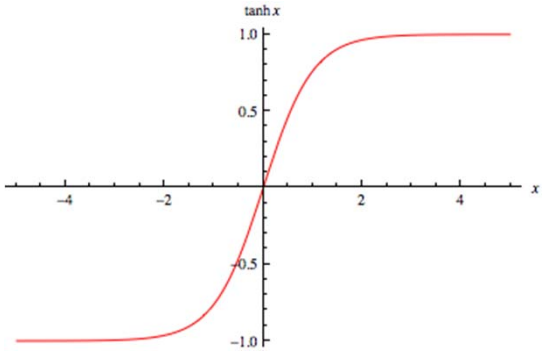
$L(w_t)$: $w_t \longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in R^3$ 我 $\longrightarrow \begin{bmatrix} 0.1 \\ 0.9 \\ 0.6 \end{bmatrix}$ 随机初始化

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

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

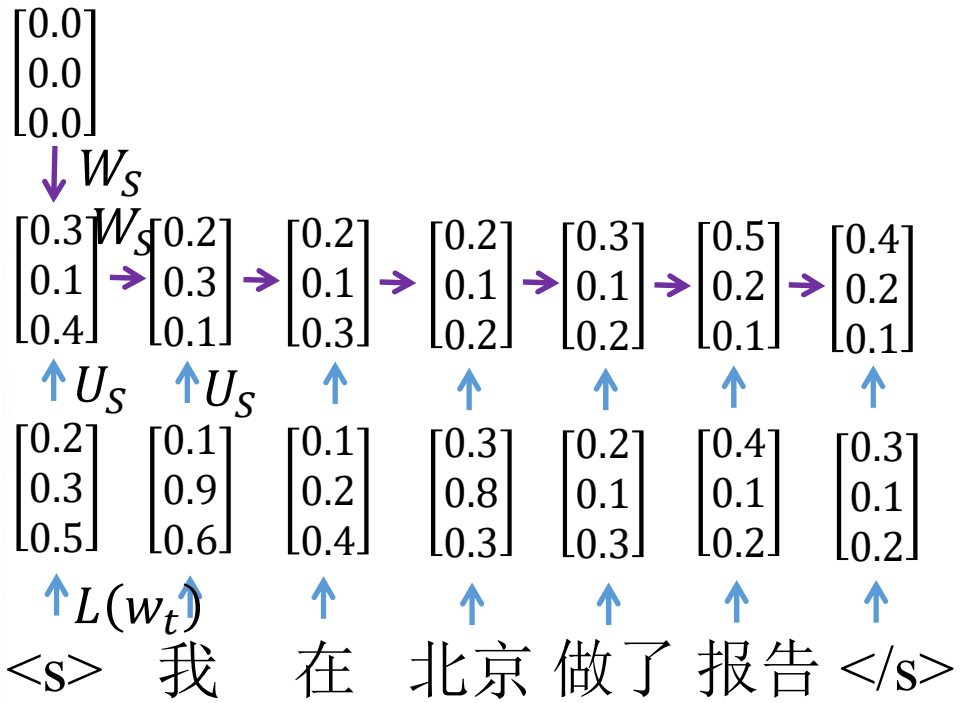
$$z = U_S L(w_t) + W_S h_{t-1} \in R^3$$

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



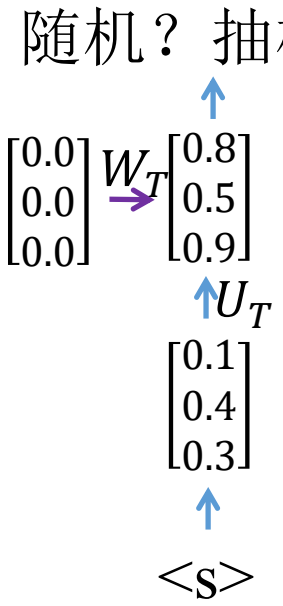
源语言编码

$$h_t = tanh(U_S L(w_t) + W_S h_{t-1})$$

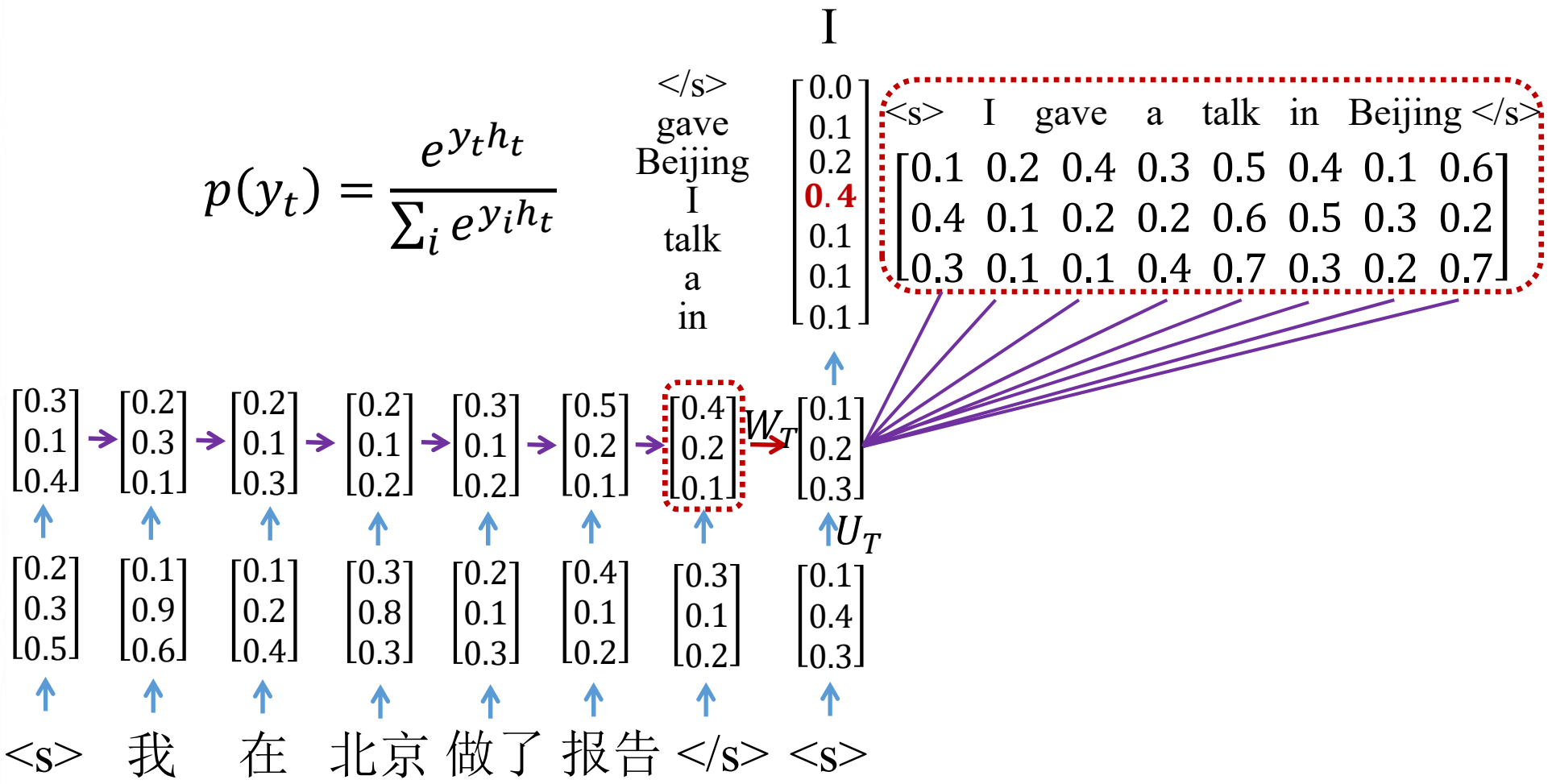


目标语言解码

$$h_t = tanh(U_T L(w_t) + W_T h_{t-1})$$

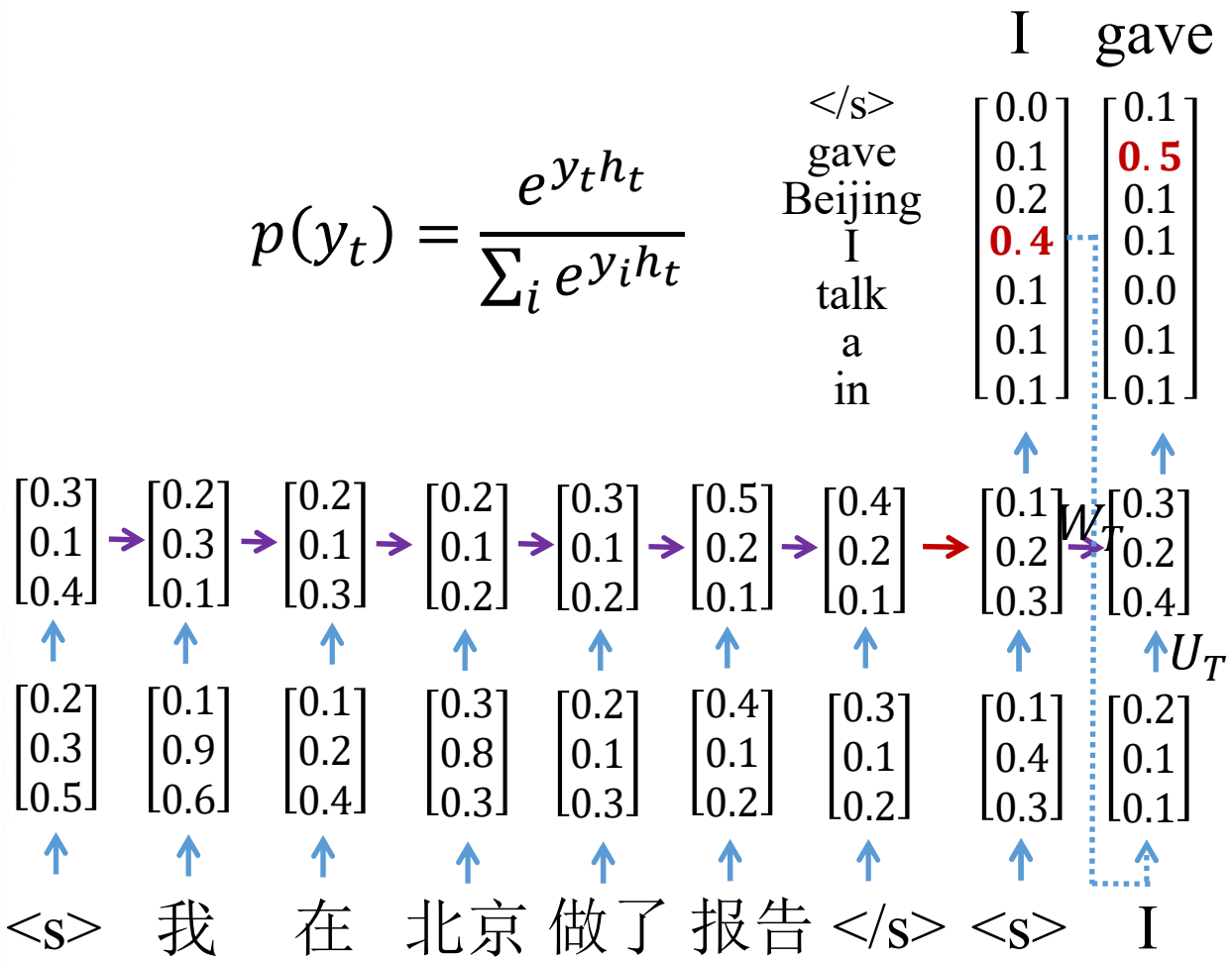


目标语言解码


$p(y_t) = \frac{e^{y_t h_t}}{\sum_i e^{y_i h_t}}$
I
</s>
gave
Beijing
I
talk
a
in
0.0 0.1 0.2 0.4 0.1 0.1 0.1
<s> I gave a talk in Beijing </s>
0.1 0.2 0.4 0.3 0.5 0.4 0.1 0.6
0.4 0.1 0.2 0.2 0.6 0.5 0.3 0.2
0.3 0.1 0.1 0.4 0.7 0.3 0.2 0.7
W_T
U_T
<s> 我 在 北京 做了 报告 </s> <s>

