

# 第11章 机器翻译

## 神经机器翻译

张家俊

---

北京市海淀区中关村东路95号  
邮编：100190



电话：+86-10-8254 4588  
邮件：jjzhang@nlpr.ia.ac.cn

# 统计机器翻译

Chinese: 我 在 北京 做了 报告

Phrase Seg:

①可解释性高

做了 报告

Phrase Trans:

人工设定的模块和特征

gave a talk

②模块随便加

③错误易追踪

English:

I

gave

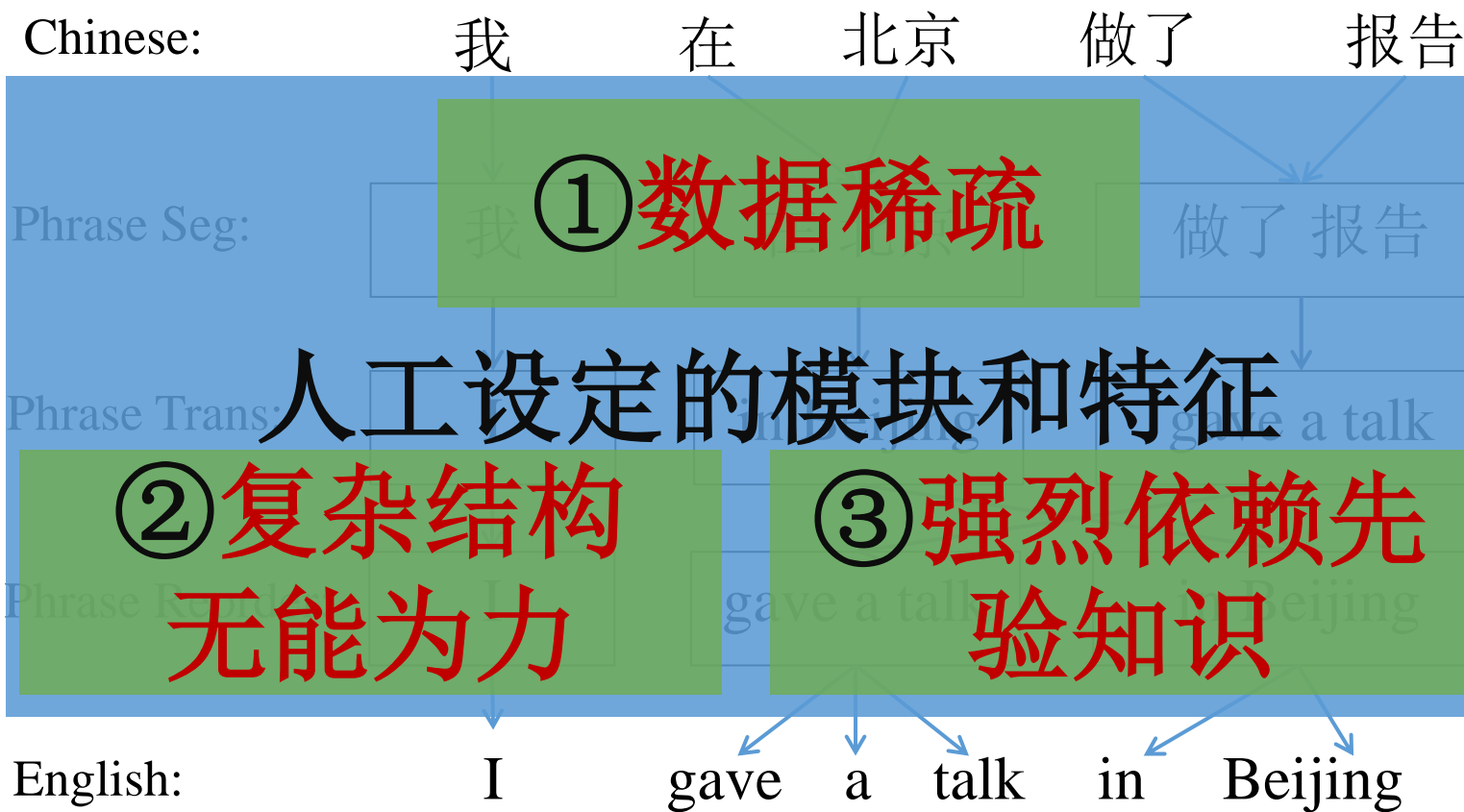
a

talk

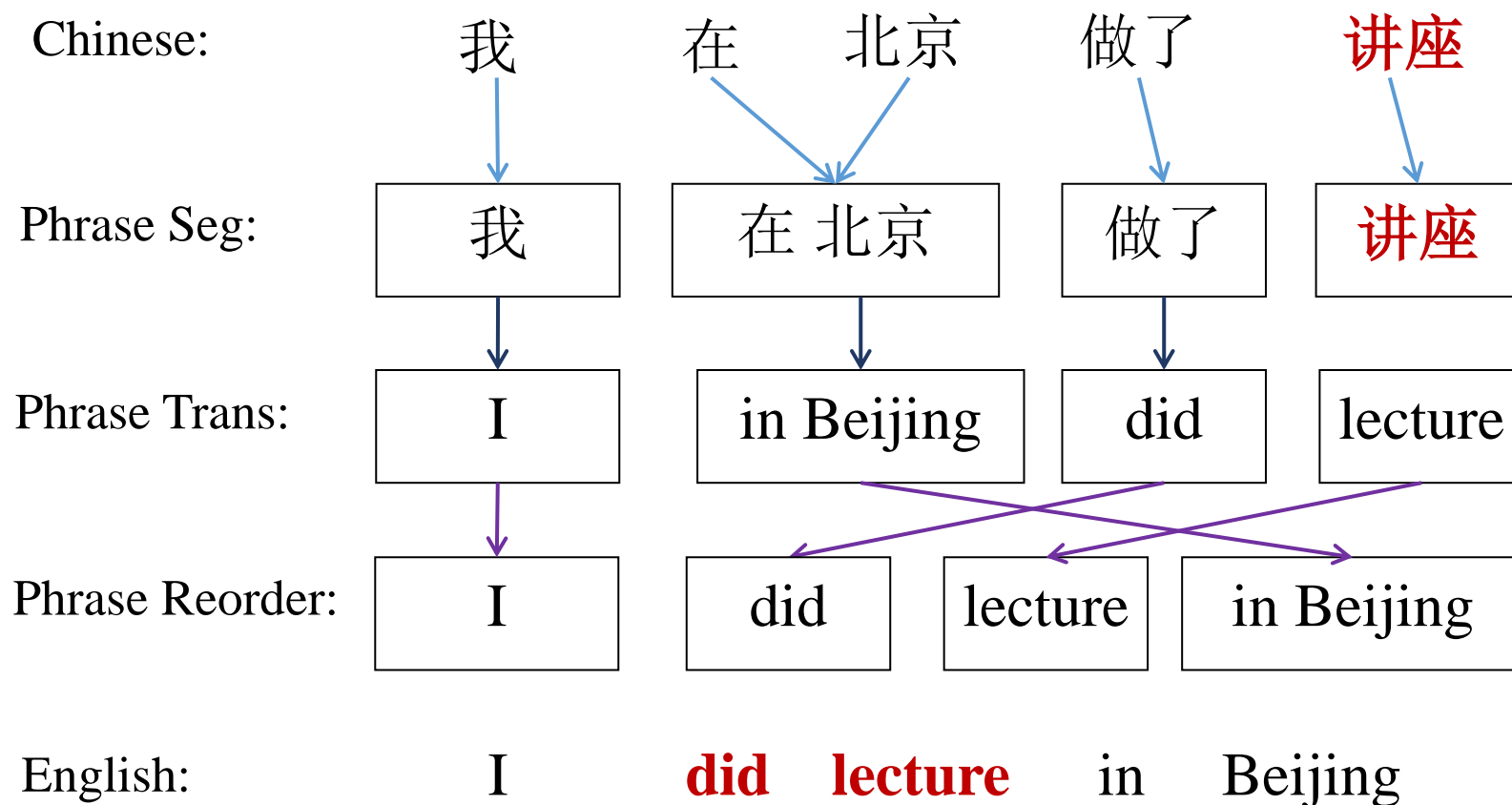
in

Beijing

# 统计机器翻译



# 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  $\Rightarrow$  连续分布式表示方法

讲座  $\otimes$  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

编码网络

仅需要两个神经网络

分布式语义表示

解码网络

English:

I gave a talk in Beijing

# 神经机器翻译

## 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。

Ä 🎤 🔊 🔍

English Chinese (Simplified) Spanish

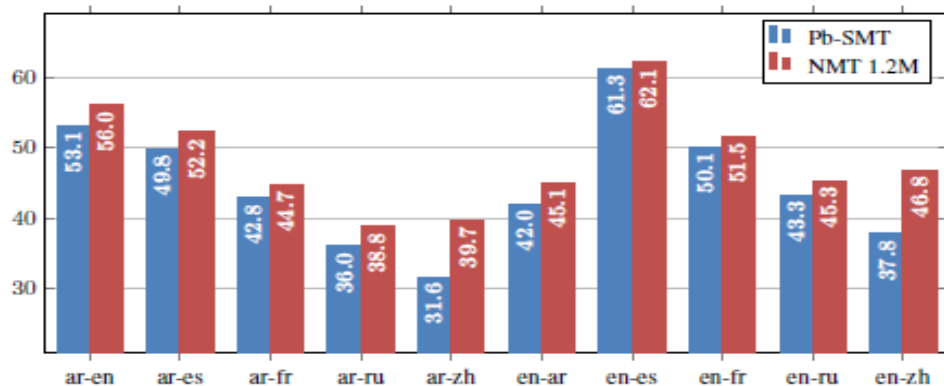
Translate

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.