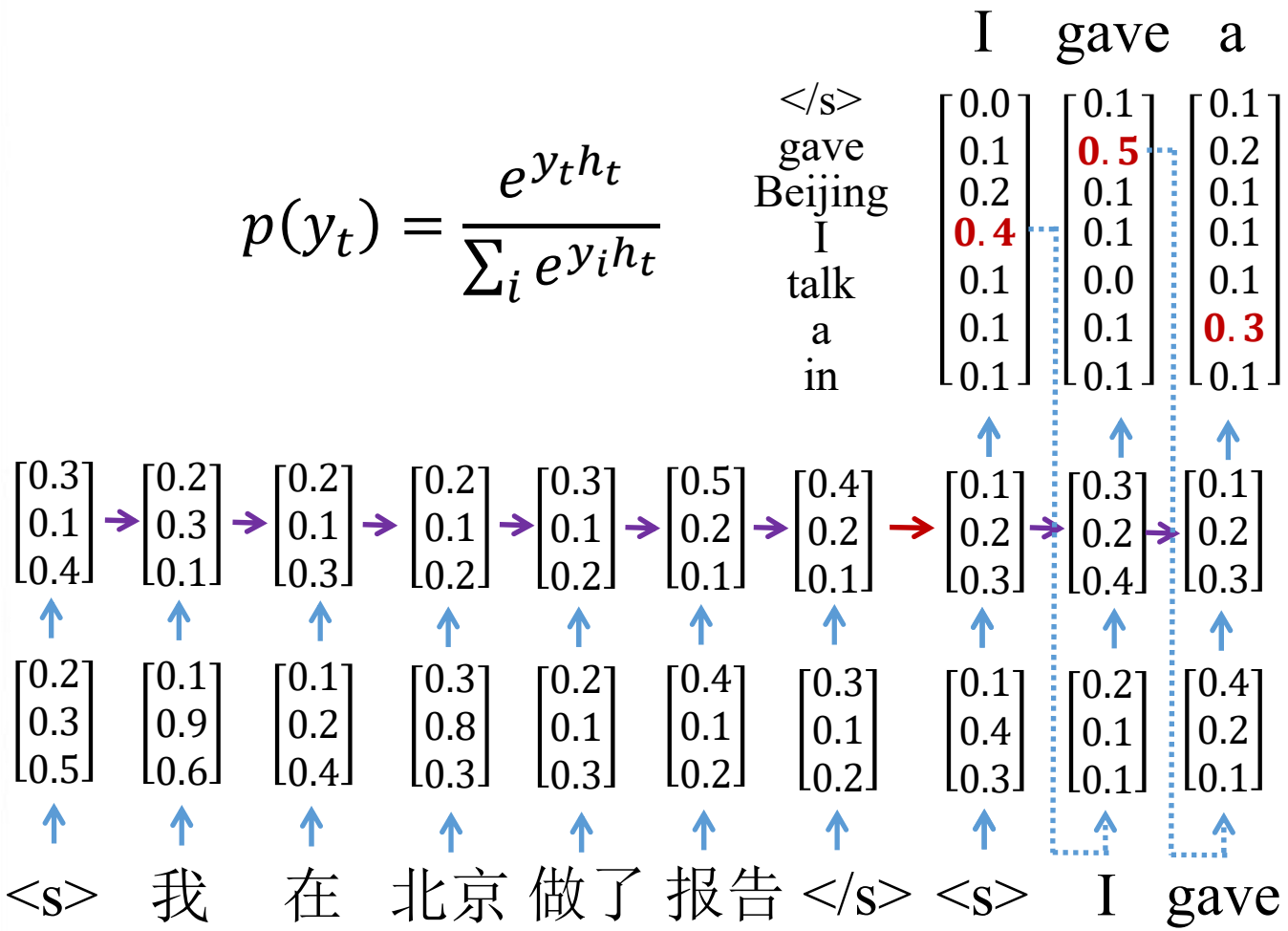
目标语言解码

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



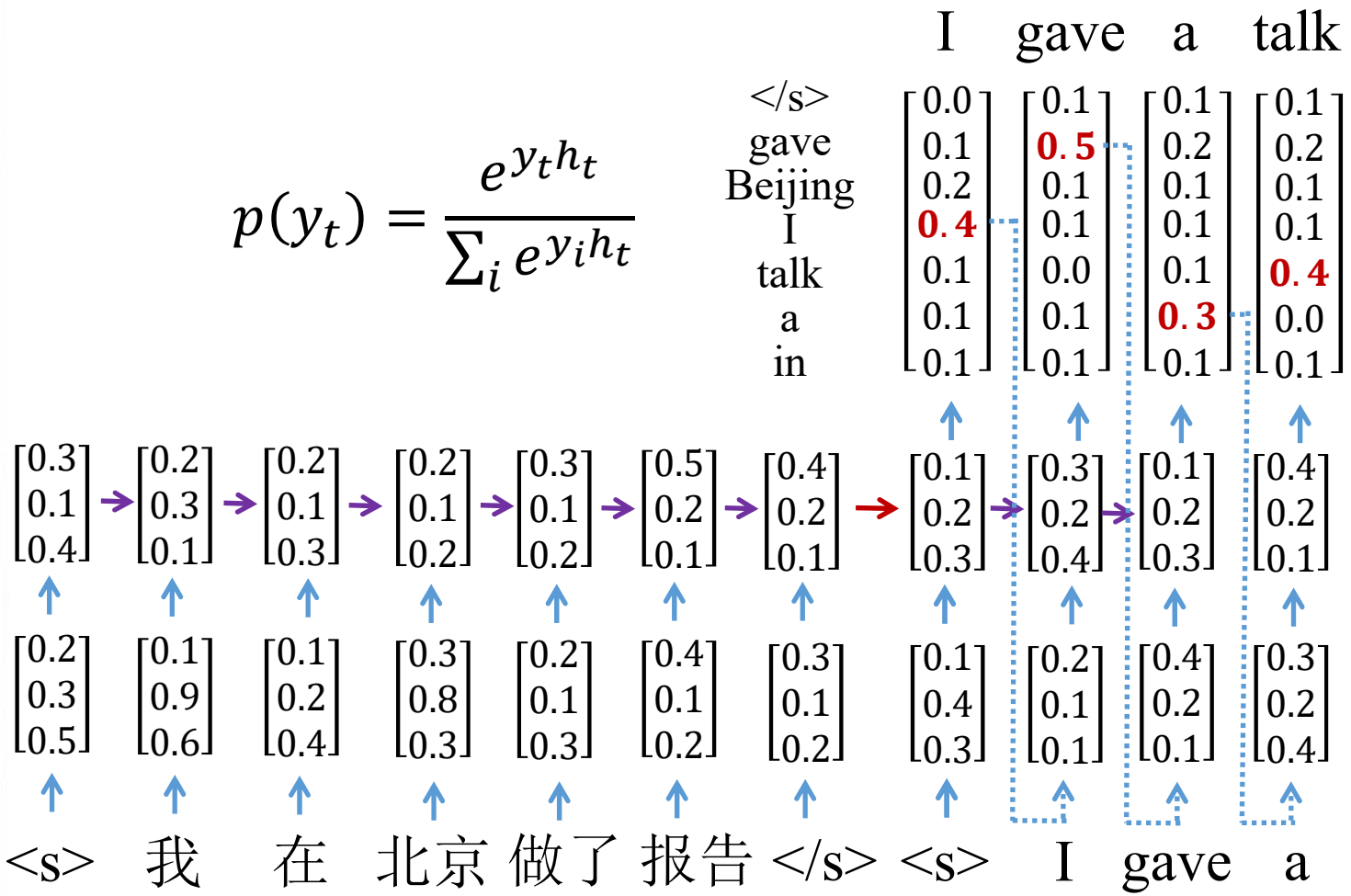
目标语言解码

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



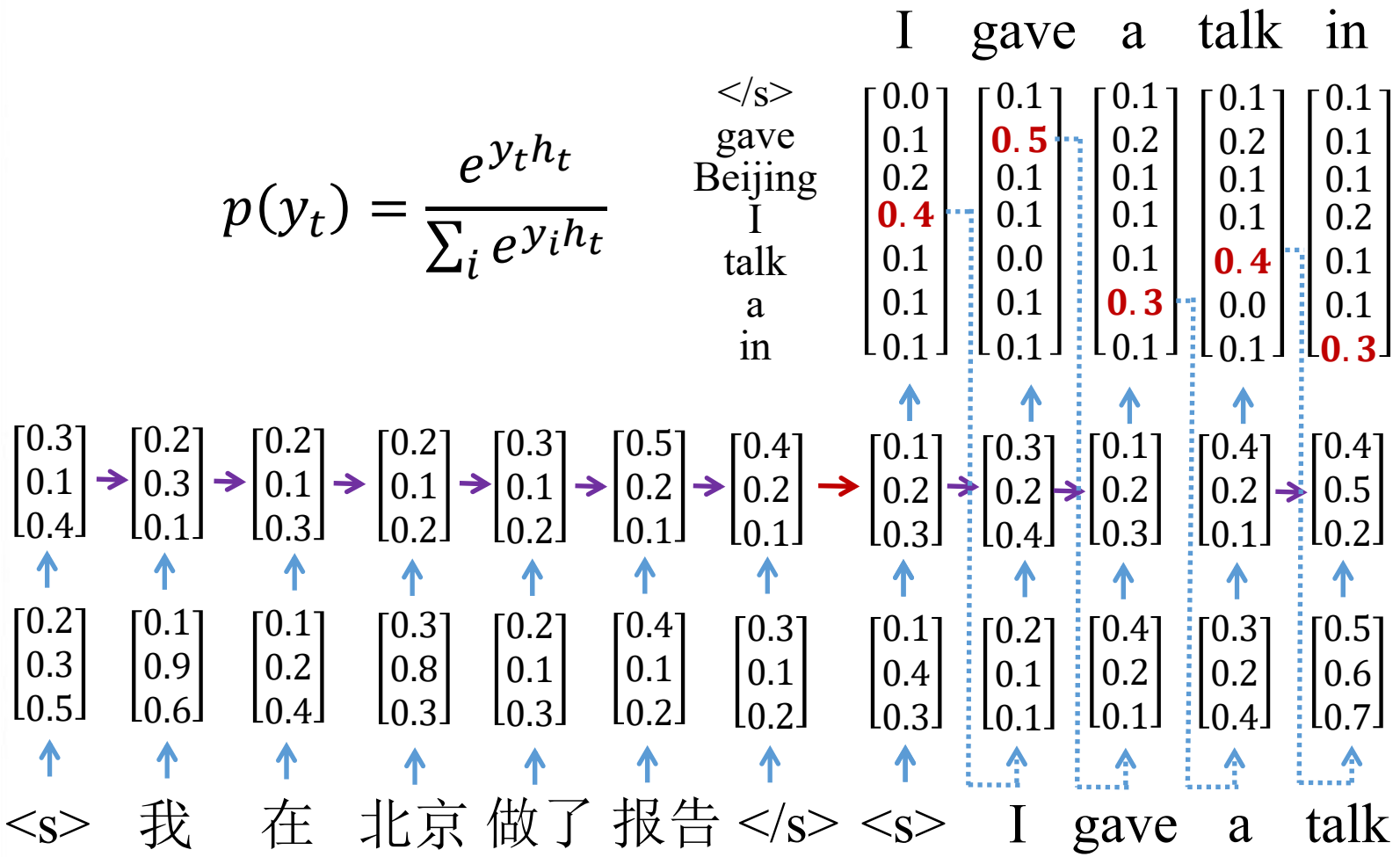
目标语言解码

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



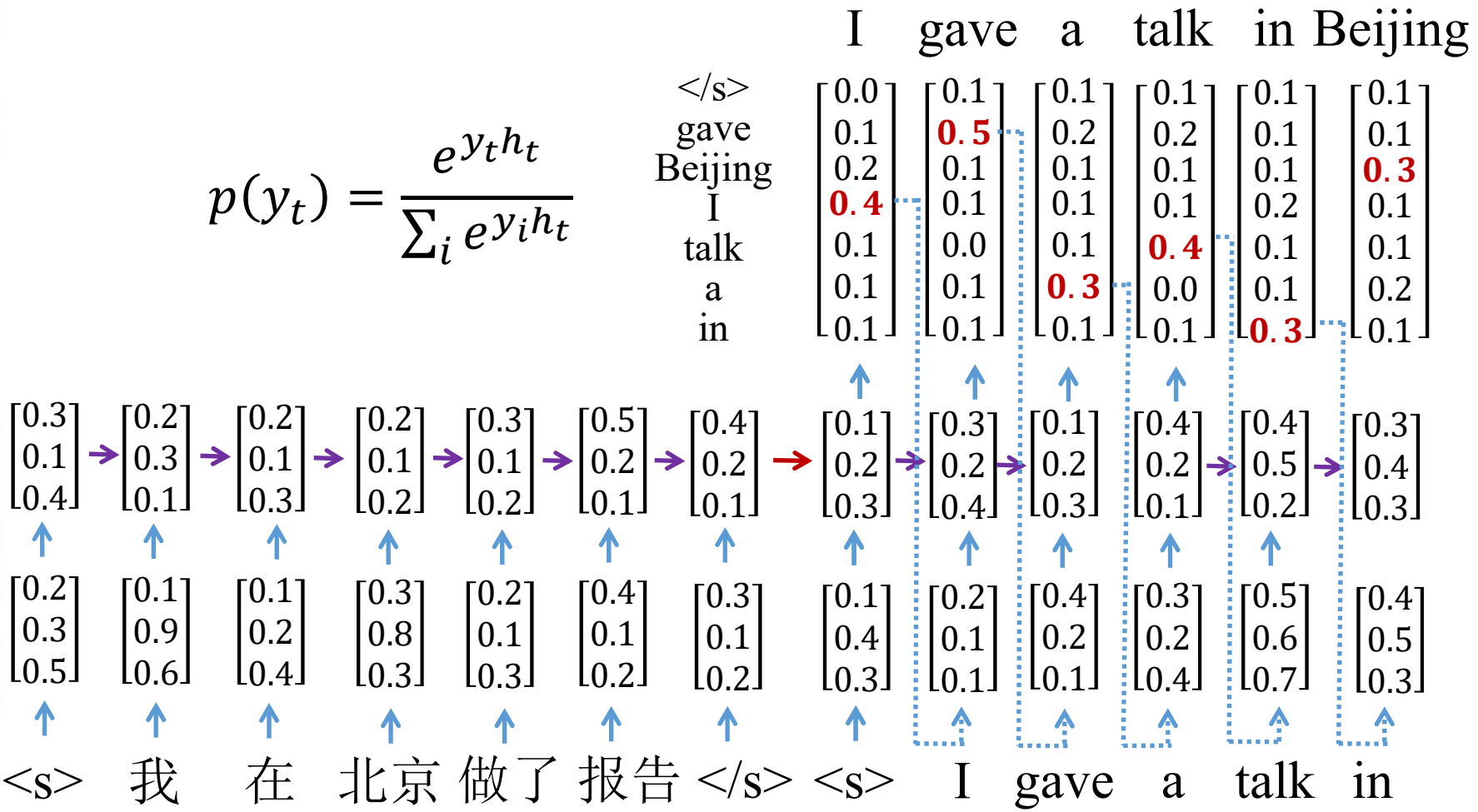
目标语言解码

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



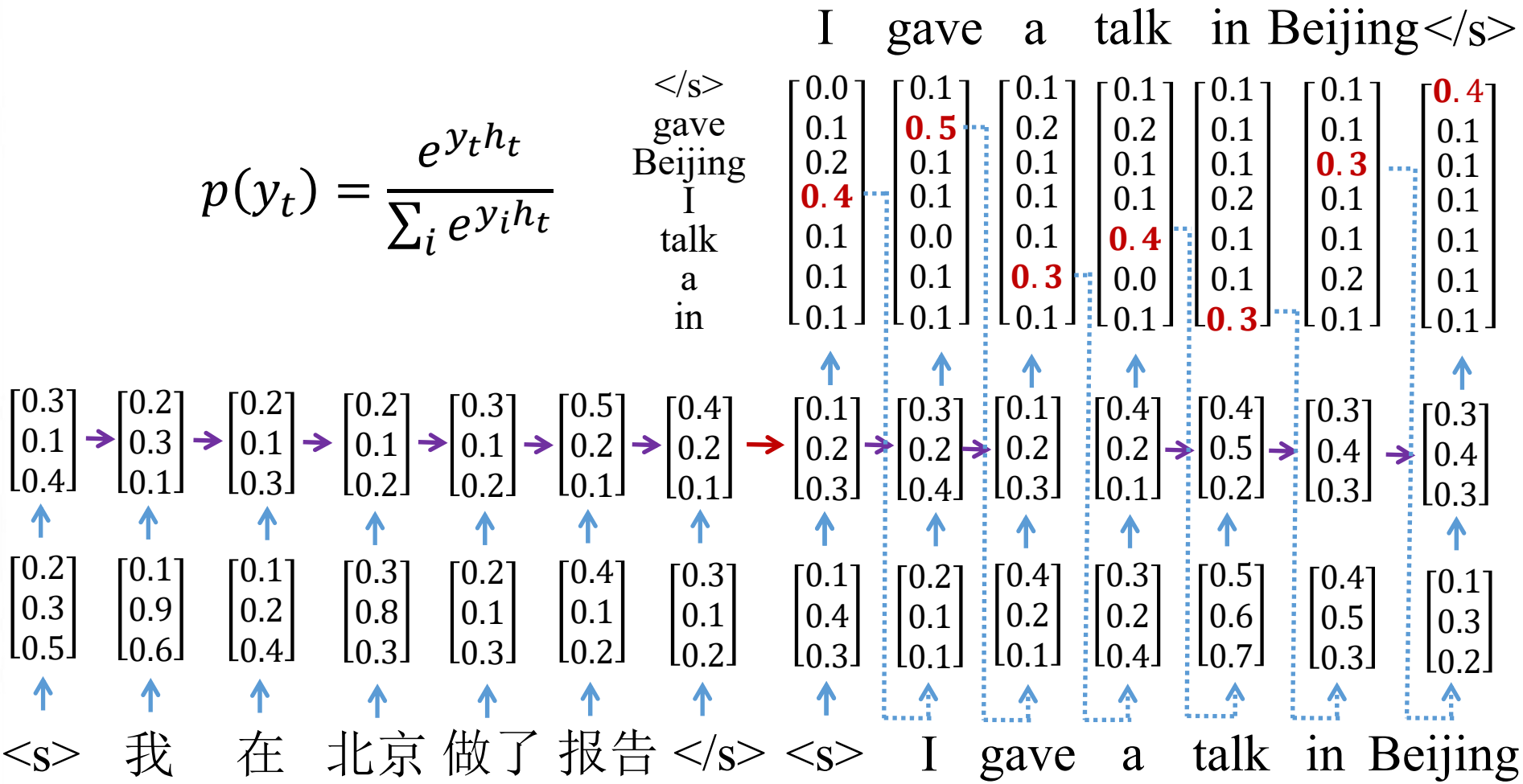
目标语言解码

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

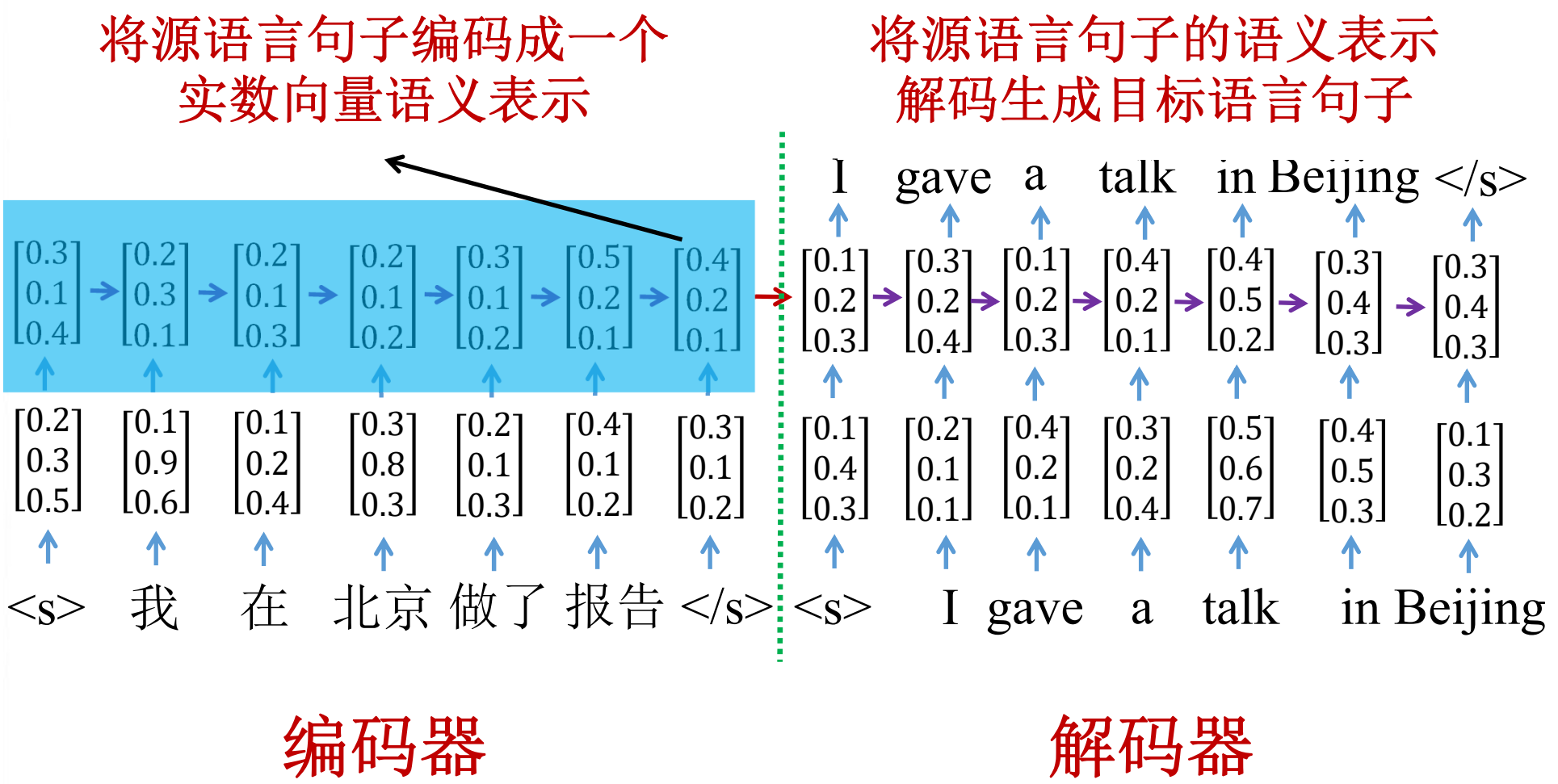


目标语言解码

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

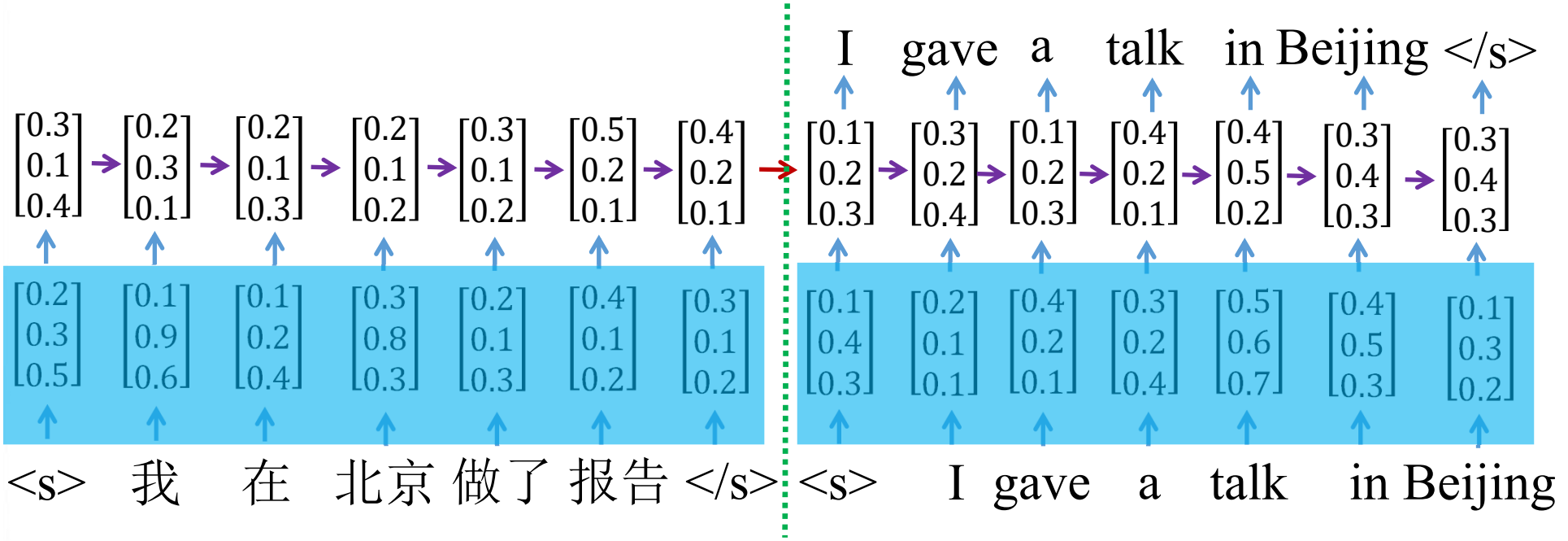


编码-解码网络


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

编码-解码网络

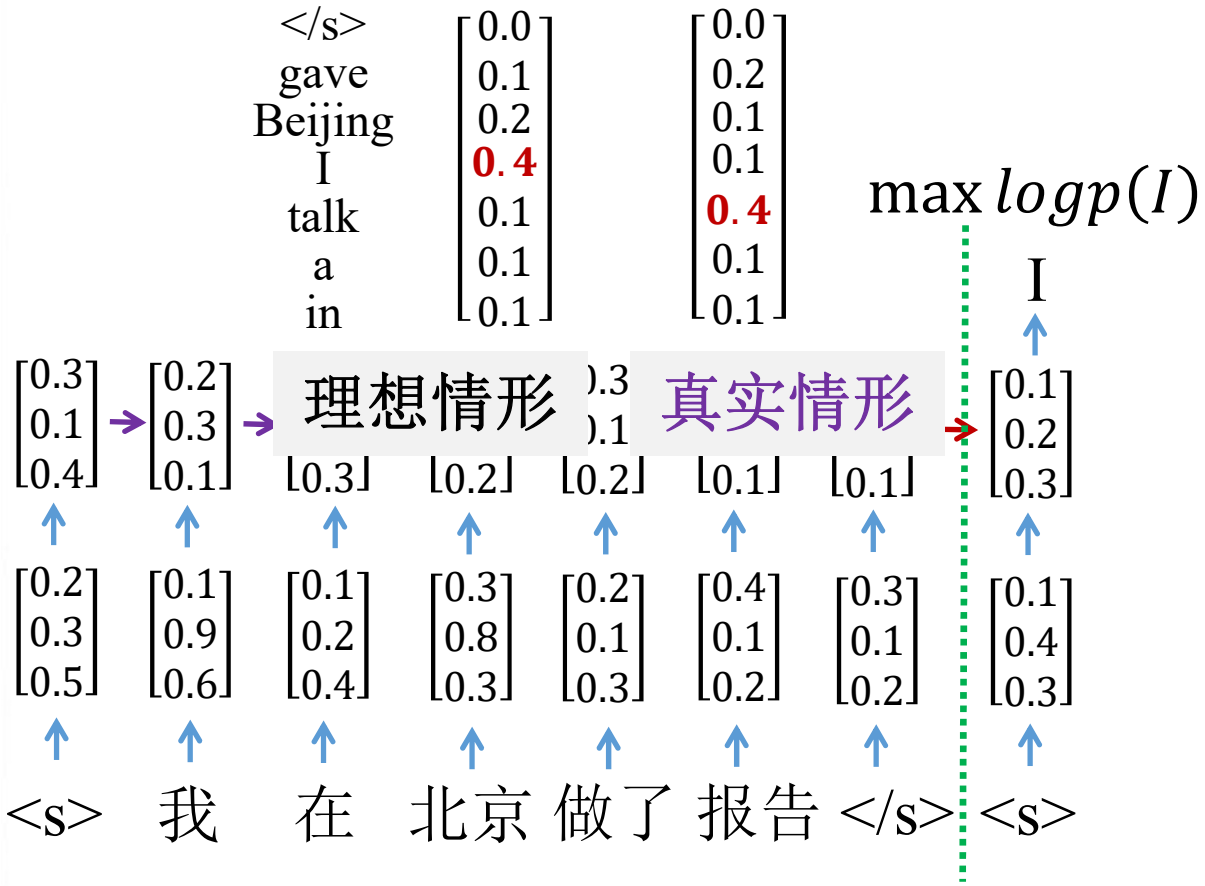
词向量随机初始化，在训练过程中进行优化！



测试 vs. 训练

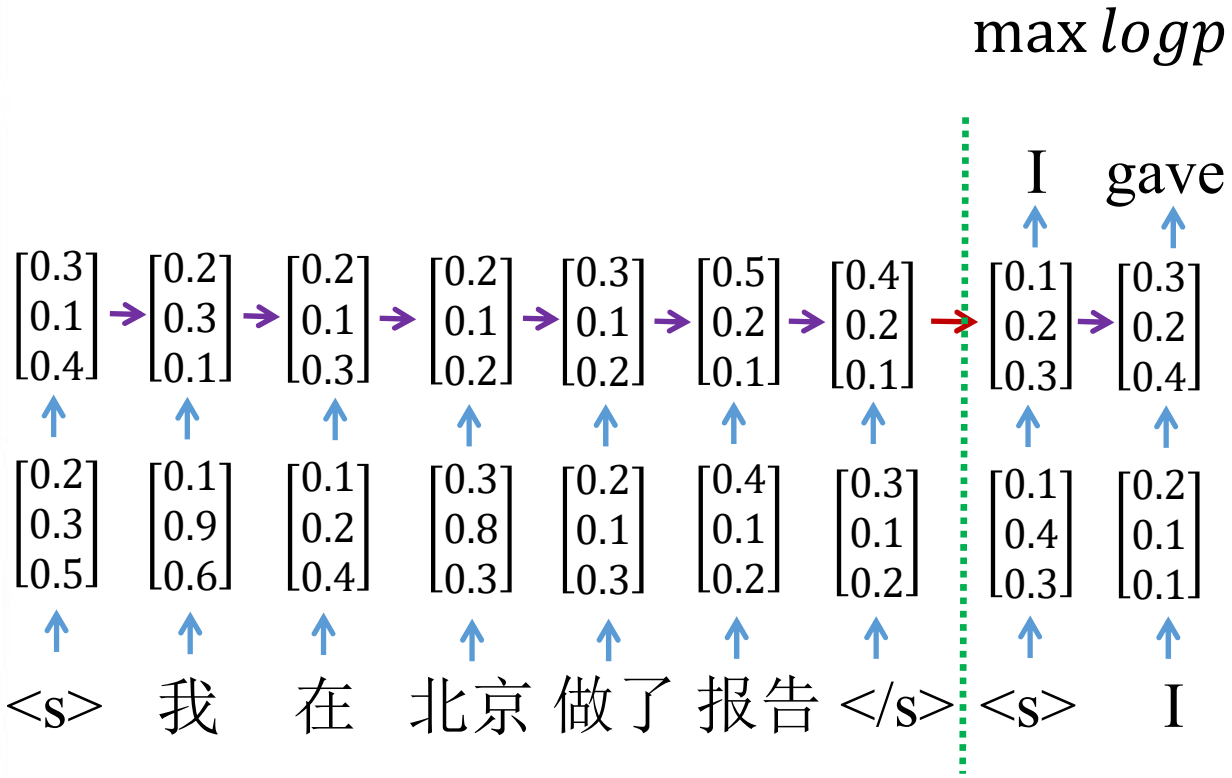

I gave a talk in Beijing</s>
测试解码输入：只有源语言句子
<s> 我 在 北京 做了 报告 </s>
模型训练输入：源语言句子和正确译文
<s> 我 在 北京 做了 报告 </s>
<s> I gave a talk in Beijing </s>
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk in Beijing

神经机器翻译参数优化



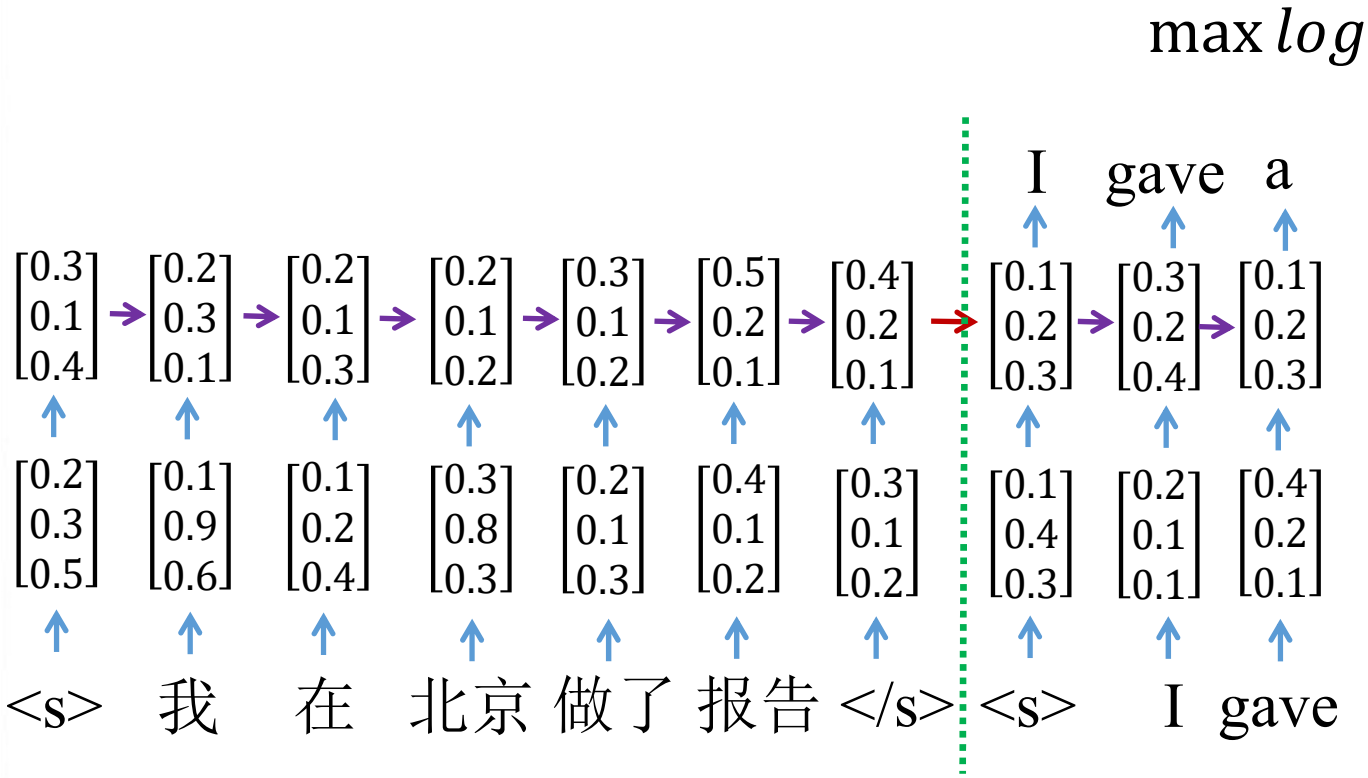
最大化 $P(target|source)$

神经机器翻译参数优化



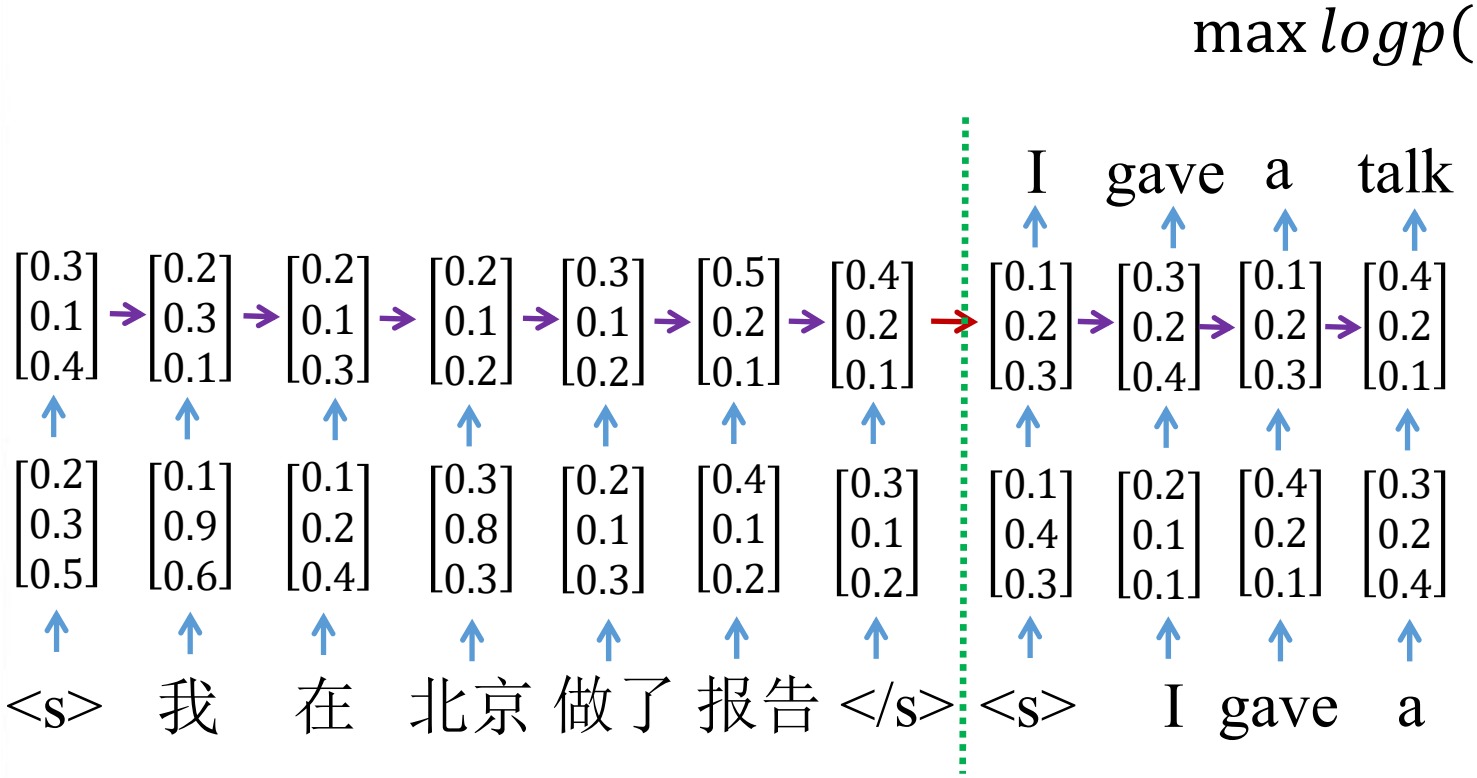
最大化 $P(\boldsymbol{target}|\boldsymbol{source})$

神经机器翻译参数优化



最大化 $P(\boldsymbol{target}|\boldsymbol{source})$

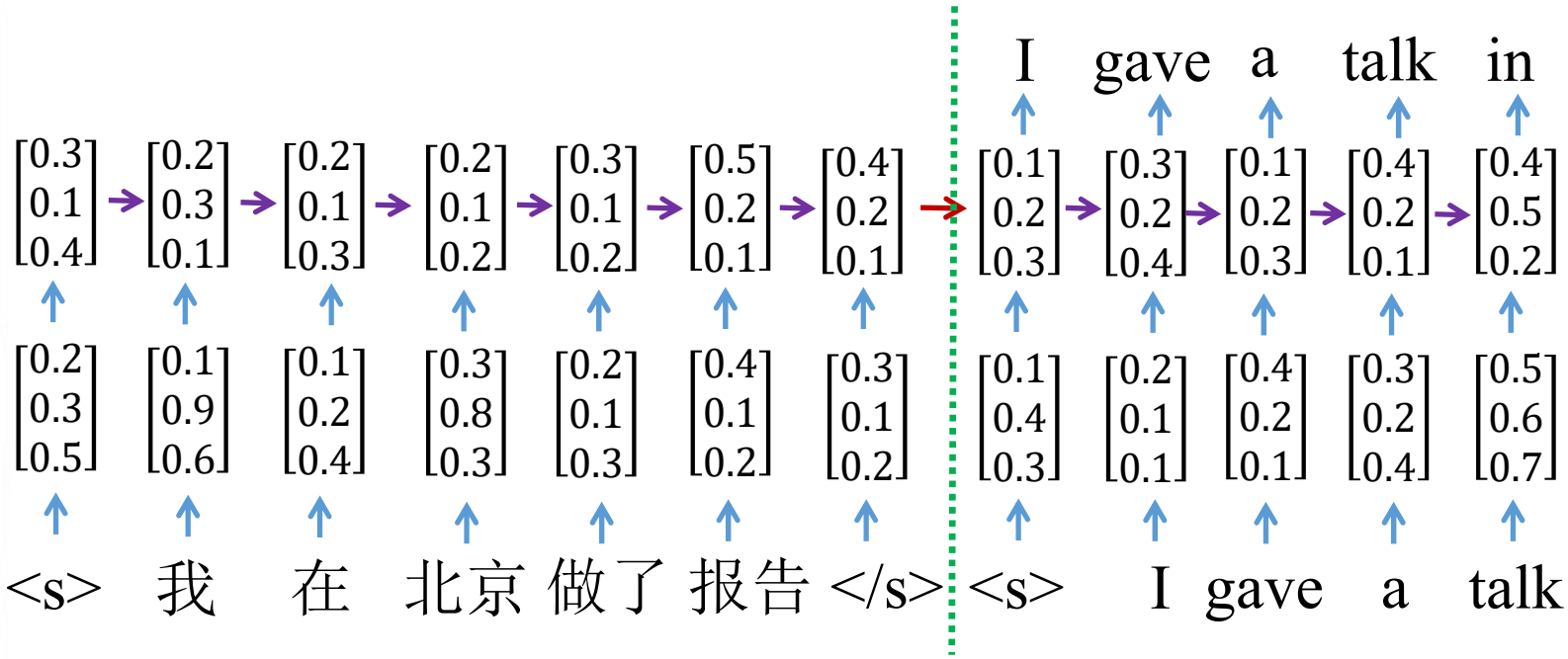
神经机器翻译参数优化



最大化 $P(\boldsymbol{target}|\boldsymbol{source})$

神经机器翻译参数优化

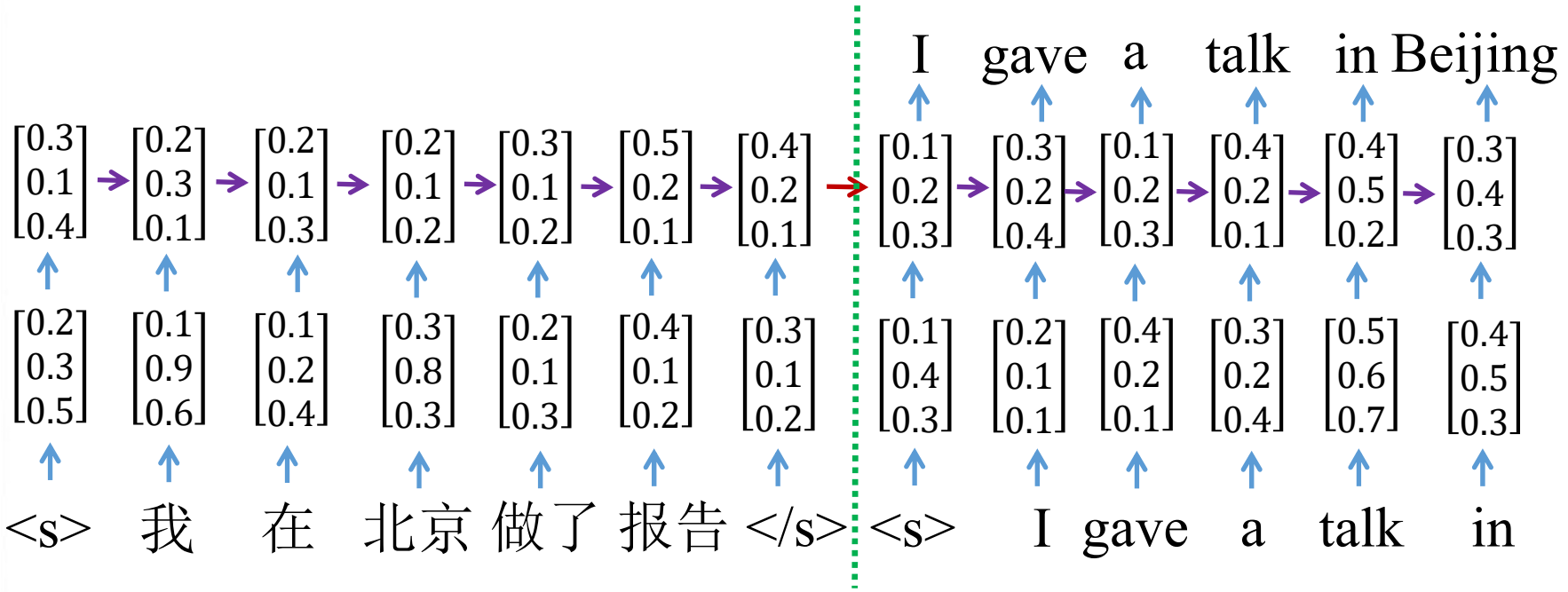
$$\max logp(in)$$



最大化 $P(\boldsymbol{target}|\boldsymbol{source})$

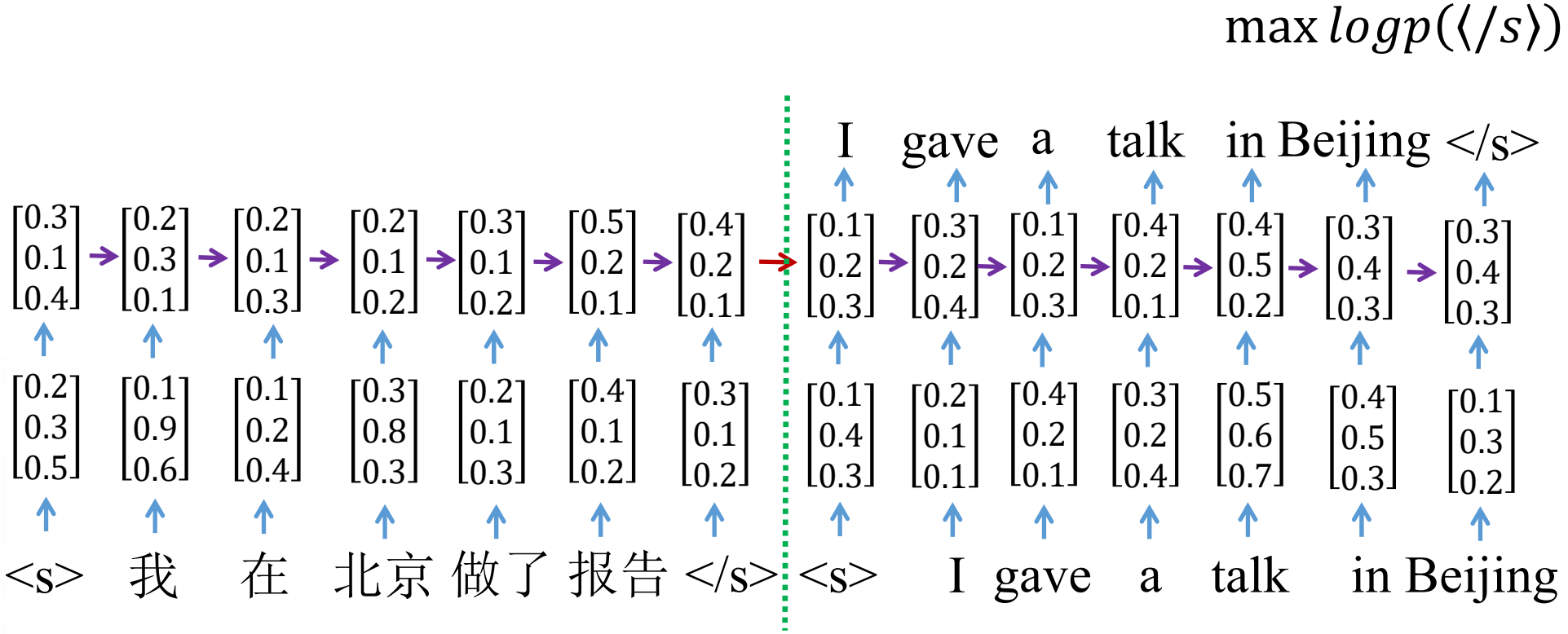
神经机器翻译参数优化

$$\max logp(Beijing)$$



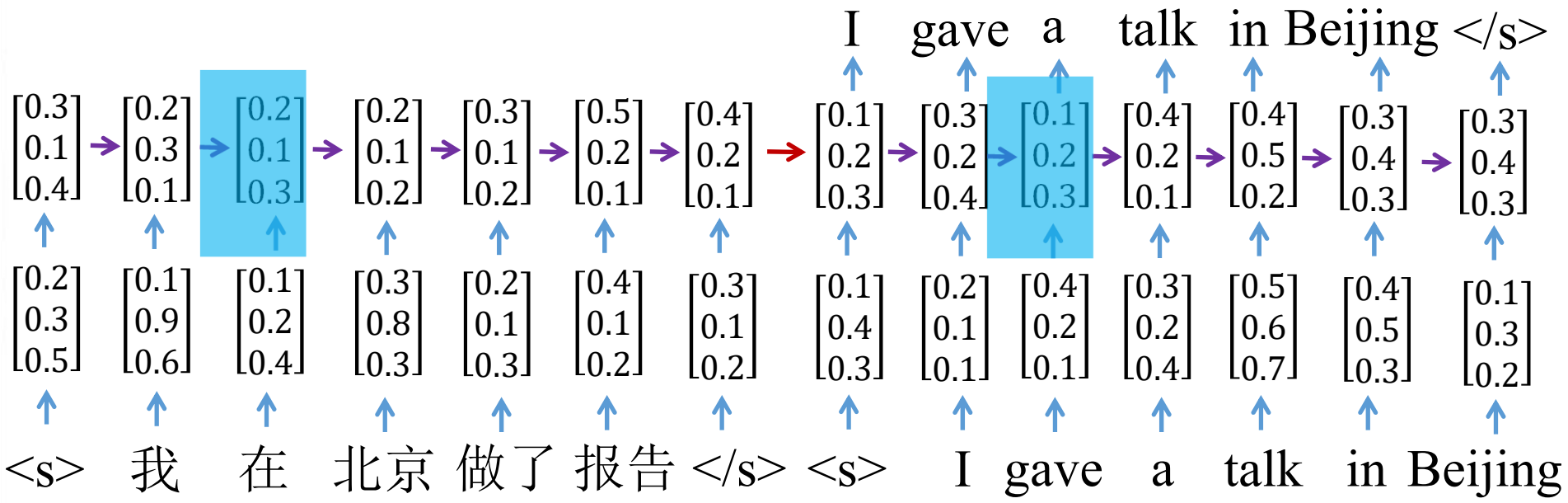
最大化 $P(\boldsymbol{target}|\boldsymbol{source})$