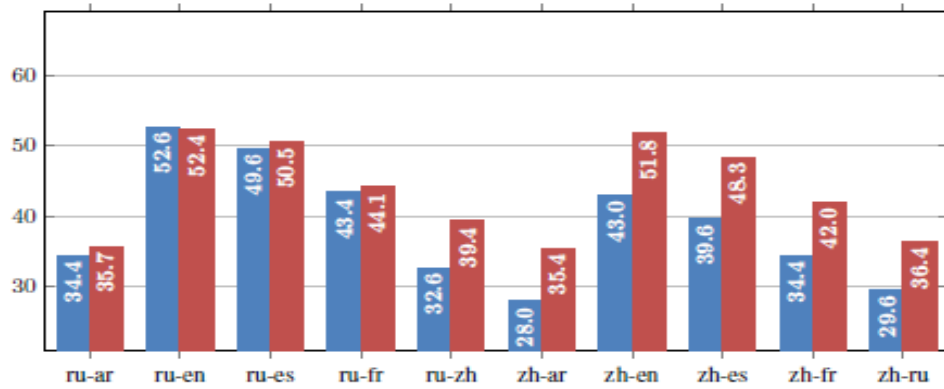
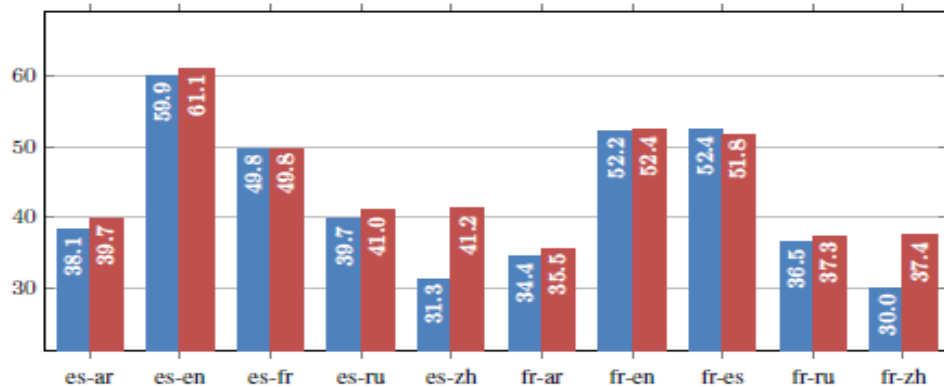
☆ 📄 🔊 🔍