神经机器翻译参数优化

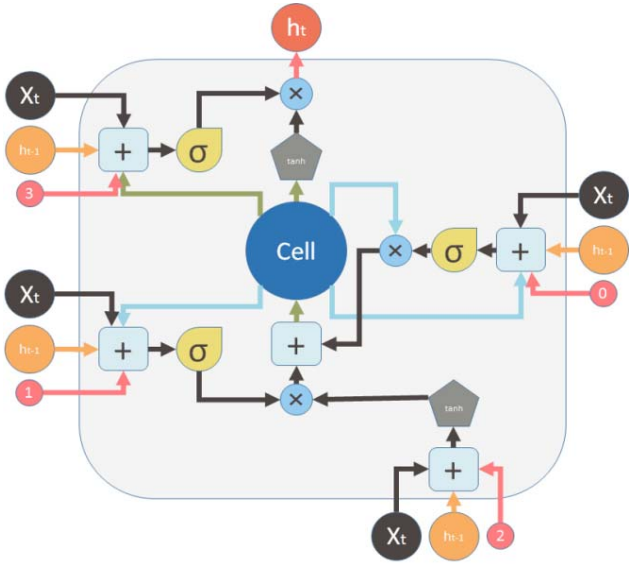

max $logp(\langle/s\rangle)$
I gave a talk in Beijing </s>
<s> 我 在 北京 做了 报告 </s>
<s> I gave a talk in Beijing</s>
最大化 $P(target|source)$

神经机器翻译-计算单元

$$h_t = tanh(U_S L(w_t) + W_S h_{t-1}) \qquad h_t = tanh(U_T L(w_t) + W_T h_{t-1})$$

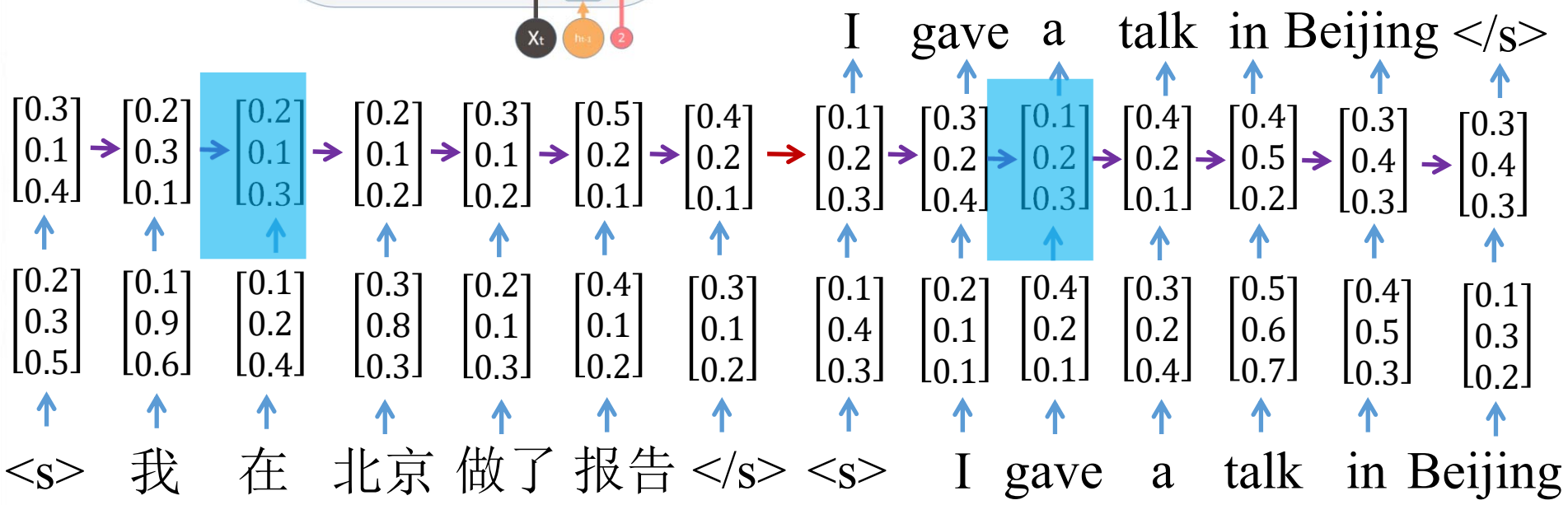


神经机器翻译-计算单元



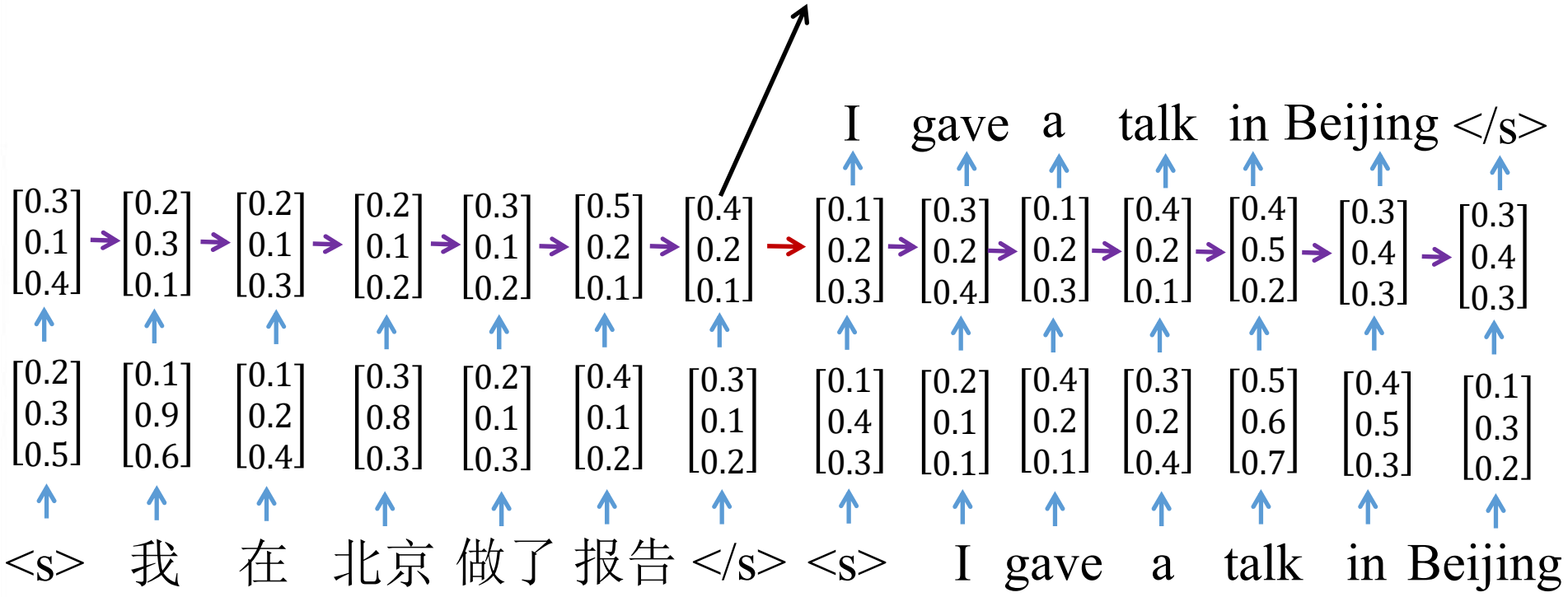
LSTM计算单元

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

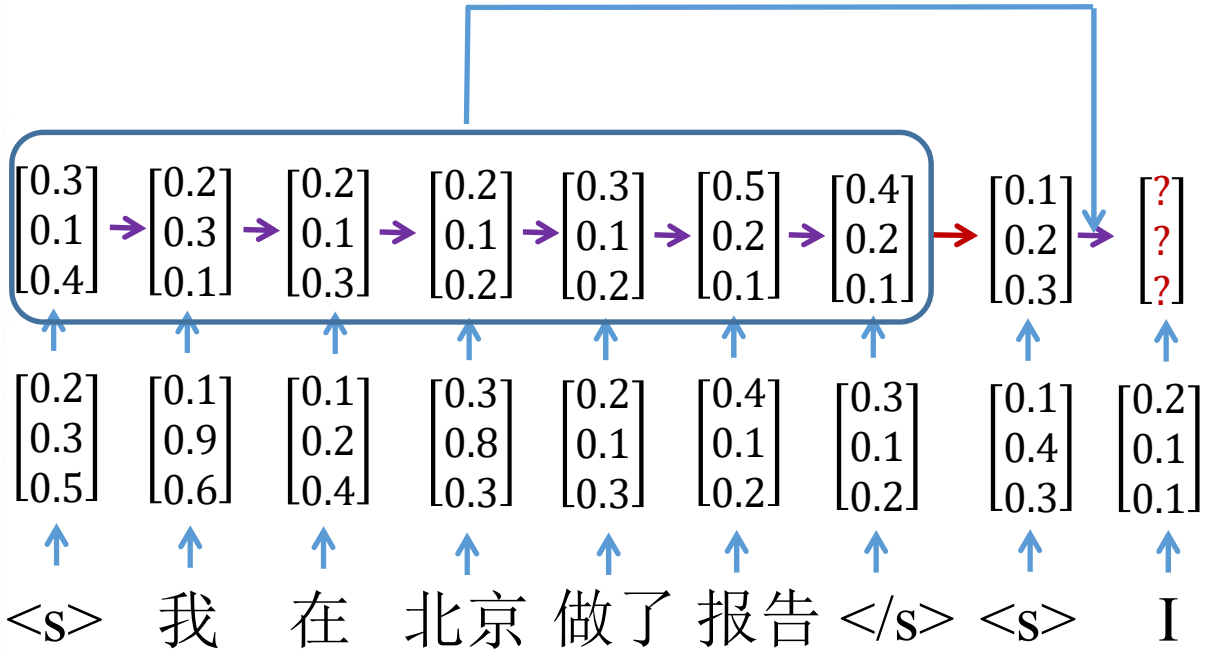


神经机器翻译-源端表示问题

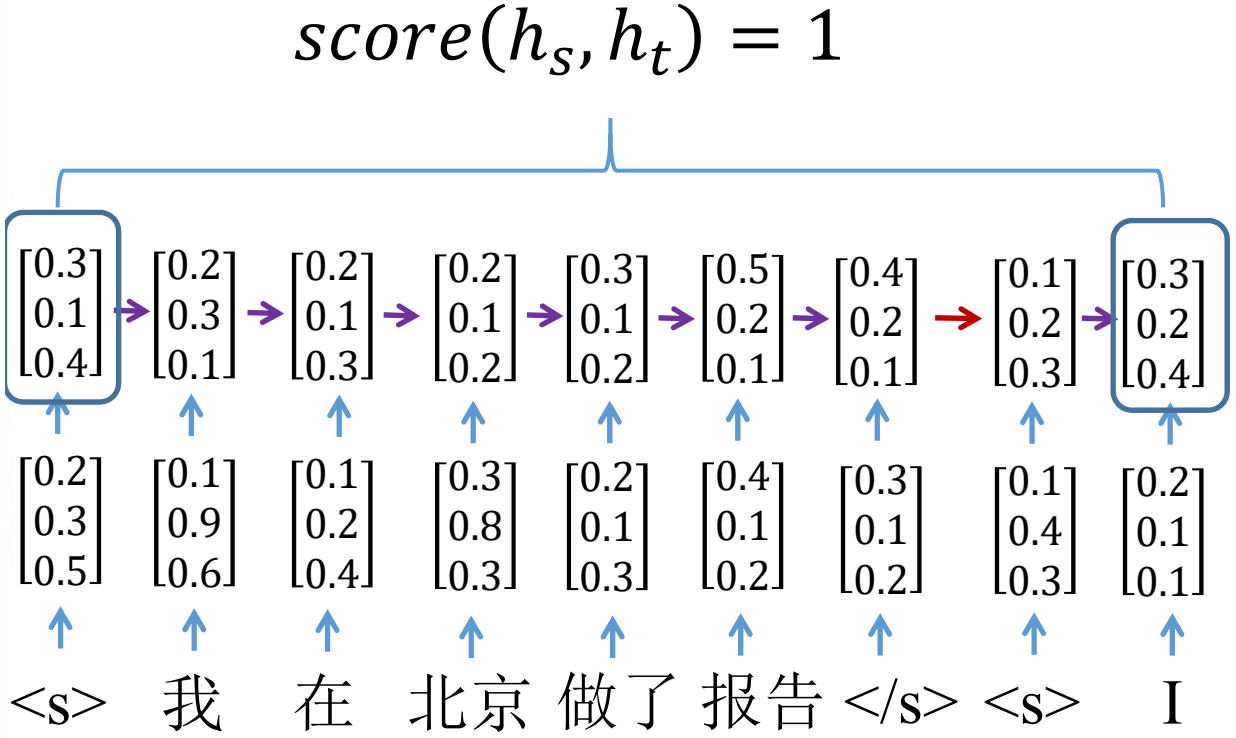
一个实数向量无法表示
源语言句子的完整语义



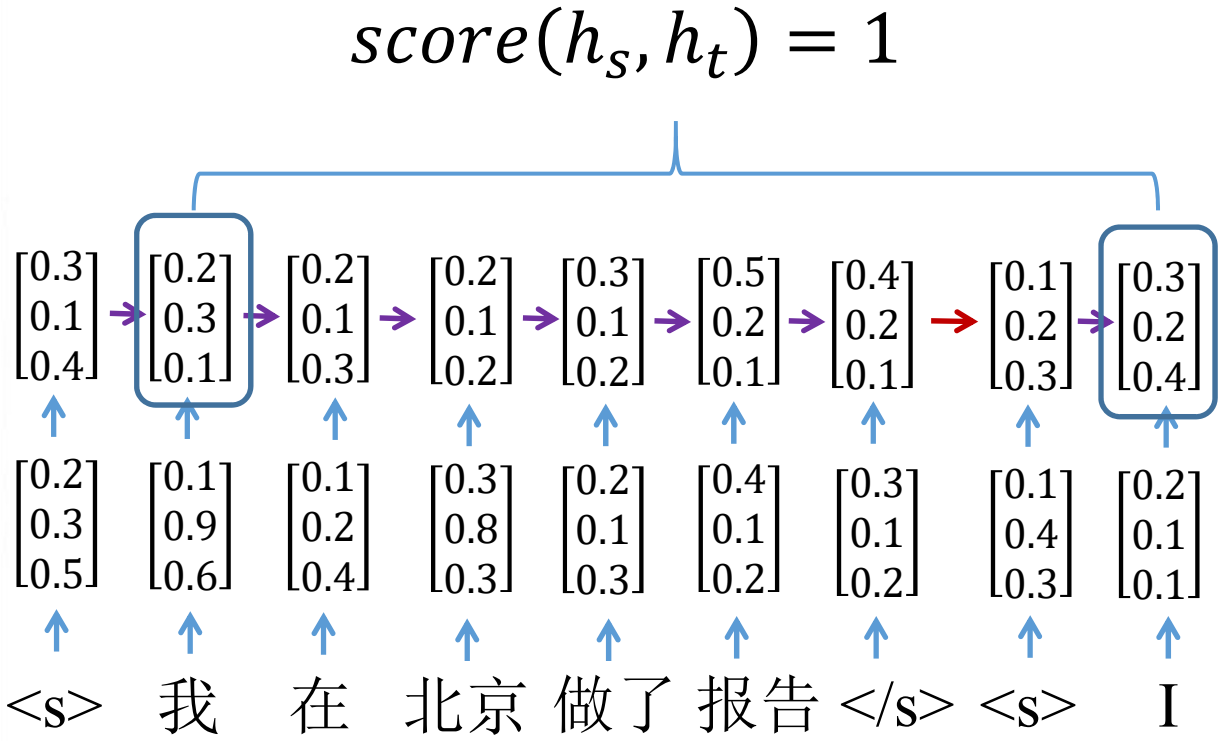
神经机器翻译-注意机制


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
0.1 0.4 0.3
0.2 0.1 0.1
<s> 我 在 北京 做了 报告 </s> <s> I