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  $\Rightarrow$  连续分布式表示方法

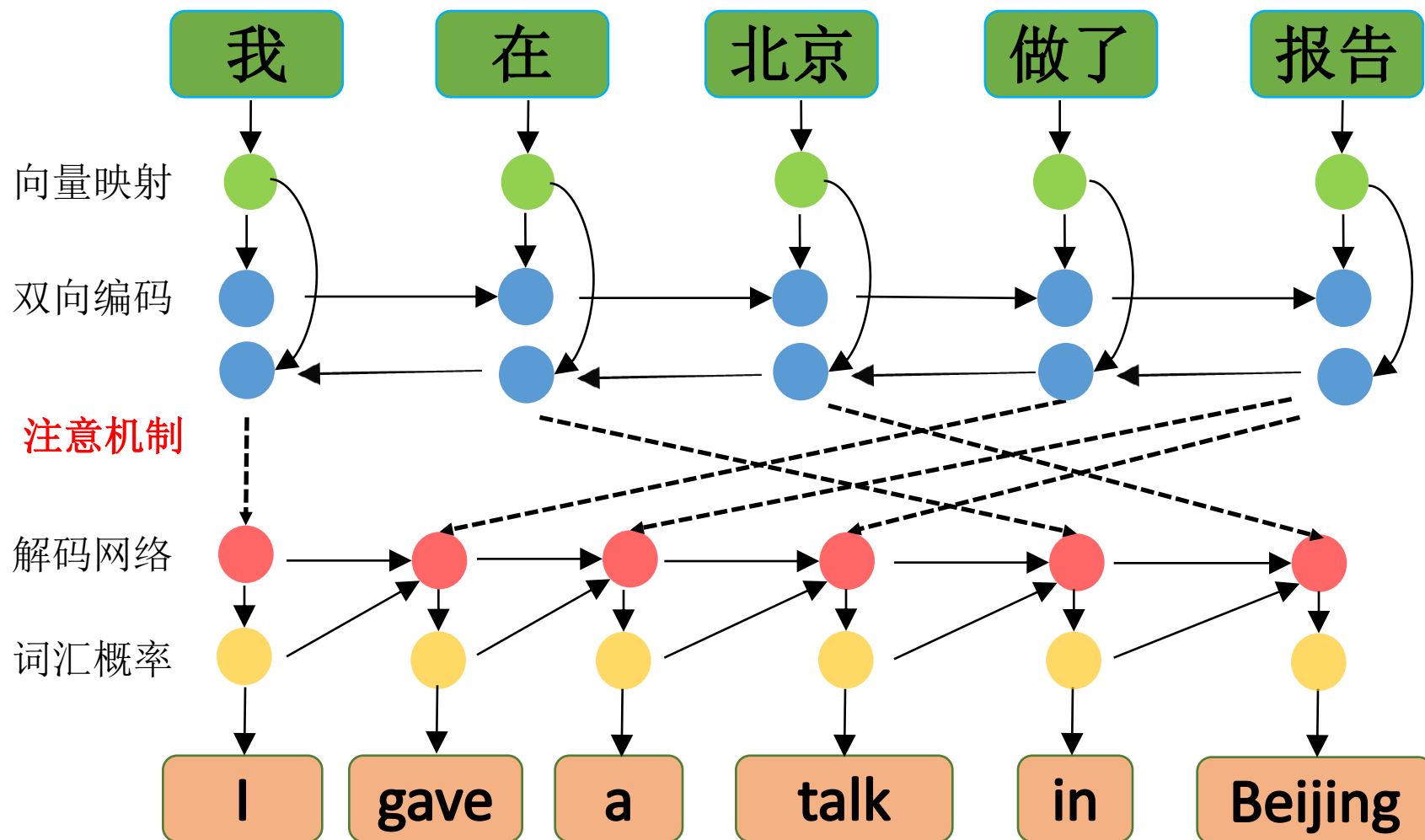
讲座

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

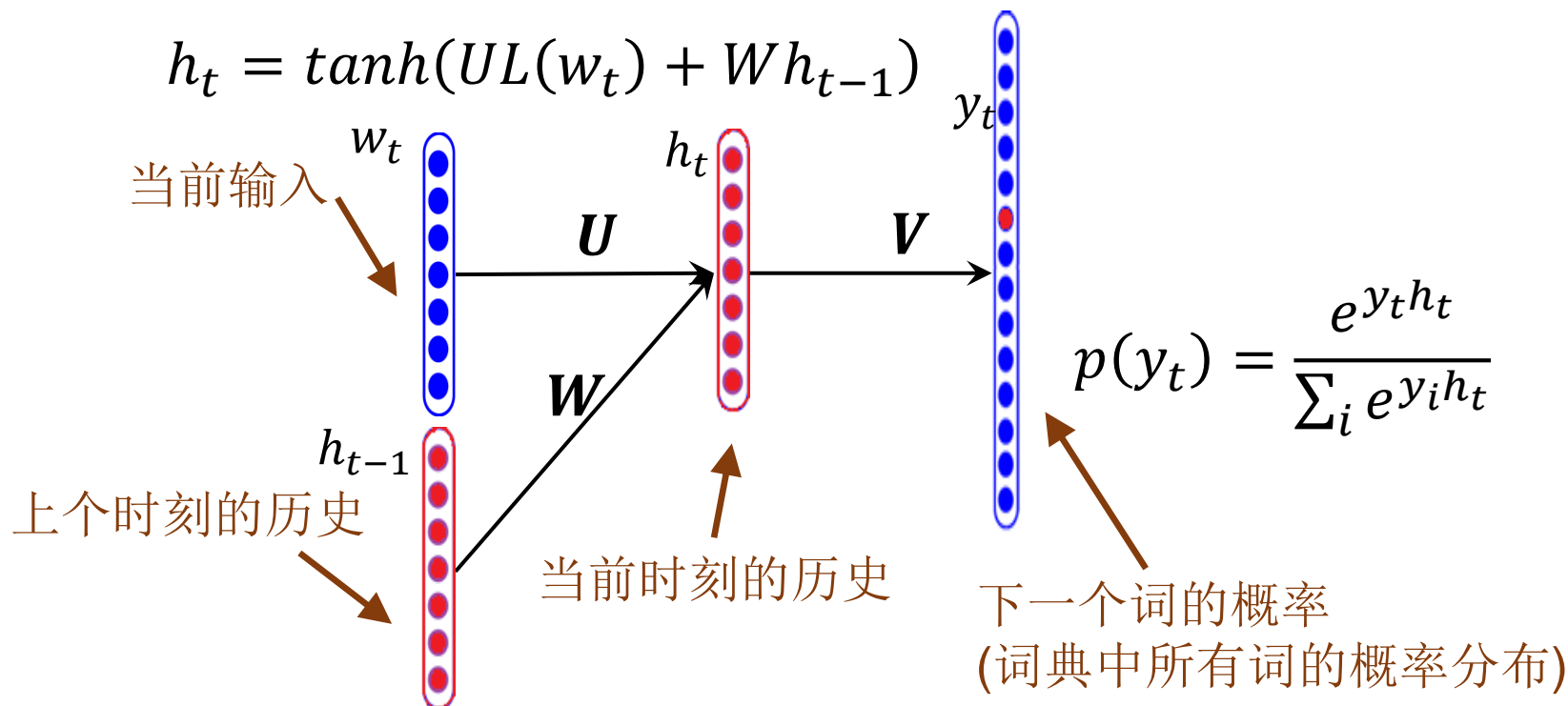
# 神经机器翻译



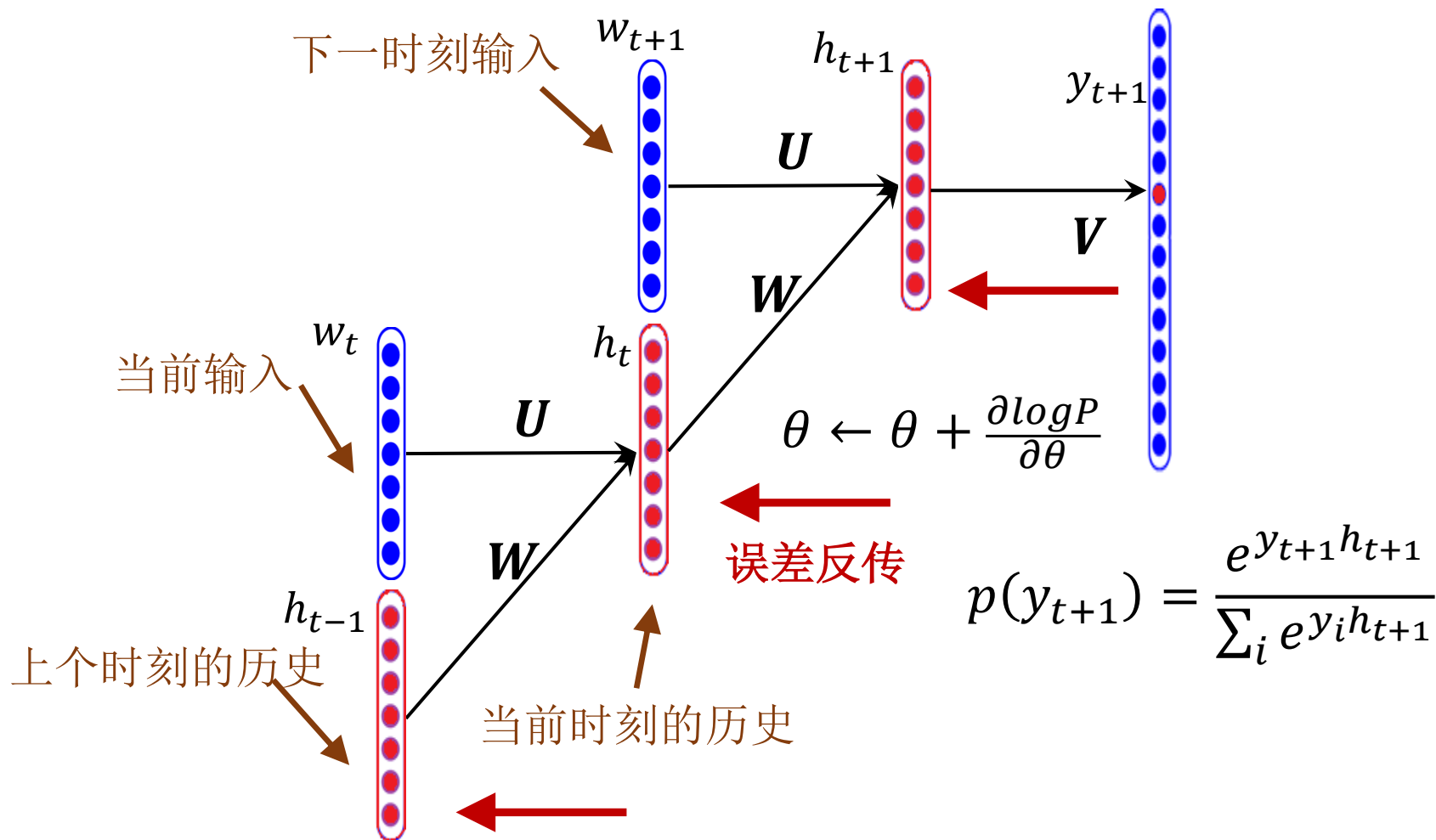


# 循环神经网络

- 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



# 循环神经网络



# 神经机器翻译

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

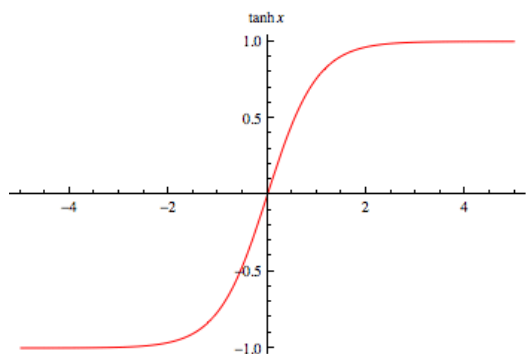
$L(w_s):$        $w_s \longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in R^3$       我  $\longrightarrow \begin{bmatrix} 0.1 \\ 0.9 \\ 0.6 \end{bmatrix}$       随机初始化