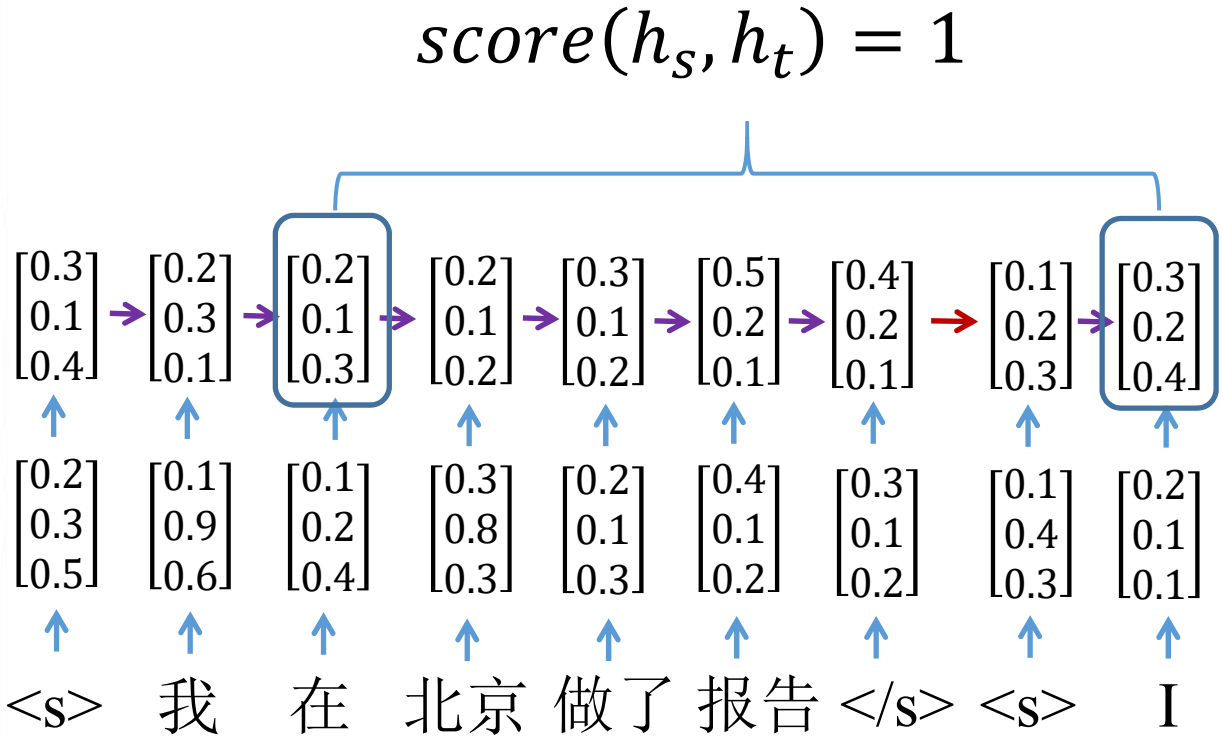
神经机器翻译-注意机制


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

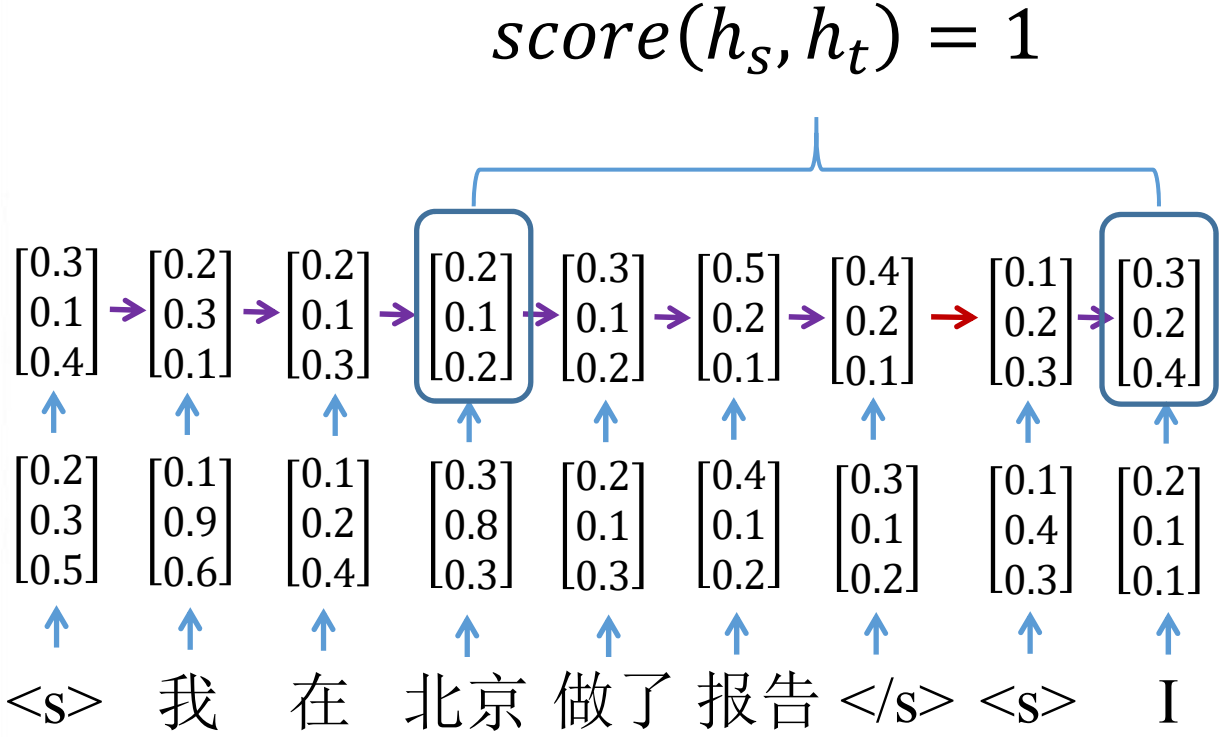
神经机器翻译-注意机制



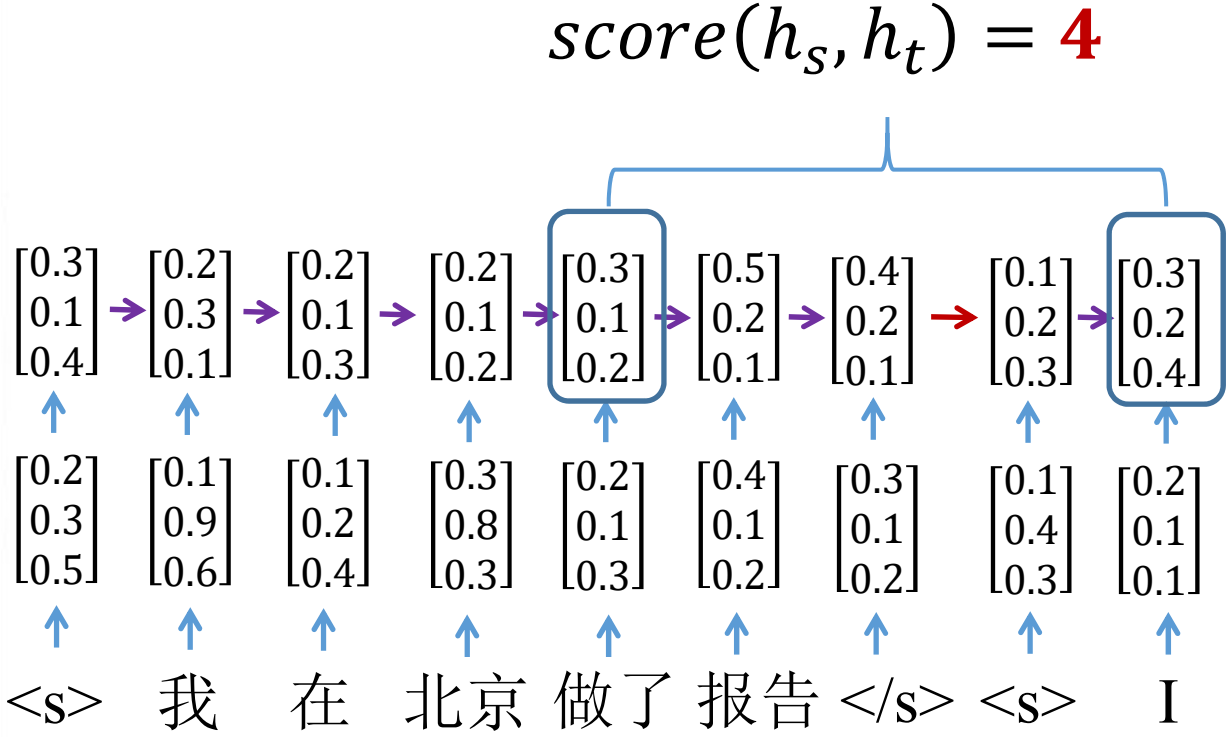
神经机器翻译-注意机制



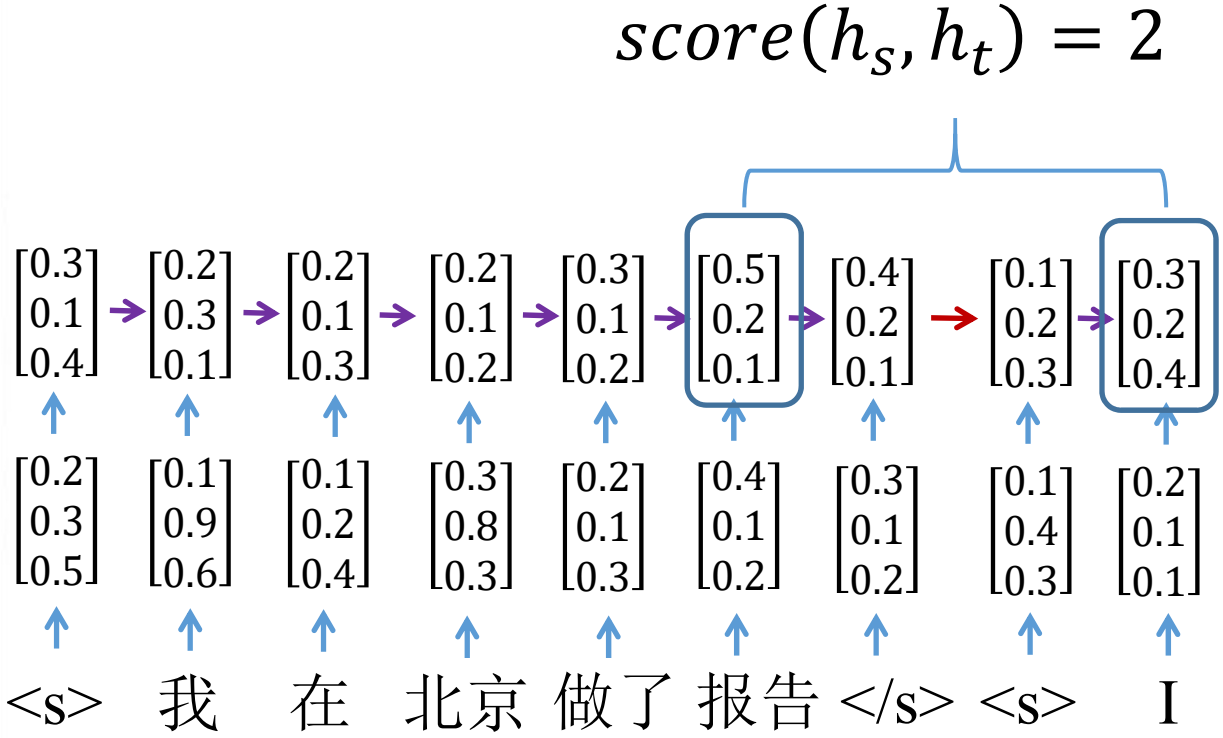
神经机器翻译-注意机制



神经机器翻译-注意机制

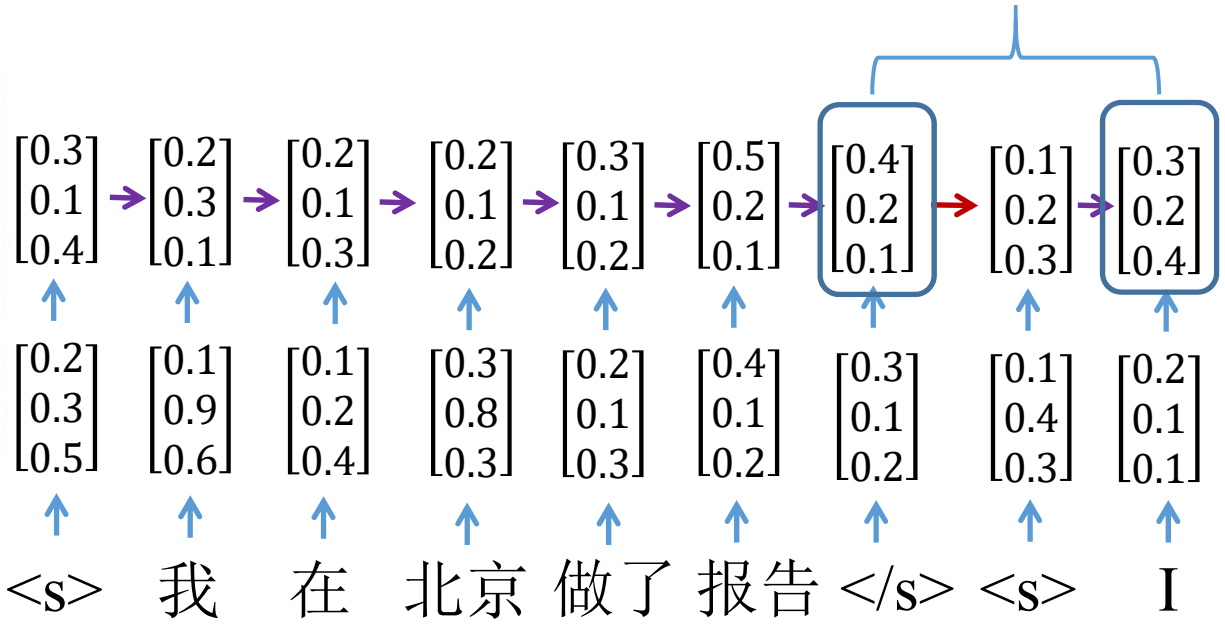


神经机器翻译-注意机制



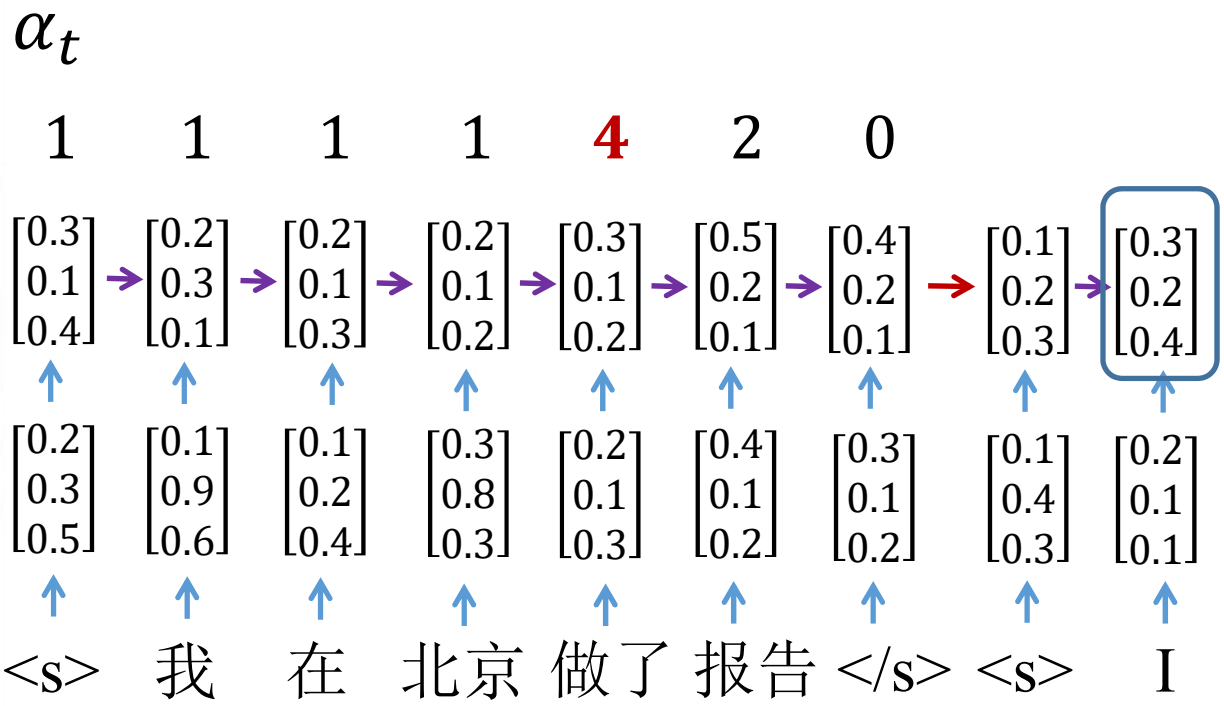
神经机器翻译-注意机制

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

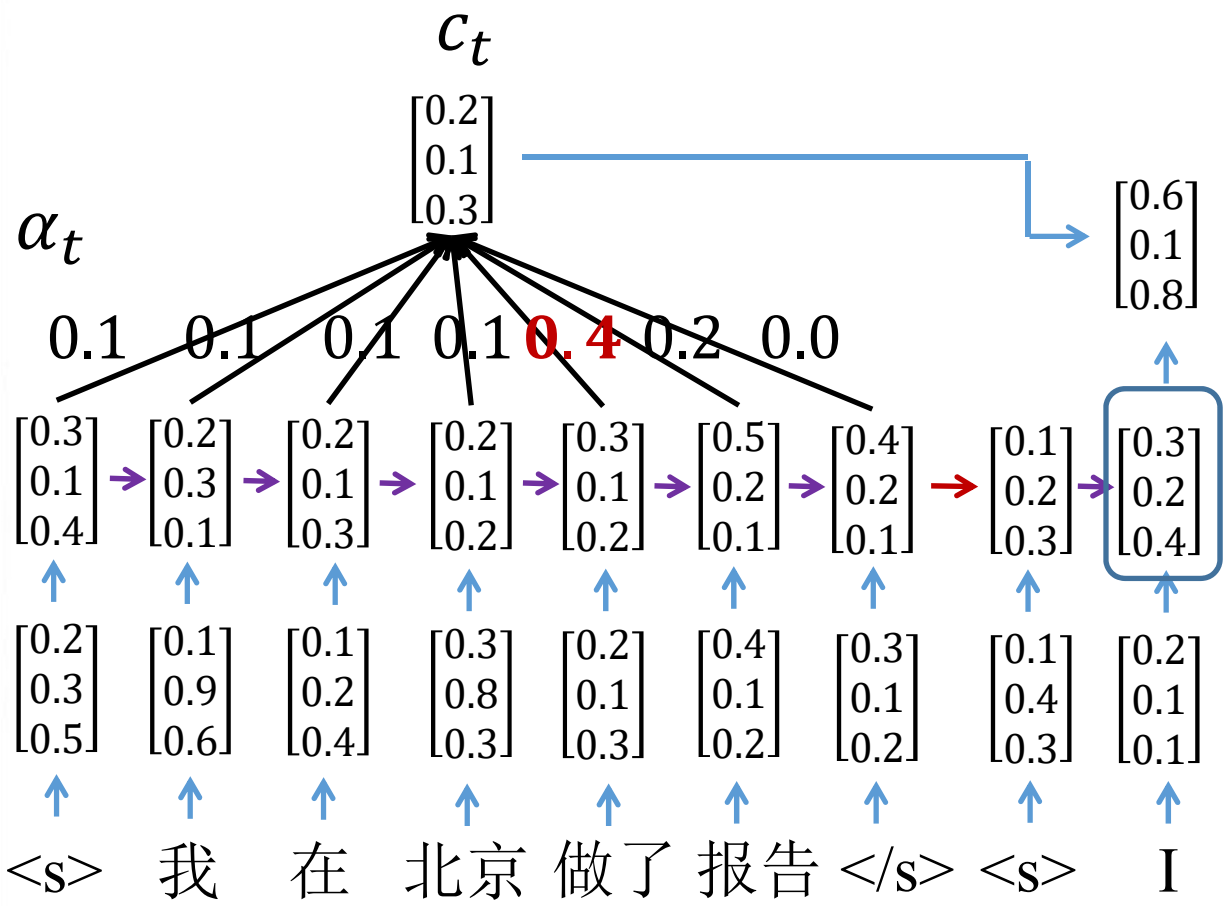


神经机器翻译-注意机制

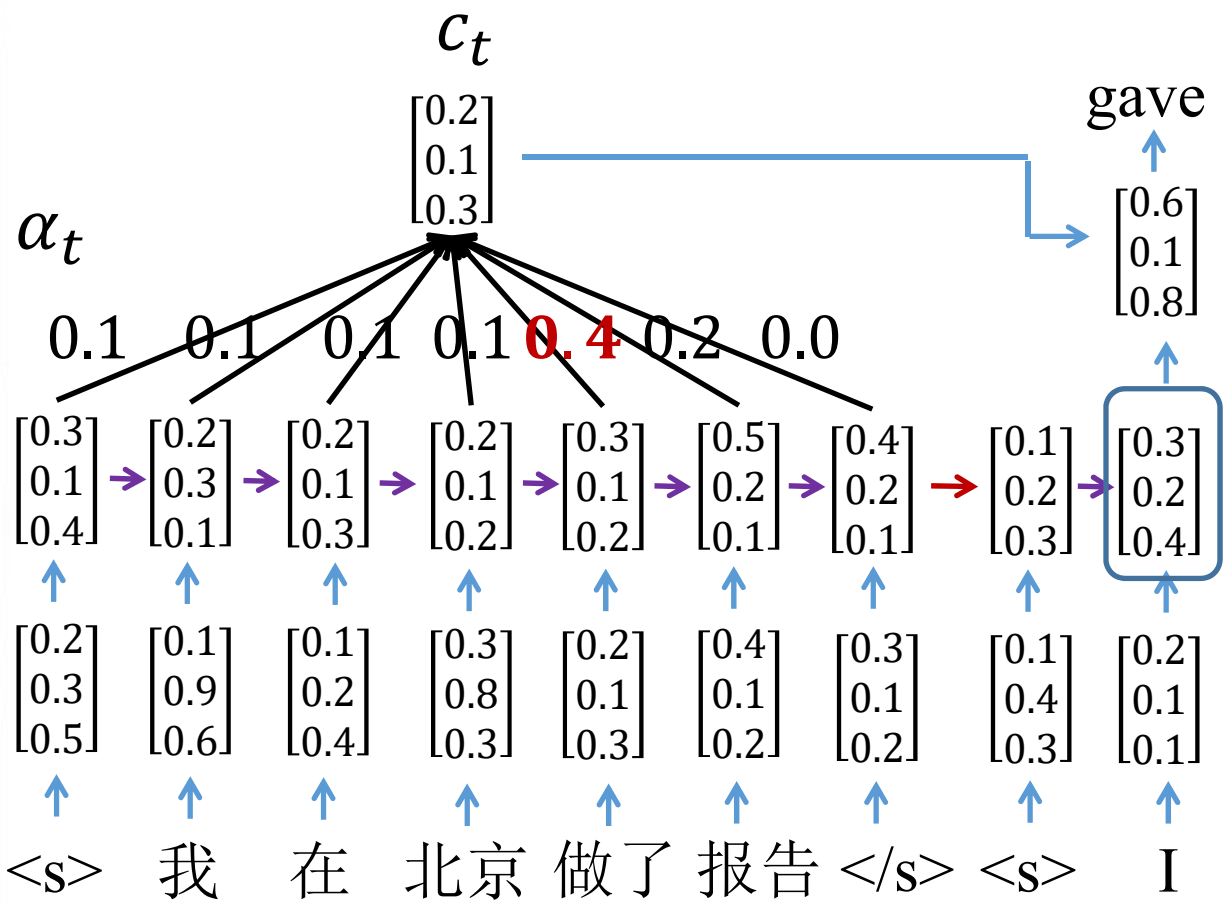
$$score(h_s, h_t)$$



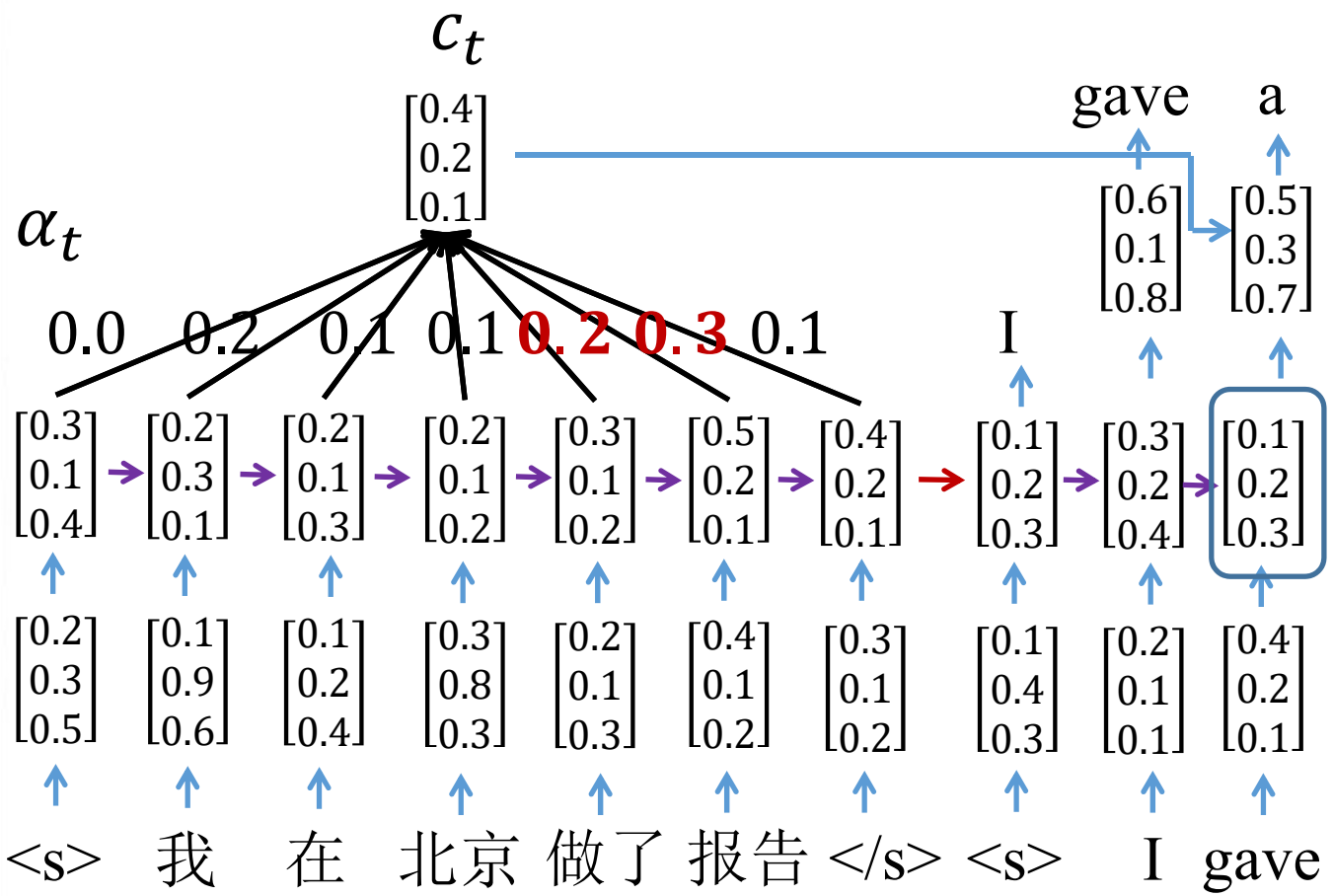
神经机器翻译-注意机制



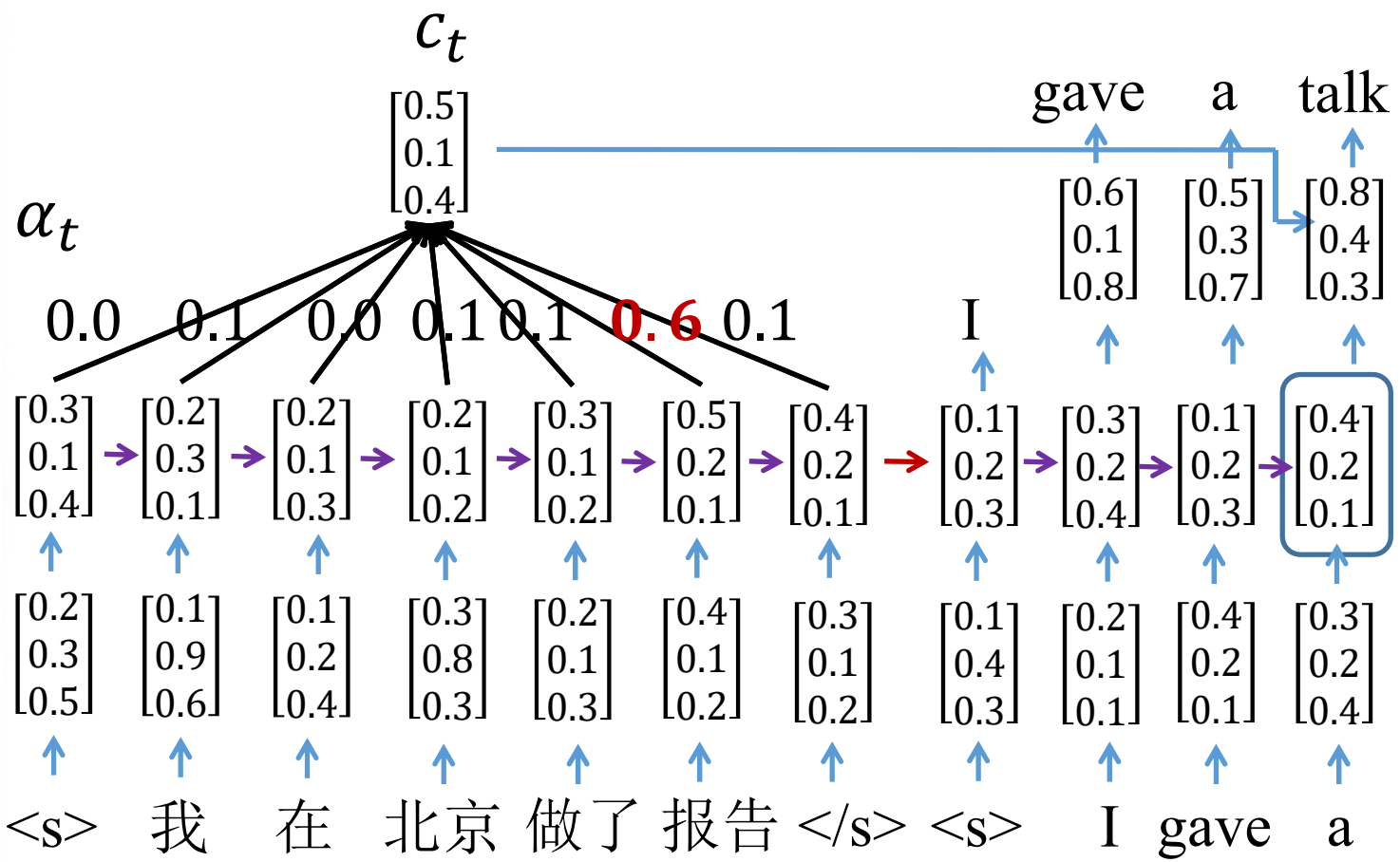
神经机器翻译-注意机制



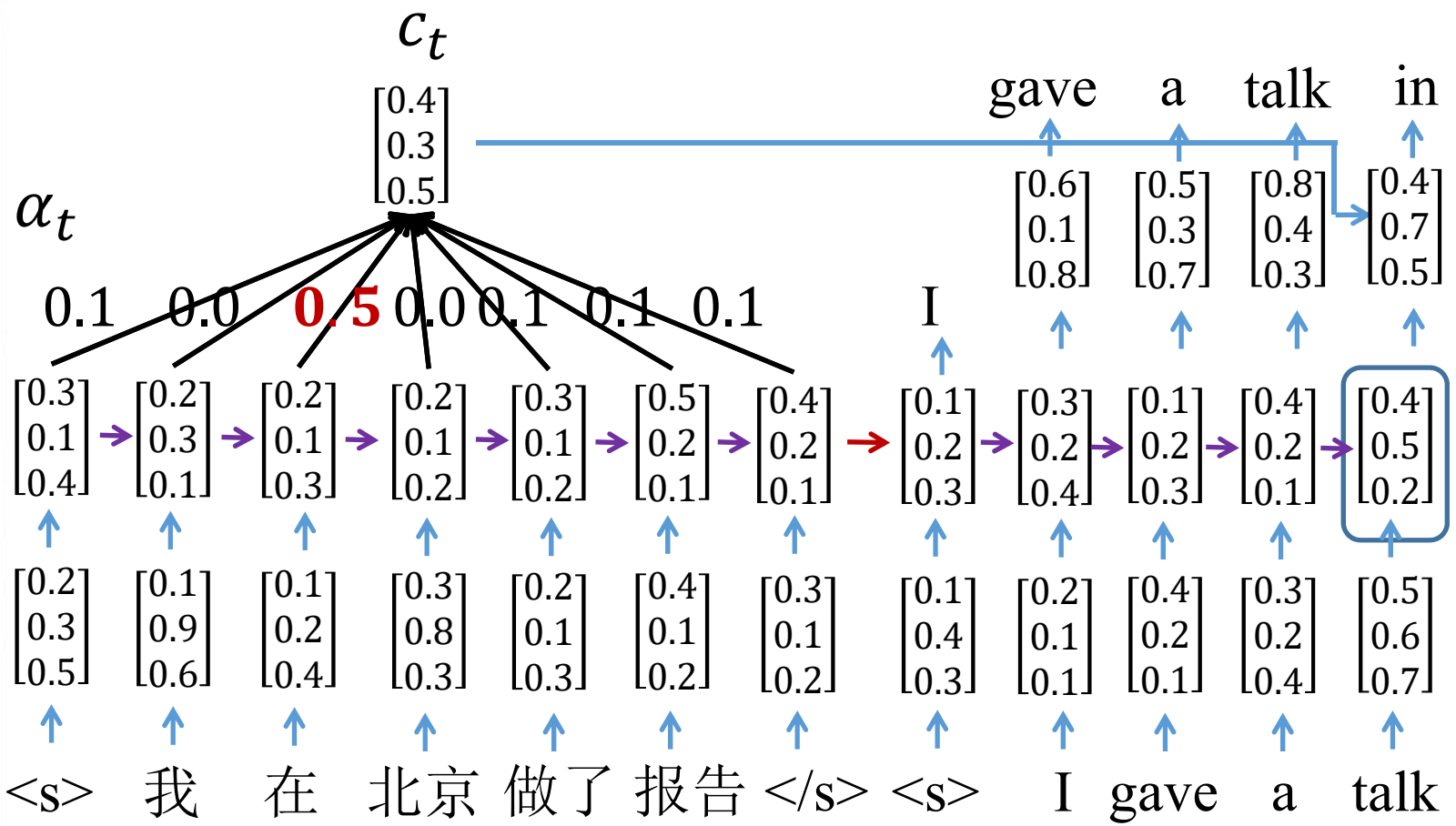
神经机器翻译-注意机制



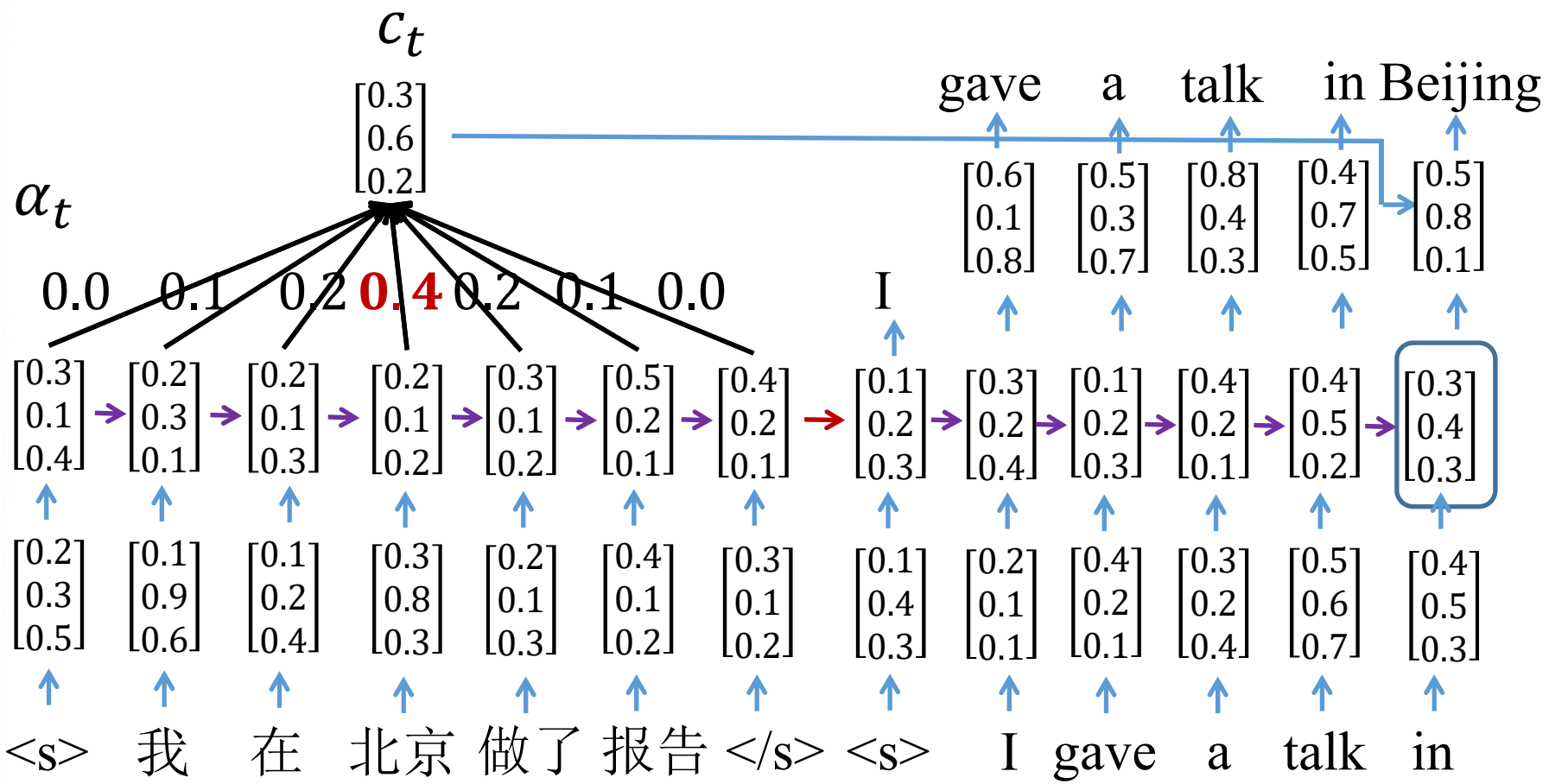
神经机器翻译-注意机制



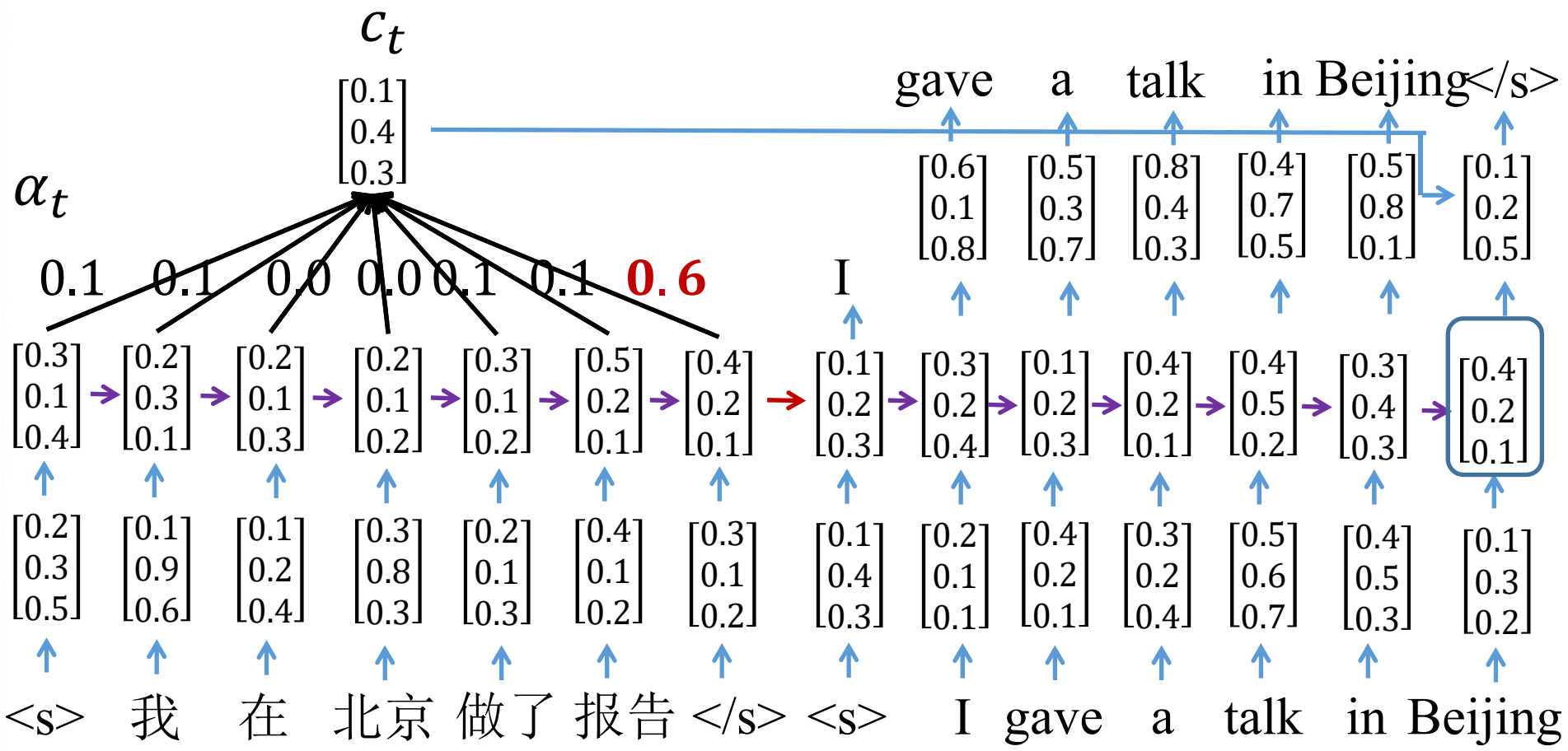
神经机器翻译-注意机制


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

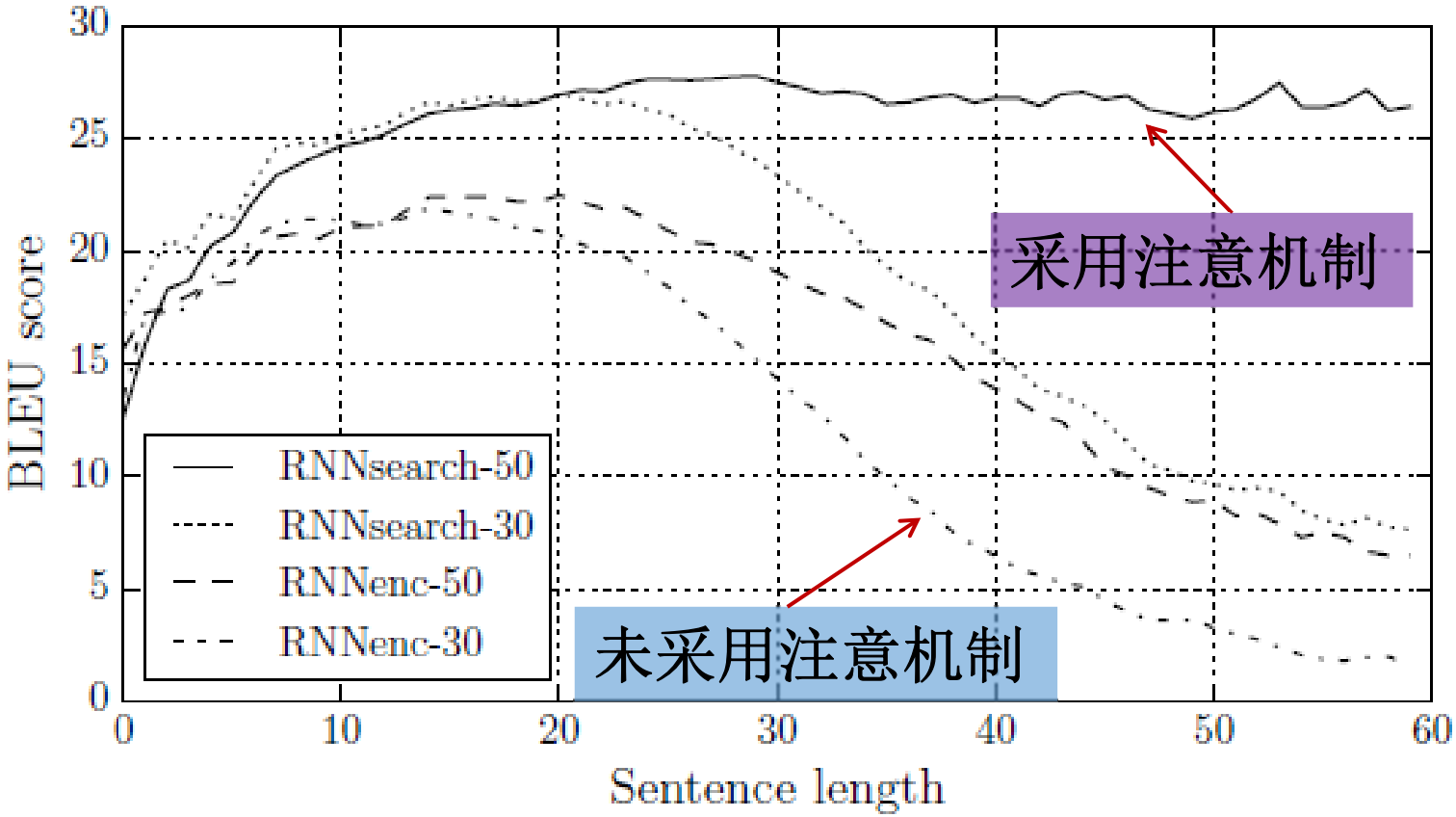
神经机器翻译-注意机制


c_t
α_t
0.0 0.1 0.2 0.4 0.2 0.1 0.0
gave a talk in Beijing
I
<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