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

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

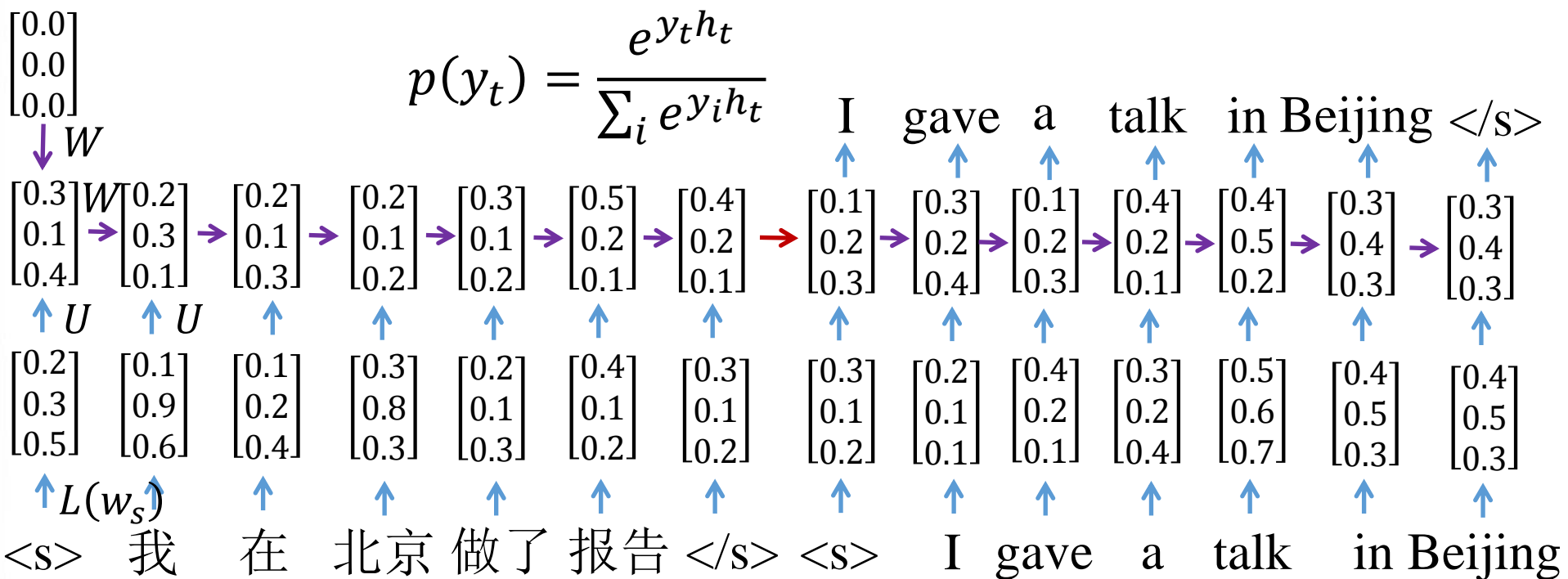
$$z = UL(w_s) + Wh_{s-1} \in R^3$$

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



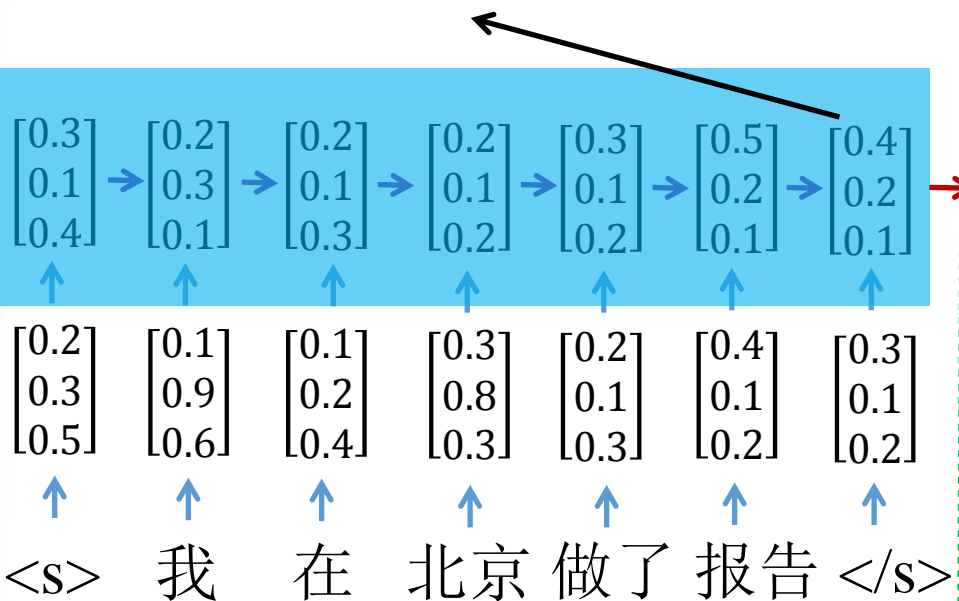
# 神经机器翻译

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



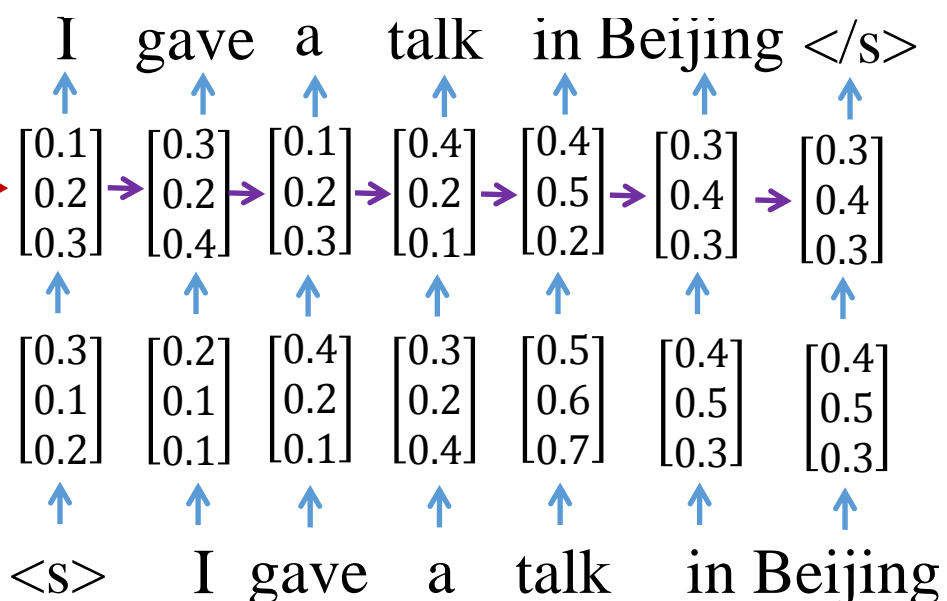
# 神经机器翻译

将源语言句子编码成一个  
实数向量语义表示



编码器

将源语言句子的语义表示  
解码生成目标语言句子

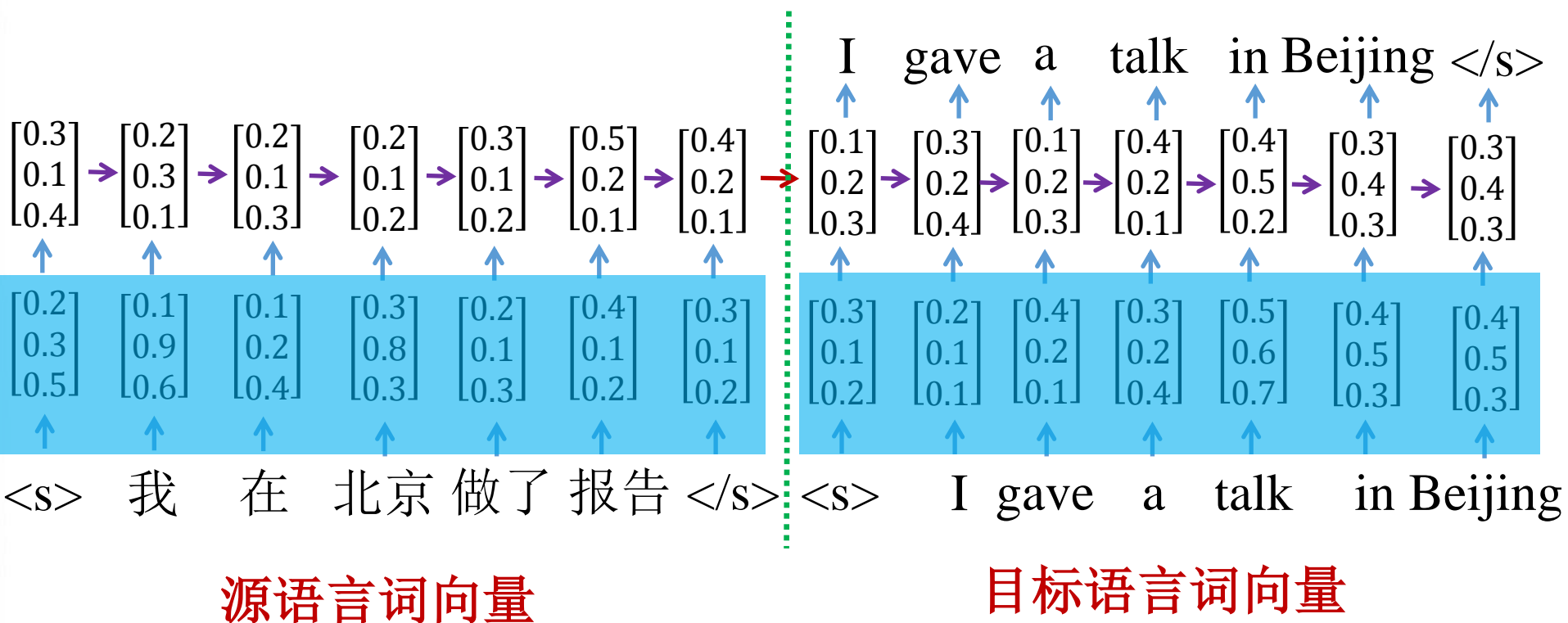


解码器

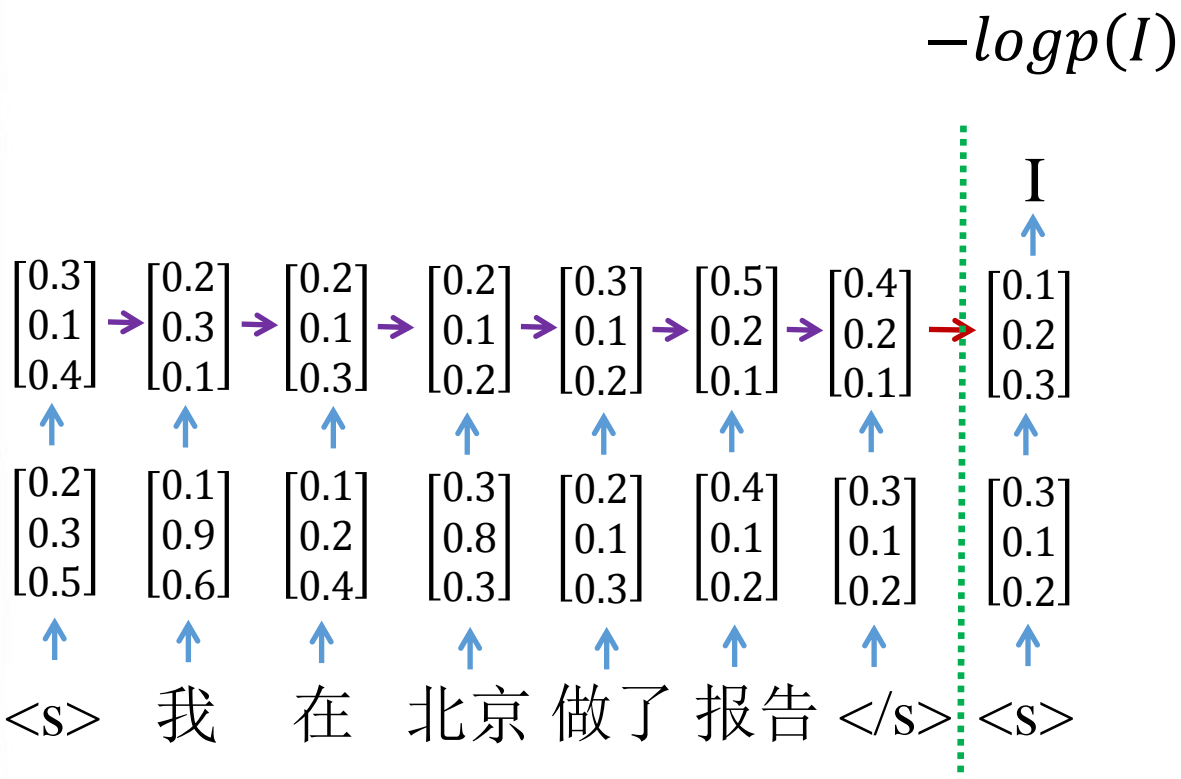


# 神经机器翻译

词向量随机初始化，在训练过程中进行优化！

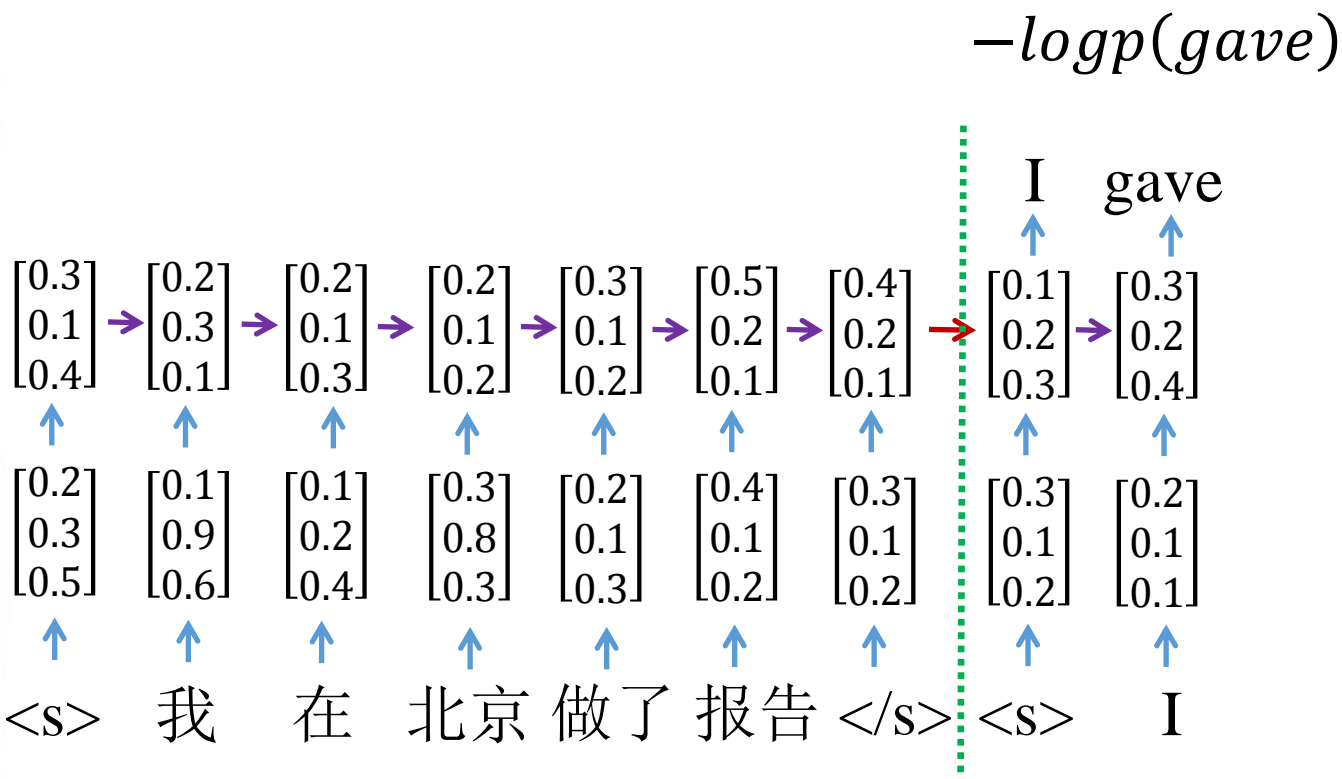


# 神经机器翻译



最大化  $P(\text{target}|\text{source})$

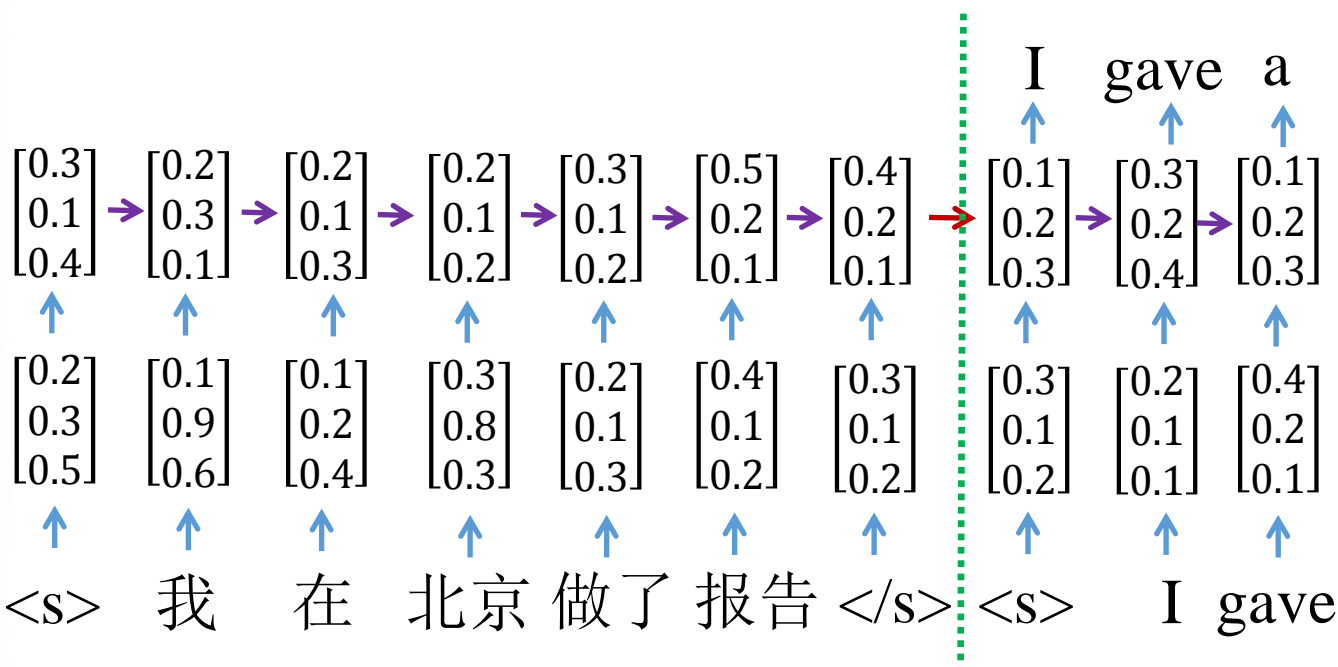
# 神经机器翻译



最大化  $P(\text{target}|\text{source})$

# 神经机器翻译

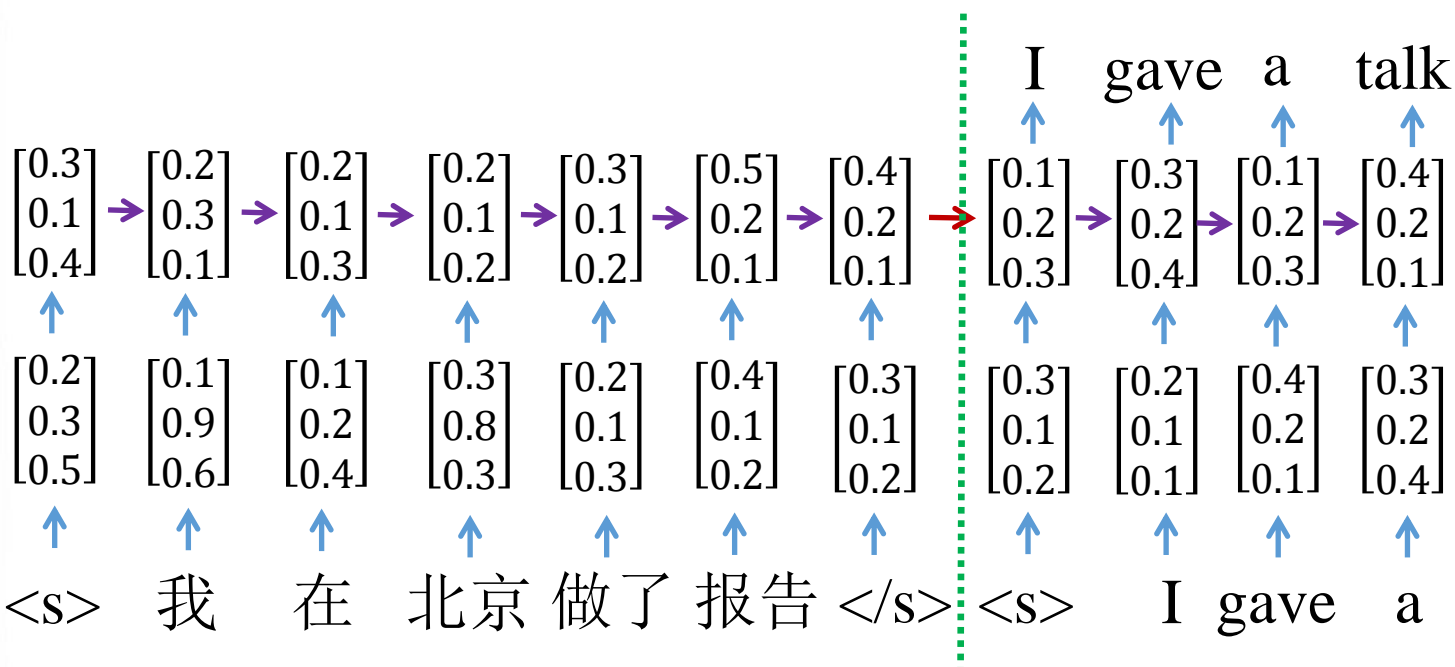
$$-\log p(a)$$



最大化  $P(\text{target}|\text{source})$

# 神经机器翻译

$$-\log p(\text{talk})$$

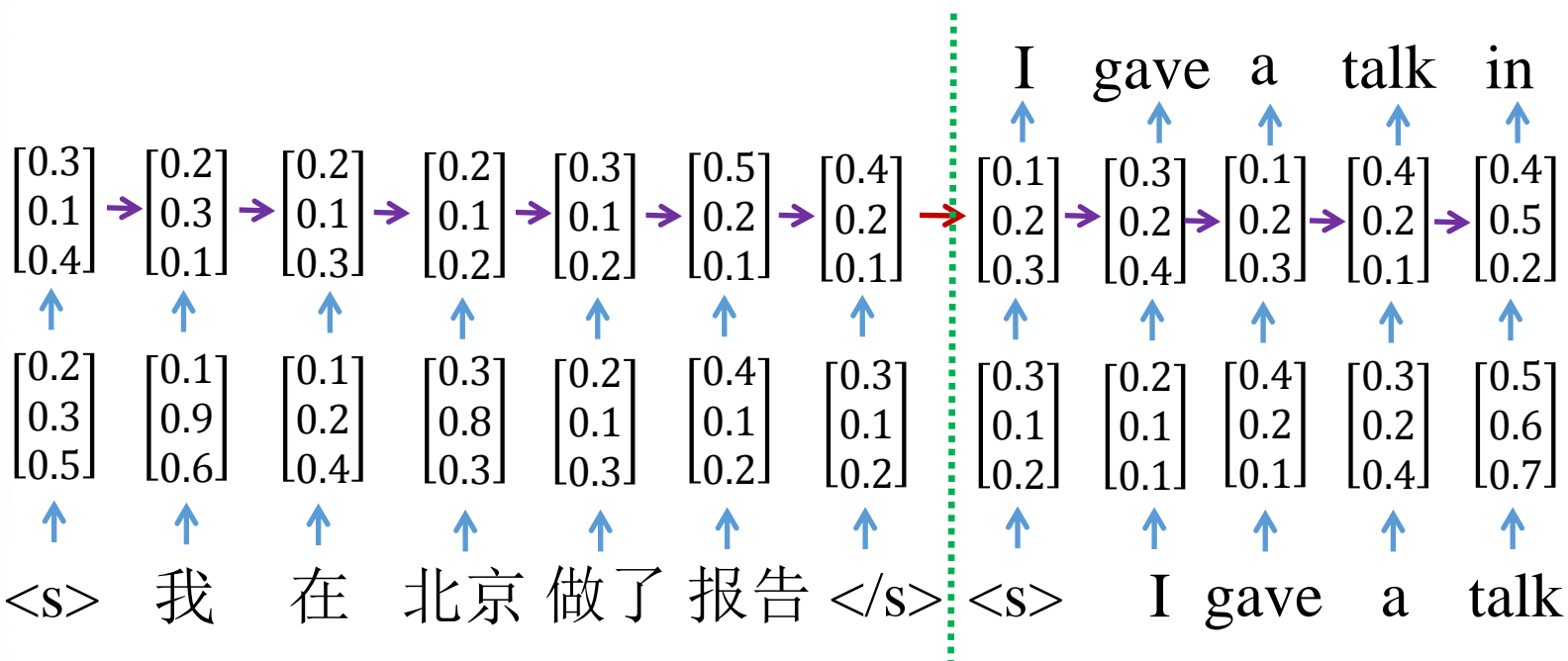


最大化  $P(\text{target}|\text{source})$