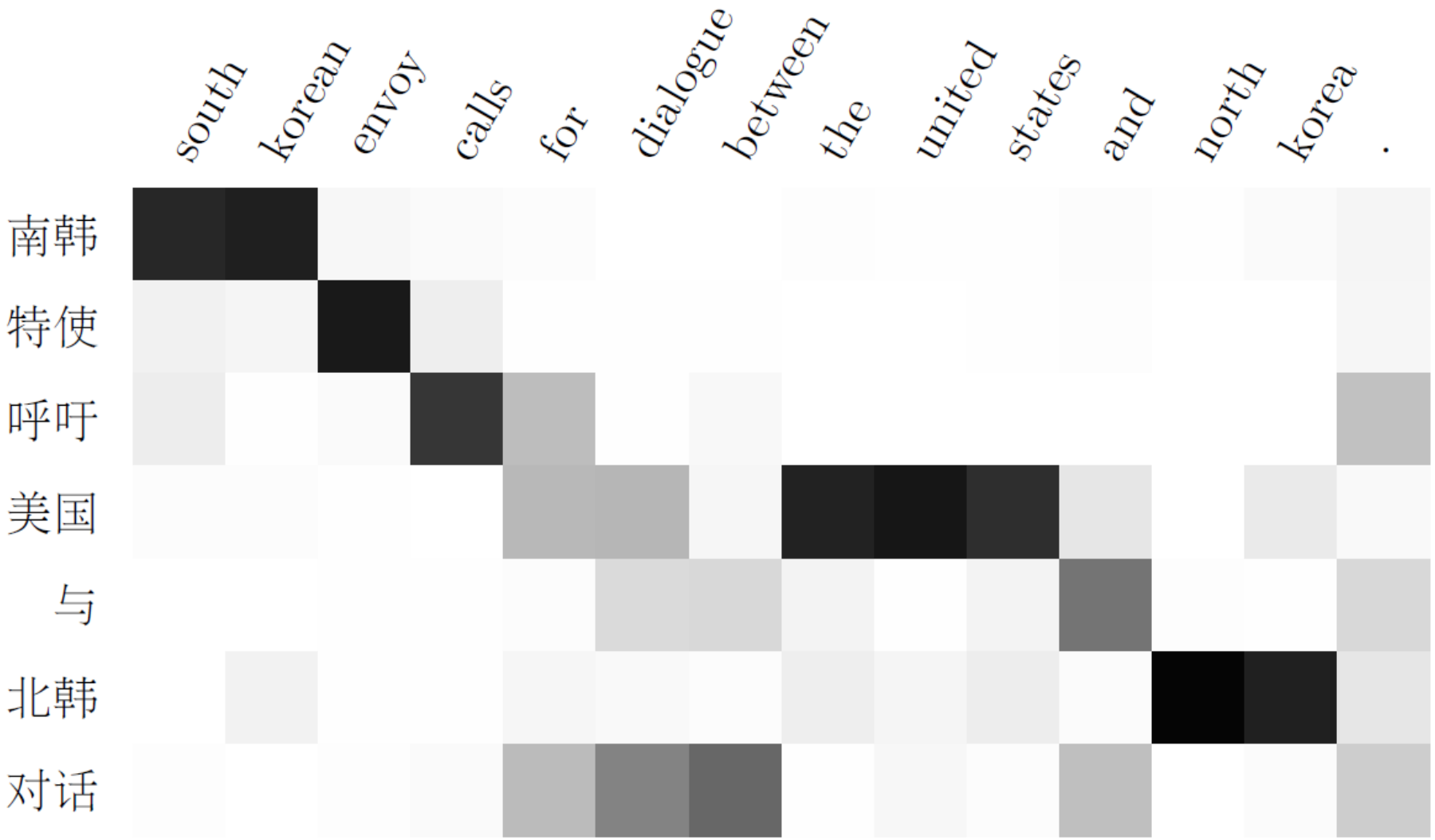

c_t
α_t
0.1 0.1 0.0 0.0 0.1 0.1 0.6
gave a talk in Beijing</s>
I
<s> 我 在 北京 做了 报告 </s> <s> I gave a talk in Beijing

神经机器翻译-注意机制

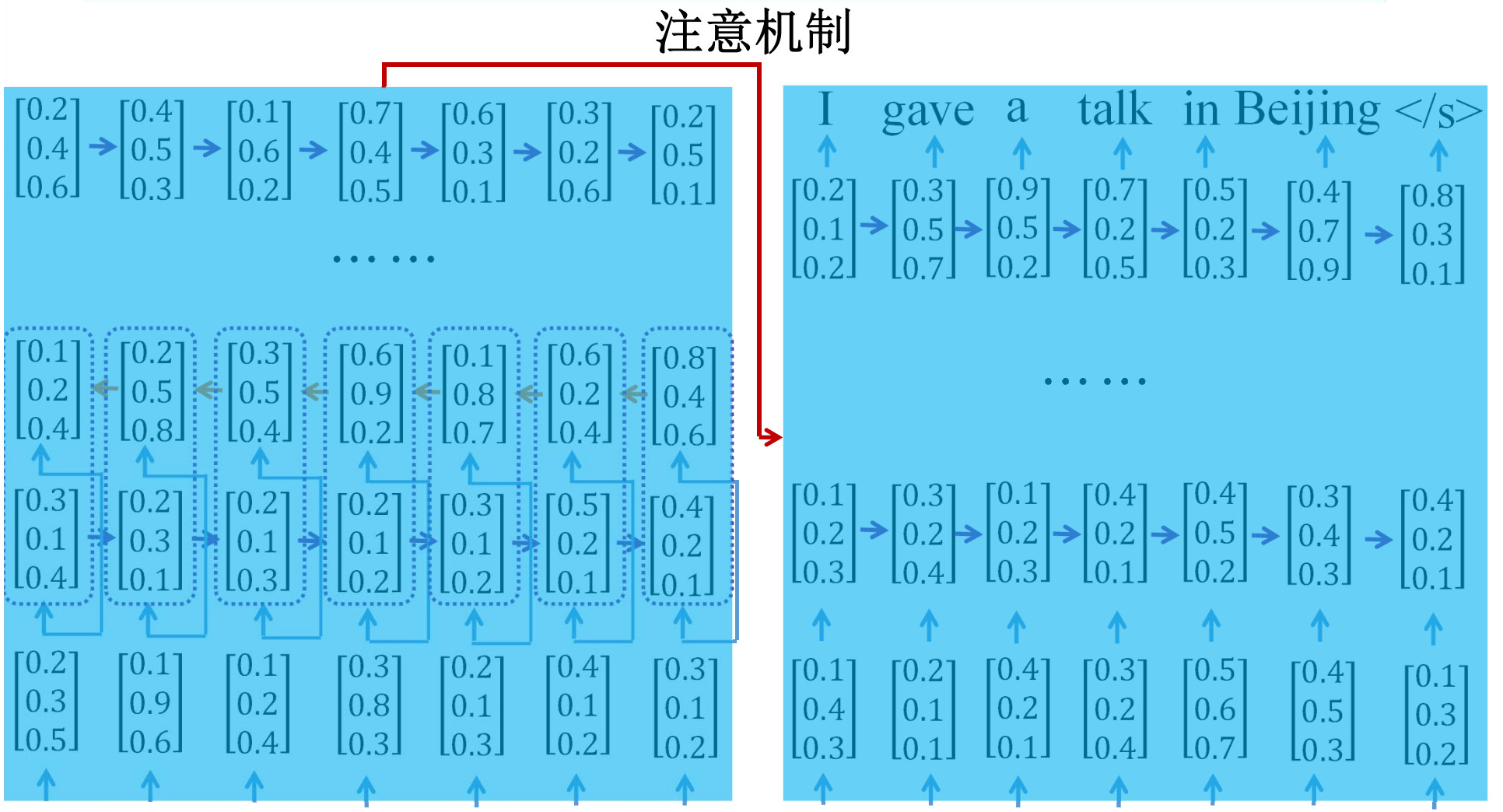


RNNenc: 无注意机制，**RNNsearch**: 采用注意机制

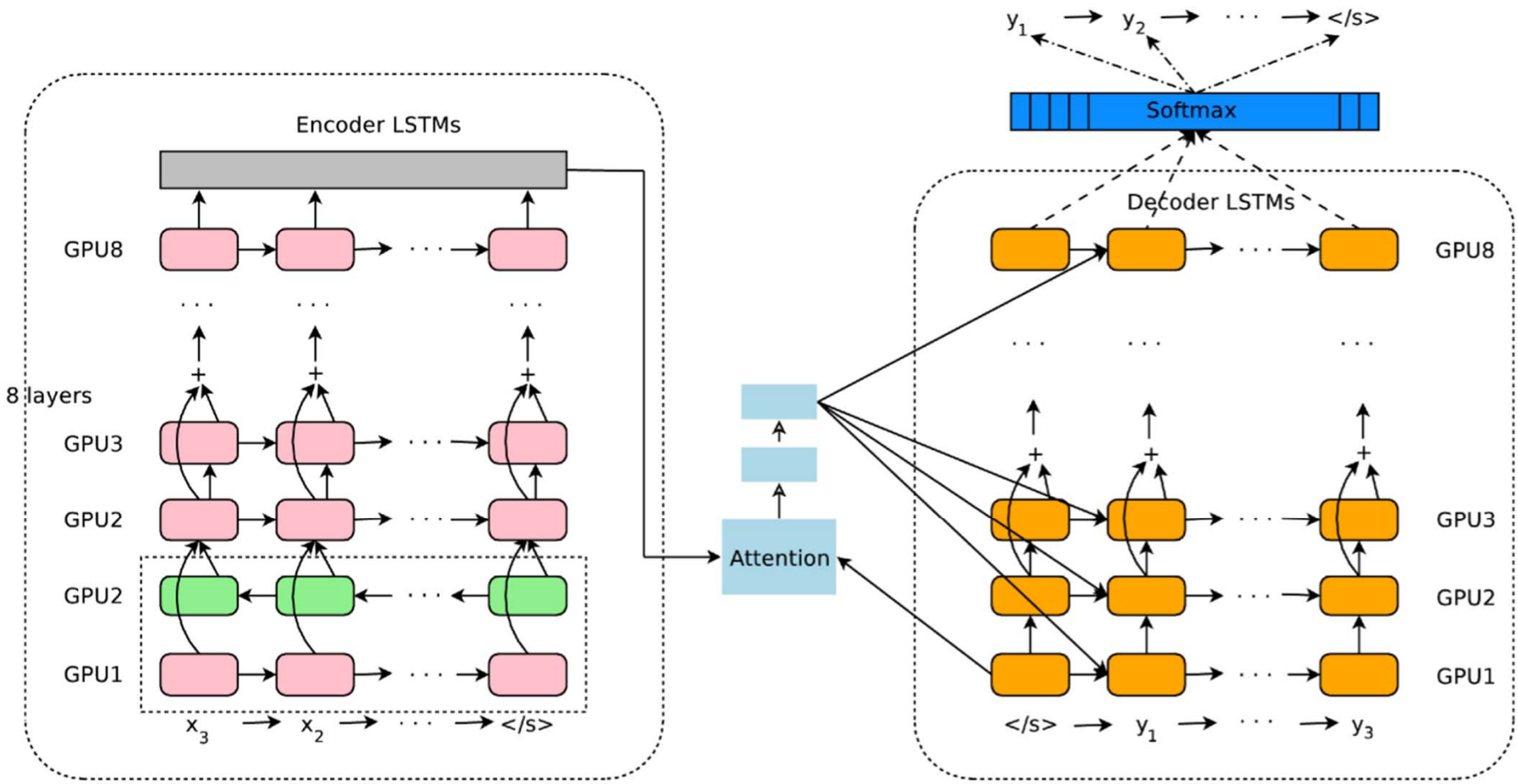
翻译实例


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

神经机器翻译-隐藏层结构



神经机器翻译-隐藏层结构



GNMT：谷歌神经翻译系统

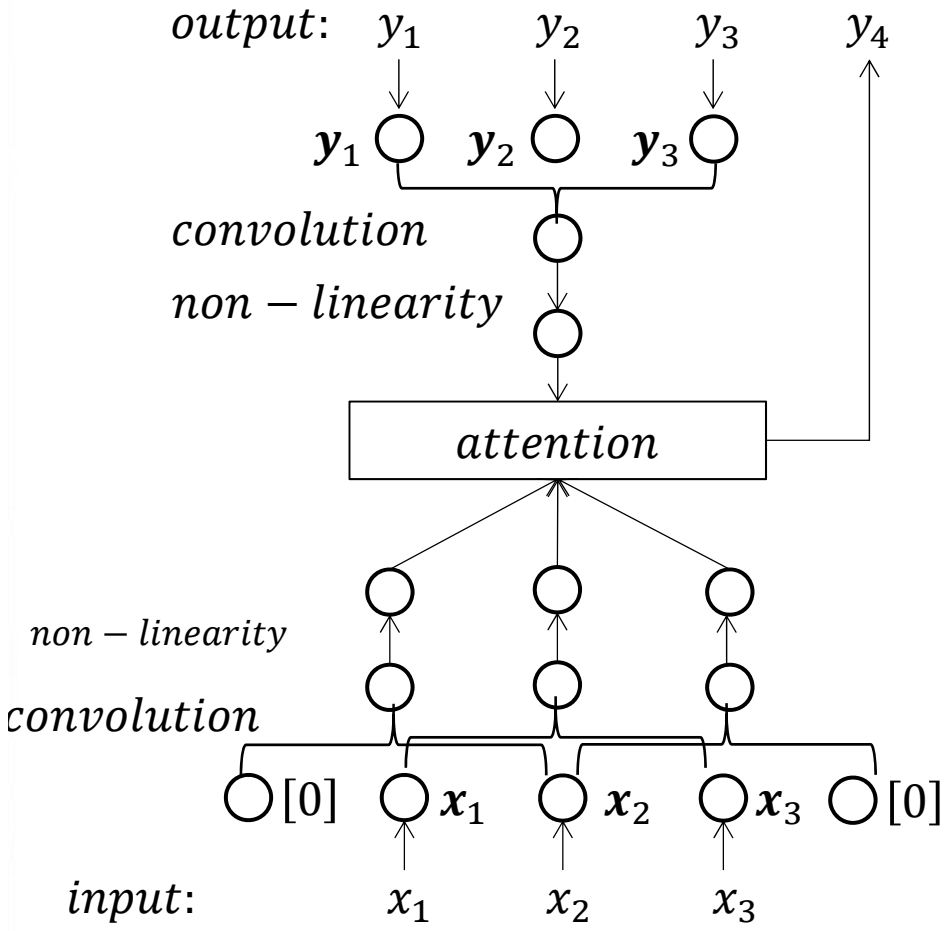
神经机器翻译 VS. 统计机器翻译

	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%

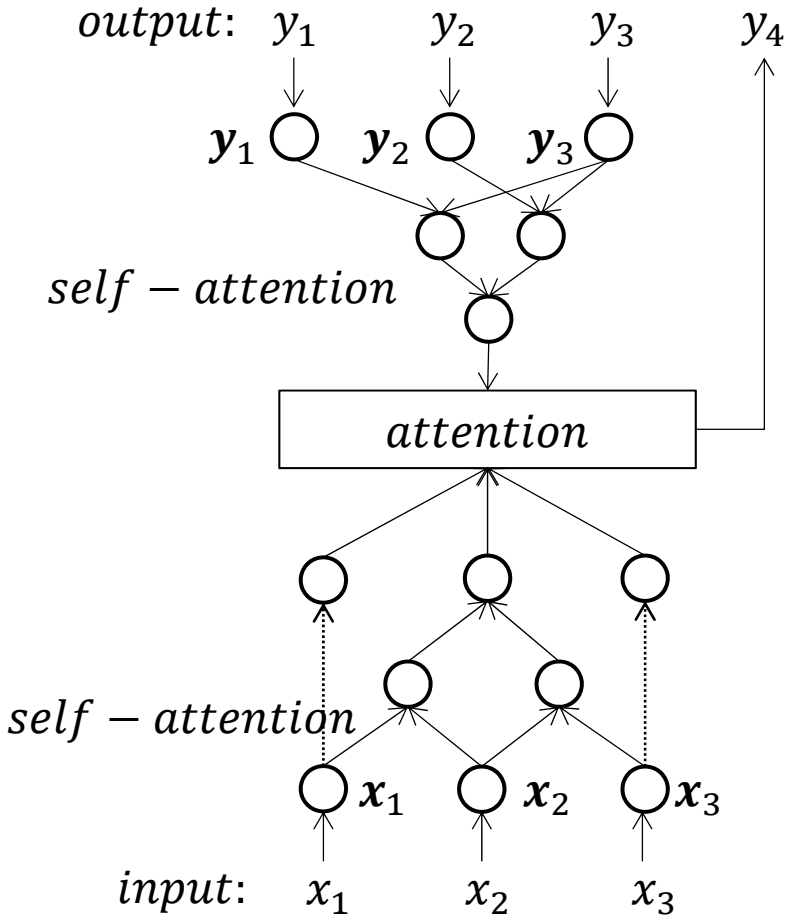
人工评测提升显著！

GNMT：谷歌神经翻译系统

其他编码-解码架构



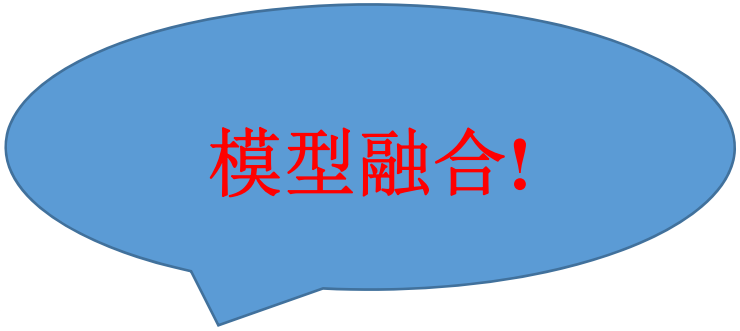
(a)基于卷积神经网络的翻译模型



(b)基于纯注意机制的翻译模型

两类机器翻译的结果对比

Source：	海珊 也 与 恐怖 组织网 建立 了 联系 。
Reference:	Hussein has also established ties with terrorist networks.
PBMT:	Hussein also has established relations and terrorist group .
HPMT:	Hussein also and terrorist group established relations .
NMT:	Hussein also established relations with <UNK> .