# 神经机器翻译

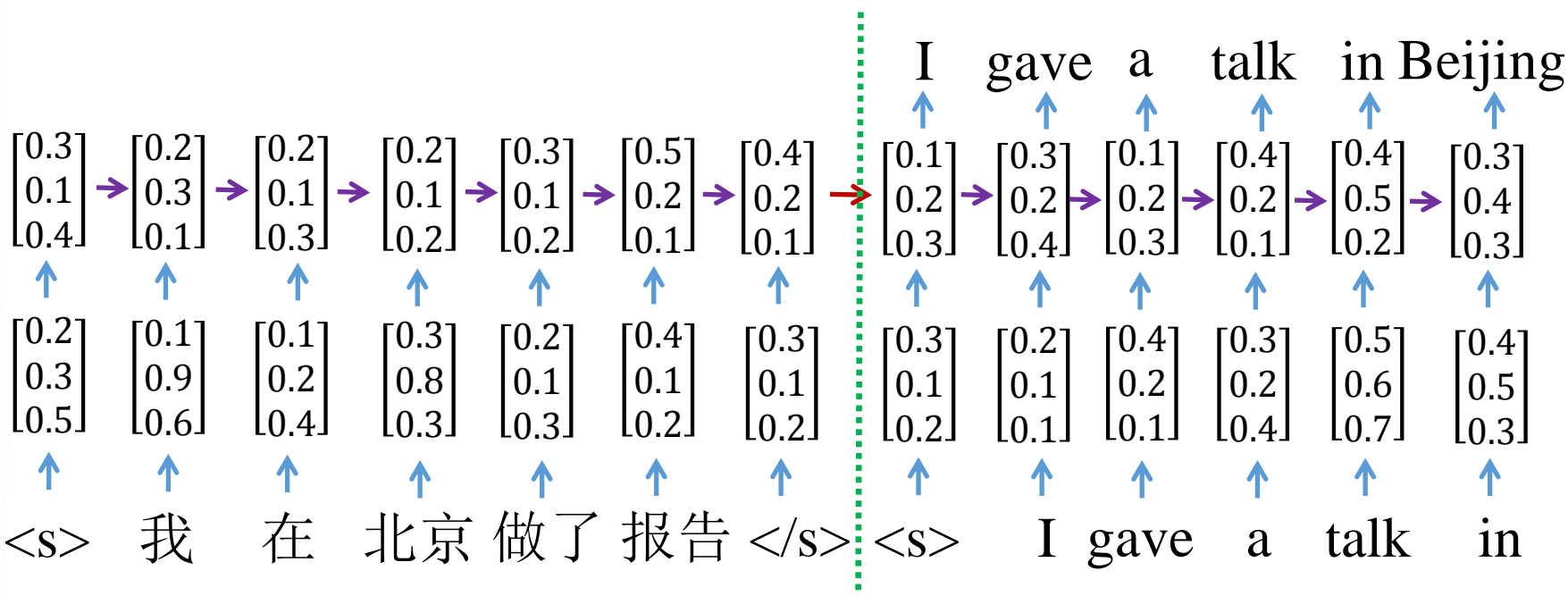
$$-\log p(in)$$



最大化  $P(\text{target}|\text{source})$

# 神经机器翻译

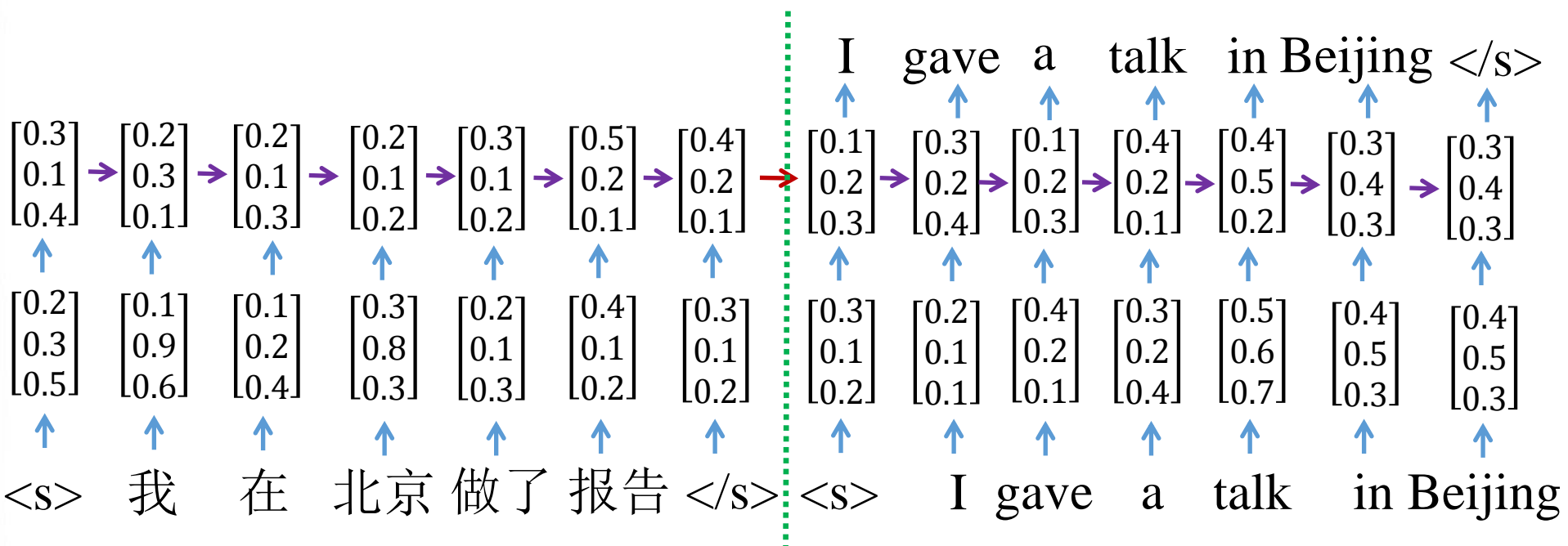
$-\log p(\text{Beijing})$



最大化  $P(\text{target}|\text{source})$

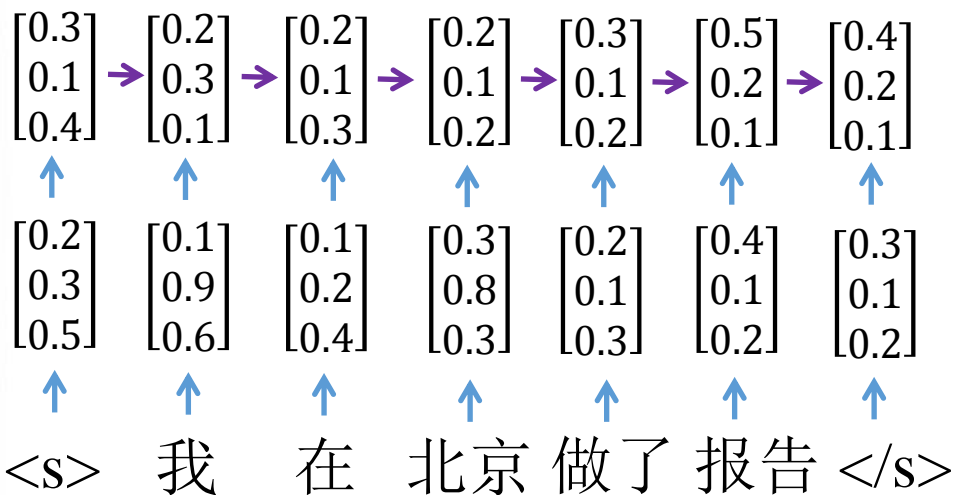
# 神经机器翻译

$$-\log p(\langle /s \rangle)$$



最大化  $P(\text{target}|\text{source})$

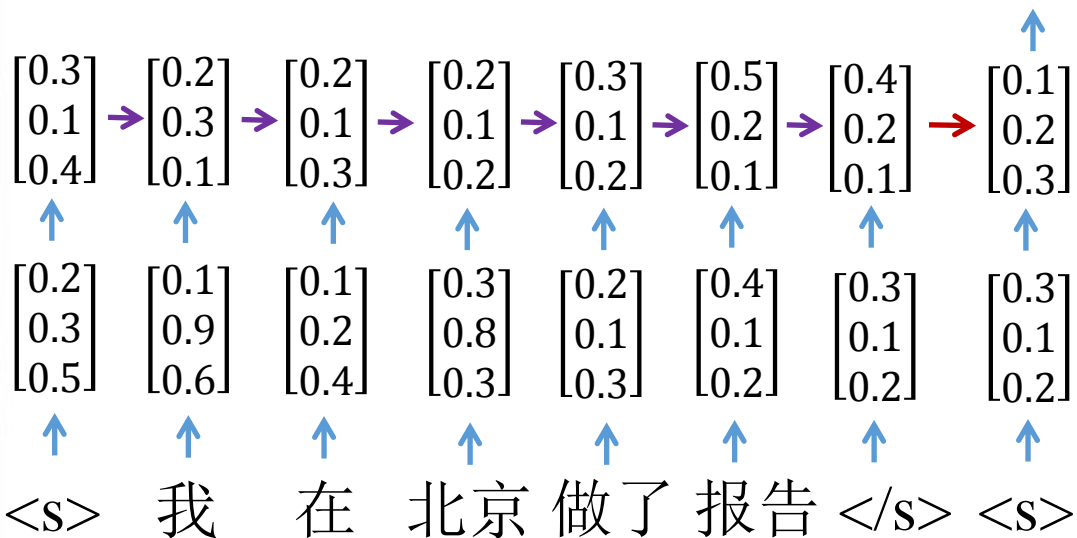
# 神经机器翻译-测试



# 神经机器翻译-测试

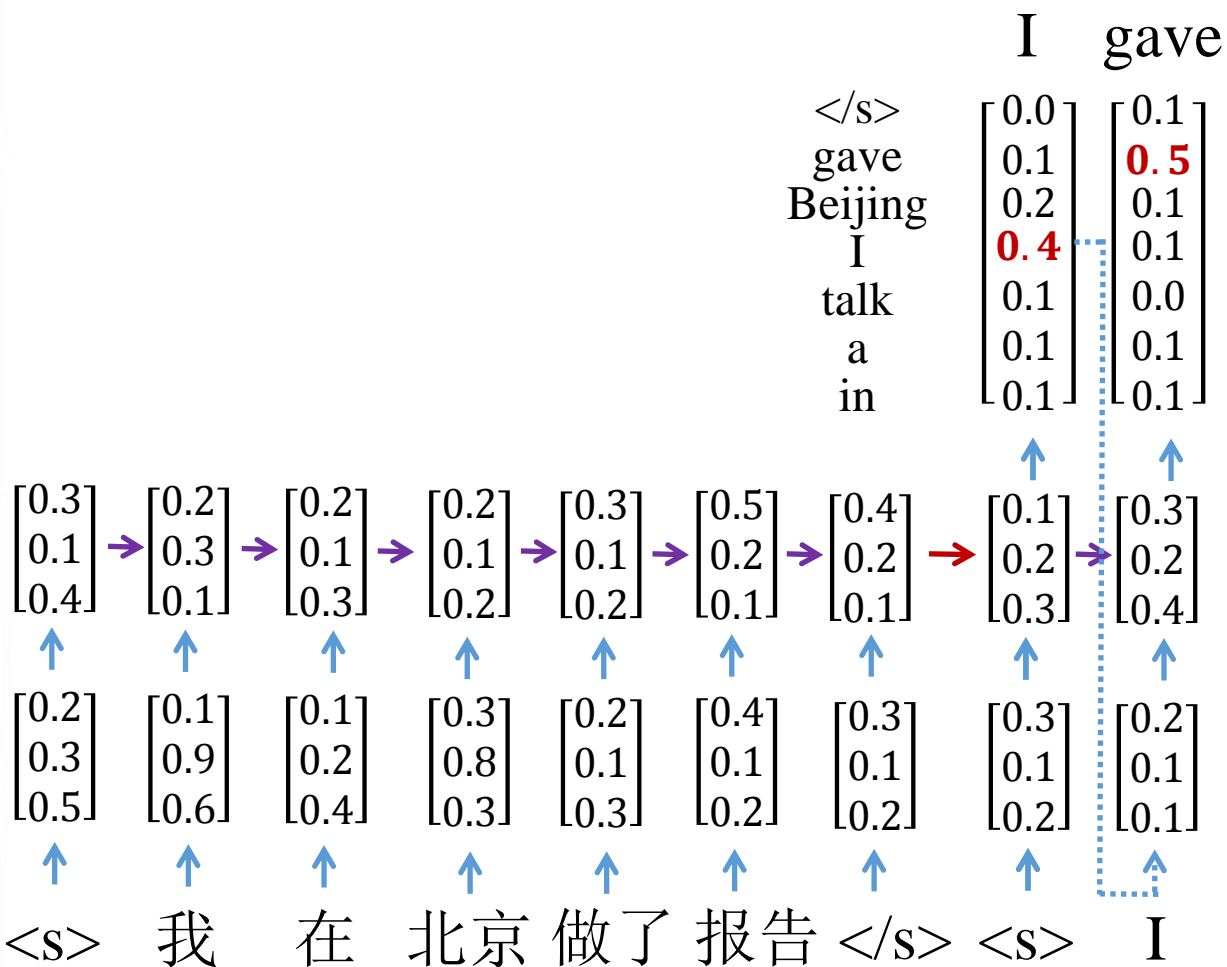
$$p(y_t) = \frac{e^{y_t h_t}}{\sum_i e^{y_i h_t}}$$

	I
</s>	$\begin{bmatrix} 0.0 \\ 0.1 \\ 0.2 \\ \mathbf{0.4} \\ 0.1 \\ 0.1 \\ 0.1 \end{bmatrix}$
gave	
Beijing	
I	
talk	
a	
in	



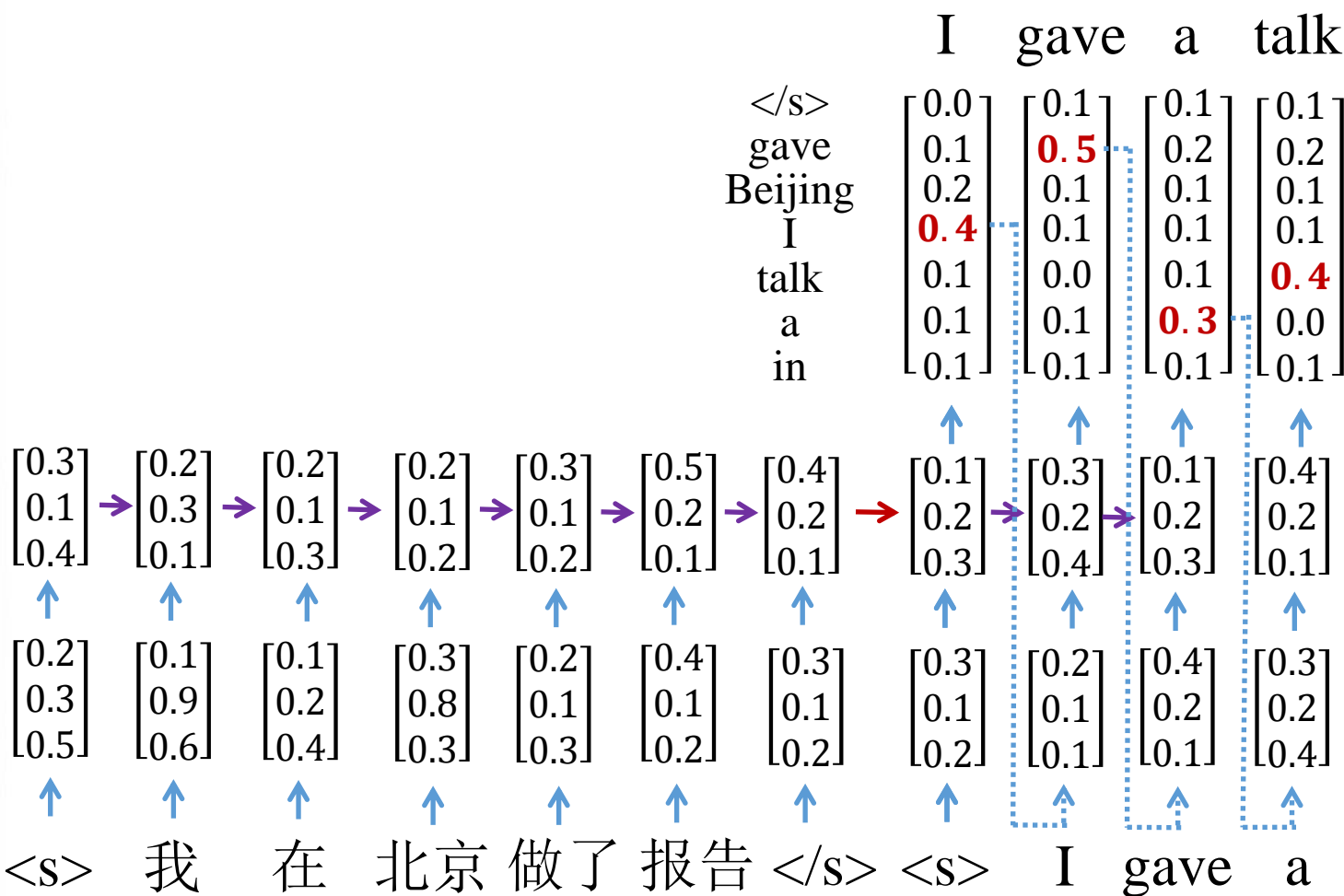


# 神经机器翻译-测试

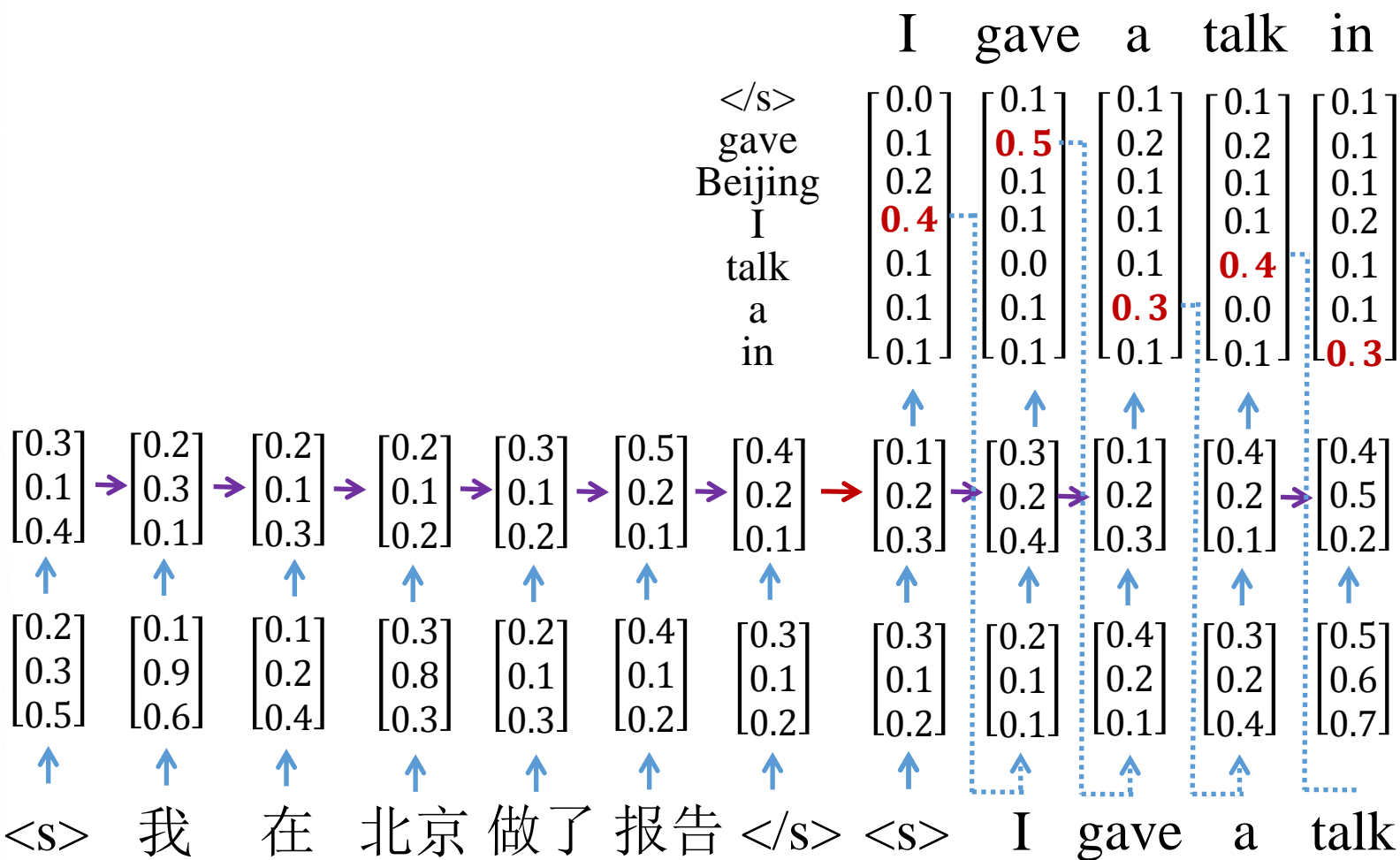




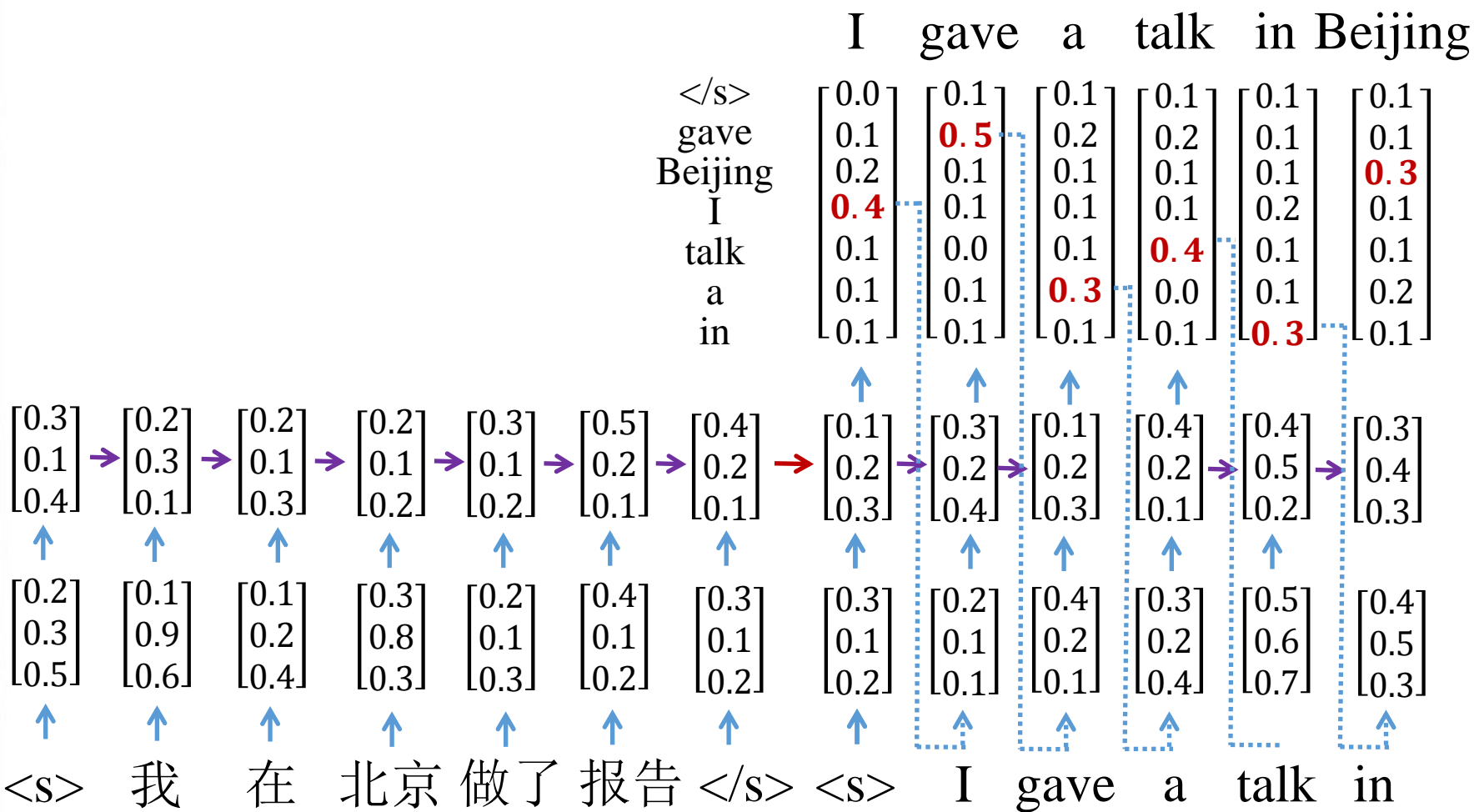