- 如何博采众长？

多粒度融合方法

- 词语级融合 (Arthur et al., 2016; Feng et al., 2017)
  - 以统计机器翻译中的词汇翻译概率为外部知识，影响神经机器翻译的译文选择和输出
- 短语级融合 (Tang et al., 2016; Wang et al., 2017; Zhao et al., 2018)
  - 将统计模型的短语翻译规则作为外部知识，影响和指导神经机器翻译的译文选择
- 句子级融合 (Niehues et al., 2016; Zhou et al., 2017)
  - 以统计机器翻译的译文和神经机器翻译的译文为多源输入，输出兼顾两者优势的译文

词汇级融合-NMT利用SMT词汇翻译表

- 动机：NMT对低频词的翻译存在大量错翻

  原因：NMT参数的“共享”机制（Feng et al., 2017）

  目的：减少参数个数，提高训练效果

  副作用：1、高频词有更多地机会去“优化”参数

  2、参数对高频词的预测较好，而对低频词则较差

例子：

Source:	阿尔卡特 宣称 第四 季 销售 成长 近 30%
Reference:	alcatel says sales in fourth quarter last year grew nearly 30 %
NMT:	he said sales grew nearly 30 percent in fourth quarter of last year

词汇级融合-NMT利用SMT词汇翻译表

- 为了缓解NMT的对低频词的错翻，在解码过程融合SMT的词汇翻译表（Arthur et al., 2016; Feng et al., 2017)
  - 原因： SMT的离散表示和非共享机制，使得SMT对低频单词的预测存在很大优势。例如：p(alcatel | 阿尔卡特) = 0.5732

  - Step 1：利用SMT得到一个词汇翻译表
  - Step 2：根据词汇翻译表以及输入句子X，得到如下矩阵

$$L=\begin{bmatrix} p_l(e=1|x_1) & \dots & p_l(e=1|x_{Tx}) \\ \dots & \dots & \dots \\ p_l(e=V_e|x_1) & \dots & p_l(e=V_e|x_{Tx}) \end{bmatrix}$$

列：源语言句子单词的个数

行：所有的目标语言单词的个数

词汇级融合-NMT利用SMT词汇翻译表

- Step 1：利用SMT得到一个词汇翻译表
- Step 2：根据词汇翻译表以及输入句子X，得到如下矩阵

$$L=\begin{bmatrix} p_l(e=1|x_1) & \dots & p_l(e=1|x_{Tx}) \\ \dots & \dots & \dots \\ p_l(e=V_e|x_1) & \dots & p_l(e=V_e|x_{Tx}) \end{bmatrix}$$

列：源语言句子单词的个数
行：所有的目标语言单词的个数

假设：源语言为“我 爱 中国”
目标语言一共5个单词：<start>、<end>、 I、 love、 china

$$L=\begin{bmatrix} p_l(start|我) & p_l(start|爱) & p_l(e=start|中国) \\ p_l(end|我) & p_l(end|爱) & p_l(e=end|中国) \\ p_l(I|我) & p_l(I|爱) & p_l(e=I|中国) \\ p_l(love|我) & p_l(love|爱) & p_l(e=love|中国) \\ p_l(china|我) & p_l(china|爱) & p_l(e=china|中国) \end{bmatrix}=\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

词汇级融合-NMT利用SMT词汇翻译表

➢Step 3：将矩阵L与attention weight相乘，得到如下SMT的预测结果

$$P_{SMT} = L * a = \begin{bmatrix} p_l(e=1|x_1) & \dots & p_l(e=1|x_{Tx}) \\ \dots & \dots & \dots \\ p_l(e=V_e|x_1) & \dots & p_l(e=V_e|x_{Tx}) \end{bmatrix} \begin{bmatrix} a_{i,1} \\ \dots \\ a_{i,Tx} \end{bmatrix}$$

假设：预测第一个目标单词时，attention weight为 $a_{11} = 0.9$, $a_{12} = 0.05$, $a_{13} = 0.05$

$$P_{SMT} = L * a = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0.9 \\ 0.05 \\ 0.05 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.9 \\ 0.05 \\ 0.05 \end{bmatrix} \begin{matrix} \text{start} \\ \text{end} \\ \text{i} \\ \text{love} \\ \text{china} \end{matrix}$$

词汇级融合-NMT利用SMT词汇翻译表

- Step 4：融合SMT的结果和NMT的结果

$$P_f = \lambda P_{NMT} + (1-\lambda) P_{SMT}$$

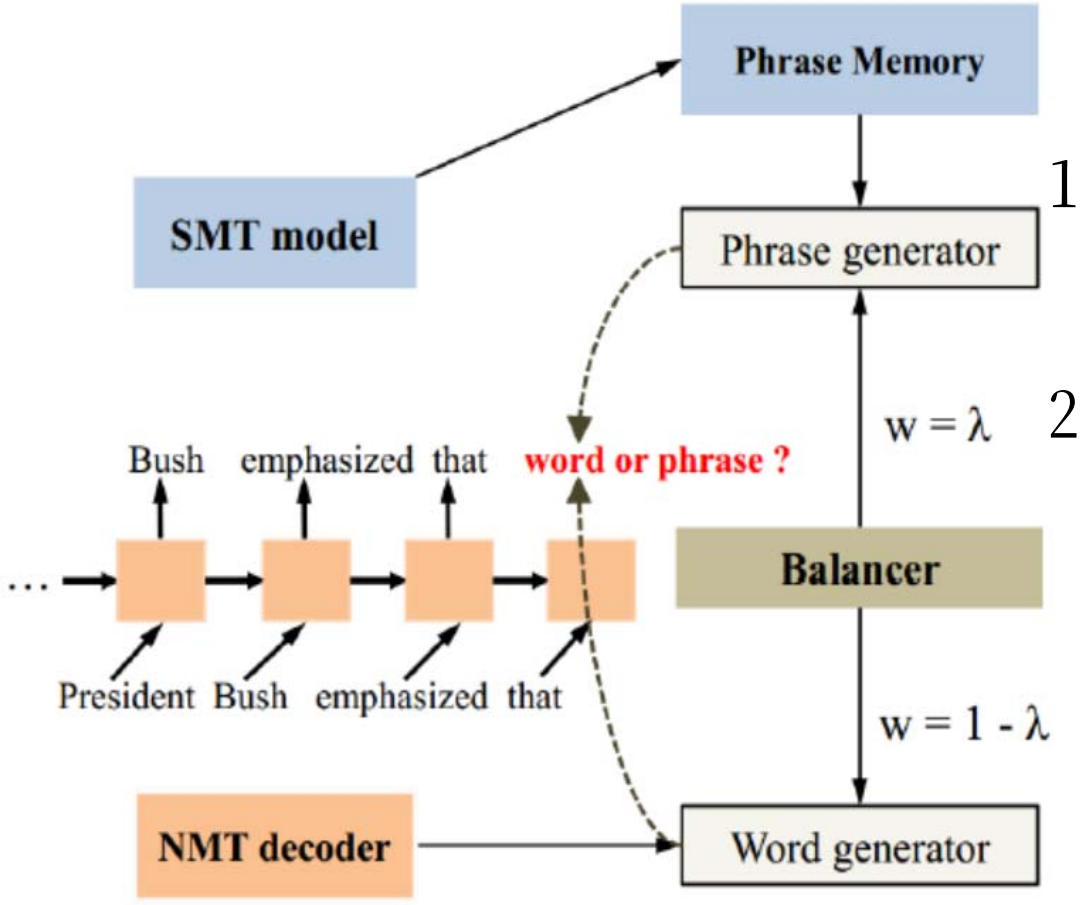
分析：SMT的预测结果只与两个因素有关

1）attention weight:翻译系统正在翻译哪个单词

2）翻译概率L:单词翻译概率越大，SMT的预测结果越高

短语级融合-NMT利用SMT短语翻译表

方法1.直接将短语作为翻译单元（Tang et al., 2016; Wang et al., 2017）



1. 通过权重调节应该是翻译单词但是翻译短语
2. 权重在网络中自动学习

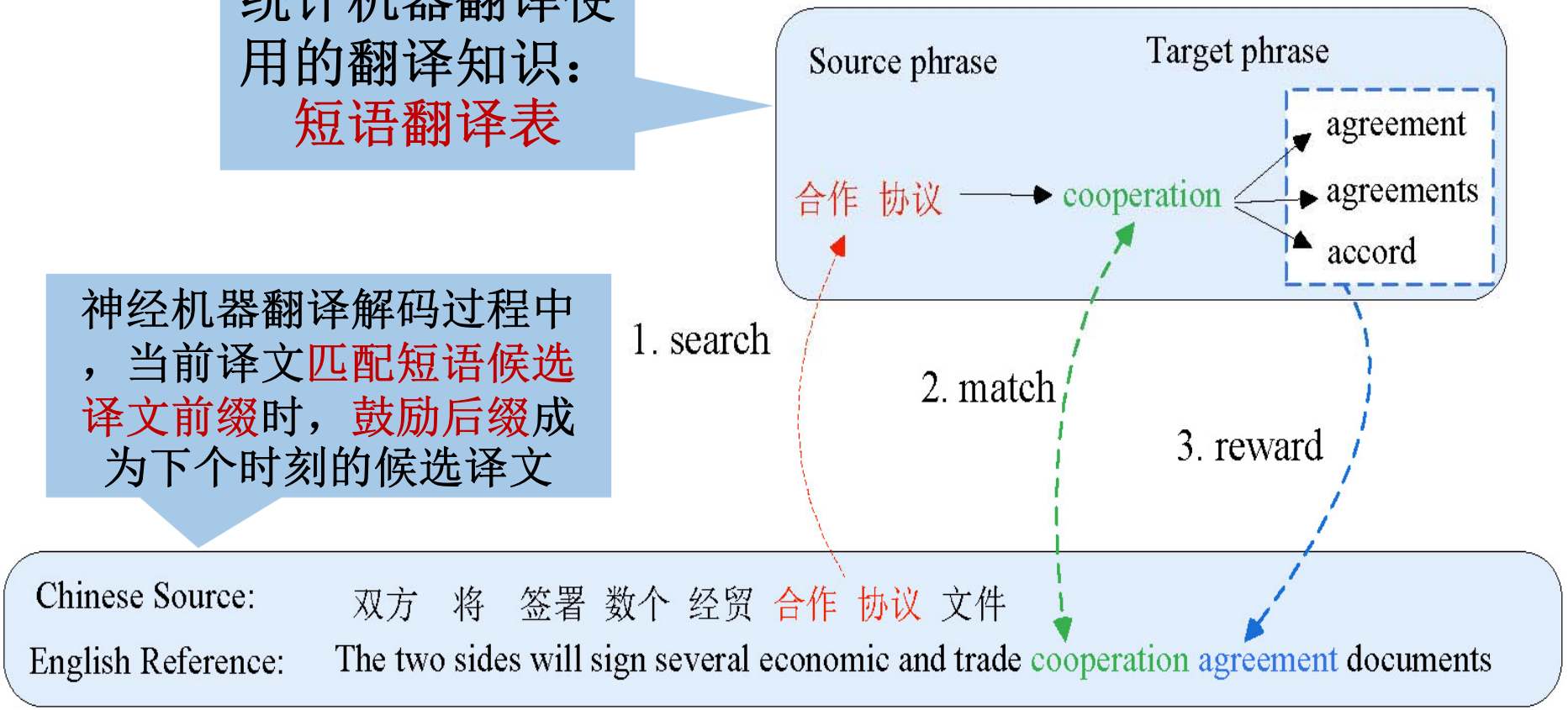
短语级融合-NMT利用SMT短语翻译表

方法2：短语与部分翻译结果进行匹配（Zhao et al. 2018）

基本思想：如果NMT当前译文的后缀能够匹配到某个短语翻译规则的目标语言部分的前缀，那么翻译规则目标语言短语匹配部分的下一个词极有可能就是NMT应该预测的词

短语级融合-NMT利用SMT短语翻译表

统计机器翻译使用的翻译知识：
短语翻译表



短语级融合-NMT利用SMT短语翻译表

$$p(y_i|y_{<i}, C) = p(y_i|y_{<i}, c_i)$$
$$= softmax\left(score\left(W_{y_i}, h_i\right)\right)$$

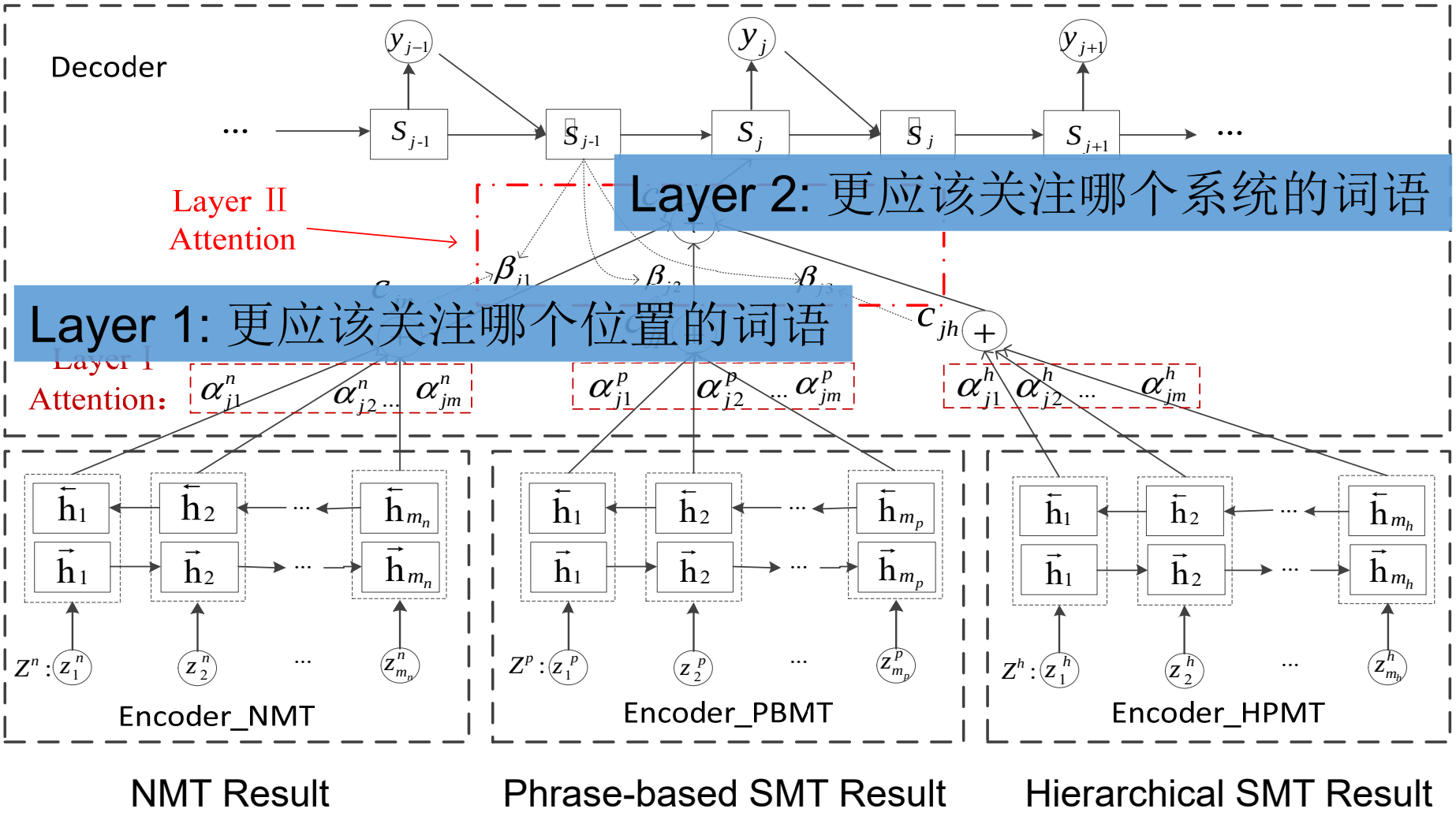
$$p(y_i|y_{<i}, C) = p(y_i|y_{<i}, c_i)$$
$$= softmax\left(\left(1 + \lambda V(R_i)\right)score\left(W_{y_i}, h_i\right)\right)$$

短语级融合-NMT利用SMT短语翻译表

#	Method	CH-EN						EN-JA	
		MT03(dev)	MT04	MT05	MT06	MT08	Ave	dev	test
1	Moses	28.35	30.02	29.10	32.92	23.20	28.72	20.06	22.40
2	Baseline	34.20	36.96	32.60	33.85	25.96	32.71	23.61	25.99
3	Arthur	34.98†	37.96†	33.36†	34.79†	26.53*	33.52	24.33*	26.72†
4	Our method	**36.48**†	**38.79**†	**35.34**†	**36.58**†	**27.49**†	**34.94**	**25.63**†	**27.95**†
5	system(no matching)	34.99†	37.54*	33.32†	34.22*	26.39*	33.29	24.11*	26.47*
6	system(no first)	35.25†	38.07†	34.13†	34.95†	26.67†	33.81	24.37†	26.93†

- 方法2比统计模型平均提升5个BLEU值，比神经网络系统平均提升2个BLEU值！

句子级融合模型 (Zhou et al., 2017)



层次注意模型

- **Layer 1:** 系统内的注意机制

第j输出与第k个系统中第i个隐层表示的注意权重

$$c_{jk} = \sum_{i=1}^{m} \alpha_{ji}^{k} h_i$$

第k个系统中第i个隐层表示

第k个系统第j个输出对应的上下文

$$\alpha_{ji}^{k} = \frac{exp(e_{ji})}{\sum_{l=1}^{m} exp(e_{jl})}$$

$$e_{ji} = v_a^T tanh(W_a \tilde{s}_{j-1} + U_a h_i)$$

层次注意模型

- **Layer 2:** 不同系统之间的注意机制

第j个输出与第k个系统之间的注意权重

$$c_j = \sum_{k=1}^{K} \beta_{jk} c_{jk}$$

第j个输出对应的上下文

第k个系统第j个输出对应的上下文

$$\beta_{jk} = \frac{exp(\tilde{s}_{j-1} \cdot c_{jk})}{\sum_{k'} exp(\tilde{s}_{j-1} \cdot c_{jk'})}$$

句子级融合模型的效果

System	MT03	MT04	MT05	MT06	Ave
PBMT	37.47	41.20	36.41	36.03	37.78
HPMT	**38.05**	**41.47**	**36.86**	36.04	**38.10**
NMT	37.91	38.95	36.02	**36.65**	37.38
Jane (Freitag et al., 2014)	39.83	42.75	38.63	39.10	40.08
Multi	40.64	44.81	38.80	38.26	40.63
Multi+Source	42.16	45.51	40.28	39.03	41.75
Multi+Ensemble	41.67	45.95	40.37	39.02	41.75
Multi+Source+Ensemble	**43.55**	**47.09**	**42.02**	**41.10**	**43.44**

3.4

5.3

- 句子级融合模型比最好的单系统平均提升**5.3**个**BLEU**值，比传统系统融合方法平均提升**3.4**个**BLEU**值！

总结

- 机器翻译从统计模型迁移到深度学习模型
- 基于注意机制的神经机器翻译模型具有更优的翻译性能
- 统计方法与神经网络方法融合，对两种范式兼容并包、博采众长

参考文献

1. Marcin Junczys-Dowmunt, Tomasz Dwojak and Hieu Hoang, 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. https://arxiv.org/pdf/1610.01108.pdf
2. Nal Kalchbrenner and Phil Blunsom, 2013. Recurrent Continuous Translation Models. *In Proc. of EMNLP 2013.*
3. Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, 2015. Neural Translation by Jointly Learning to Align and Translate. *http://arxiv.org/pdf/1409.0473.pdf*
4. Ilya Sutskever, Oriol Vinyals and Quoc V. Le, 2014. Sequence to Sequence Learning with Neural Networks. *In Proc. of NIPS 2014.*
5. Marcin Junczys-Dowmunt, Tomasz Dwojak and Hieu Hoang, 2016. Is Neural Machine Translation Ready for Deployment? A Case Study on 30 Translation Directions. arxiv.org/pdf/1610.01108
6. Philip Arthur, Graham Neubig, and Satoshi Nakamura. Incorporating discrete translation lexicons into neural machine translation. In Proceedings of EMNLP 2016.
7. Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. Memory-augmented neural machine translation. In proceedings of EMNLP 2017.

参考文献

8. Jiajun Zhang and Chengqing Zong. 2016b. Bridging Neural Machine Translation and Bilingual Dictionaries. https://arxiv.org/pdf/1610.07272.pdf

9. Yang Zhao, Yining Wang, Jiajun Zhang and Chengqing Zong, 2018. Phrase Table as Recommendation Memory for Neural Machine Translation. In Proc. of IJCAI 2018.

10. Yaohua Tang, Fandong Meng, Zhengdong Lu, Hang Li, and Philip LH Yu. Neural machine translation with external phrase memory. arXiv preprint arXiv:1606.01792, 2016

11. Xing Wang, Zhaopeng Tu, Deyi Xiong, and Min Zhang. Translating phrases in neural machine translation. In proceedings of EMNLP, 2017.

12. Jan Niehues, Eunah Cho, Thanh-Le Ha and Alex Waibel. Pre-translation for Neural Machine Translation. In Proceedings of COLING 2016.

13. Long Zhou, Wenpeng Hu, Jiajun Zhang and Chengqing Zong, 2017. Neural System Combination for Machine Translation. *In Proc. of ACL 2017.*

14. 宗成庆. 2008. 《统计自然语言处理》，清华大学出版社。

谢谢!

Q&A