# 神经机器翻译-测试



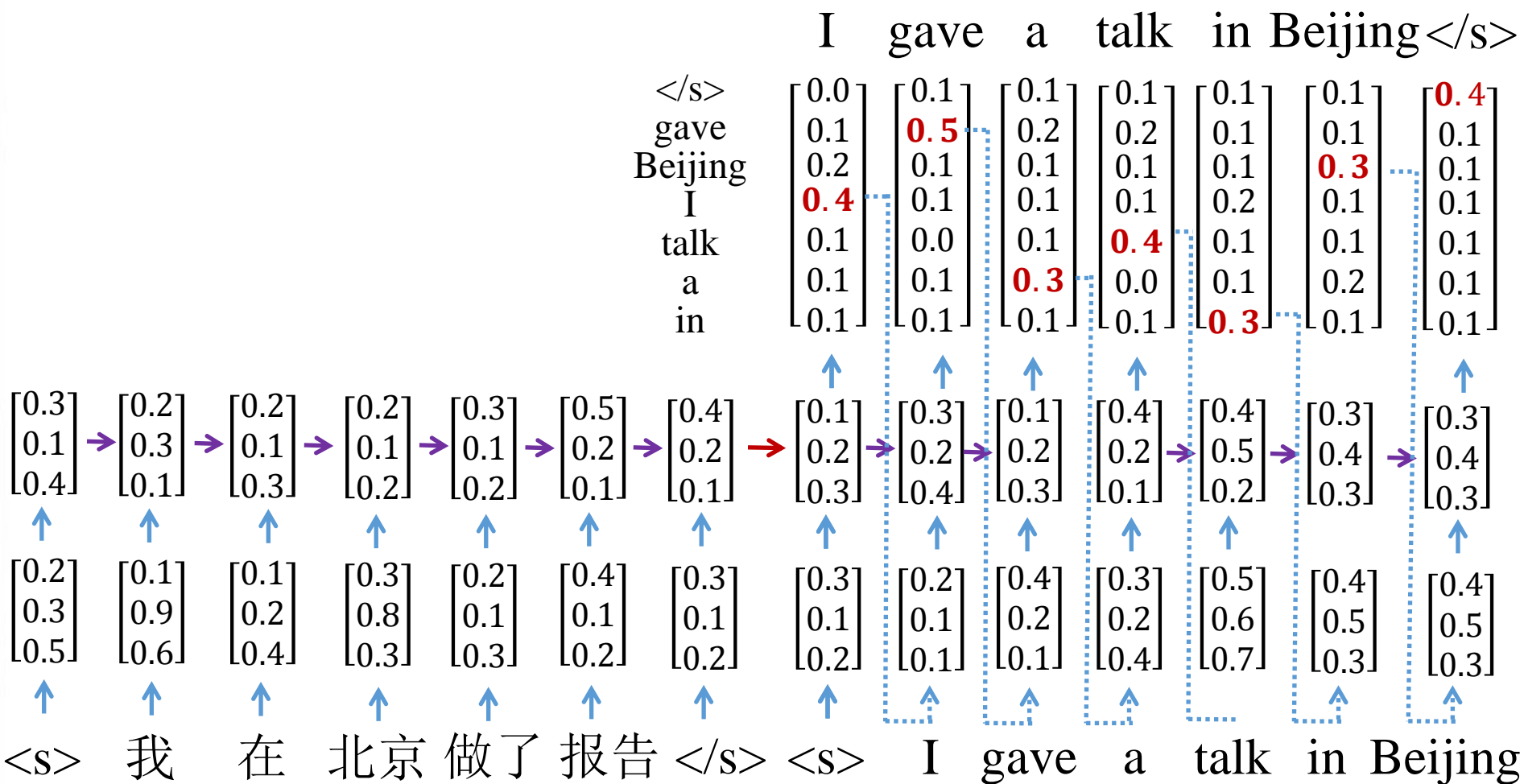
# 神经机器翻译-测试



# 神经机器翻译-测试

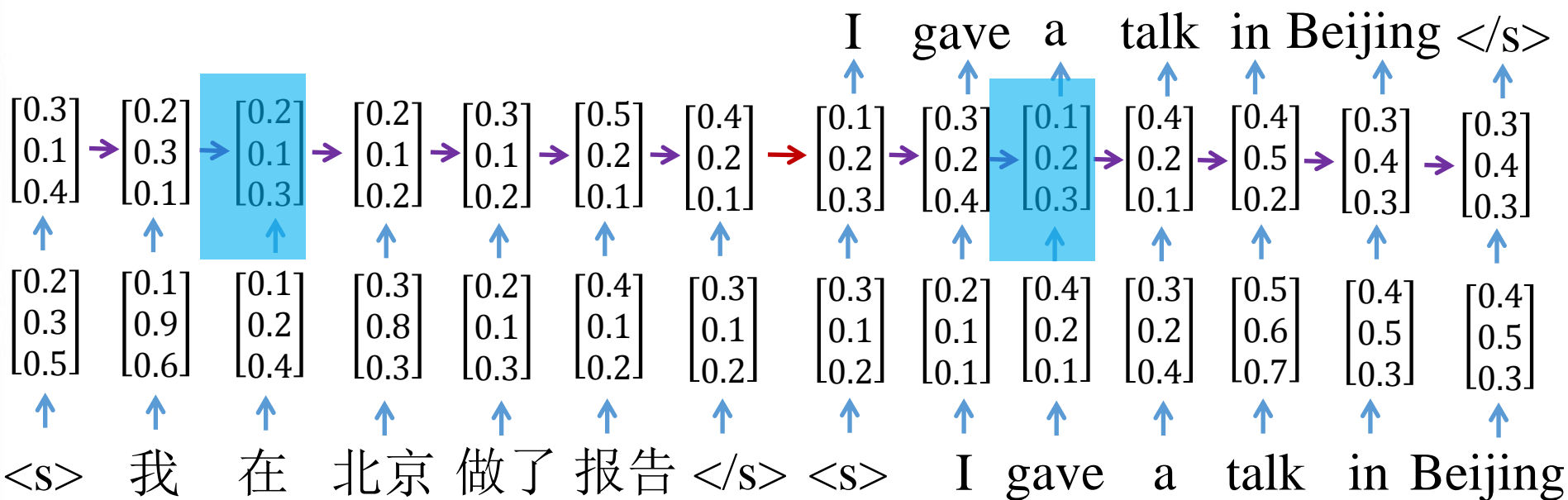


# 神经机器翻译-测试

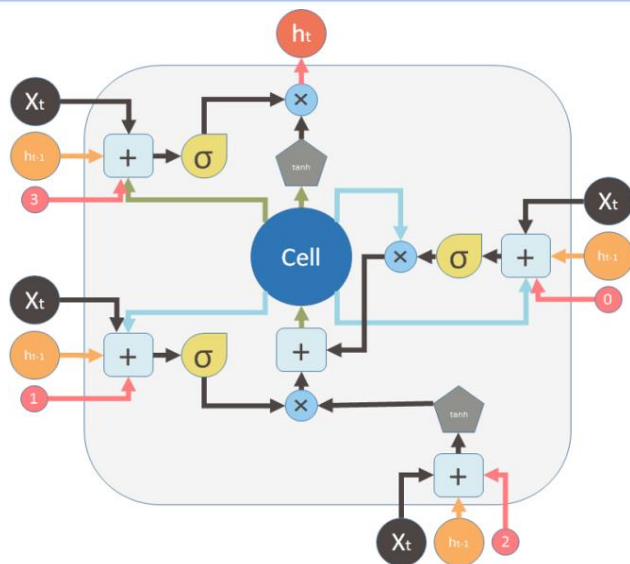


# 神经机器翻译-计算单元

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



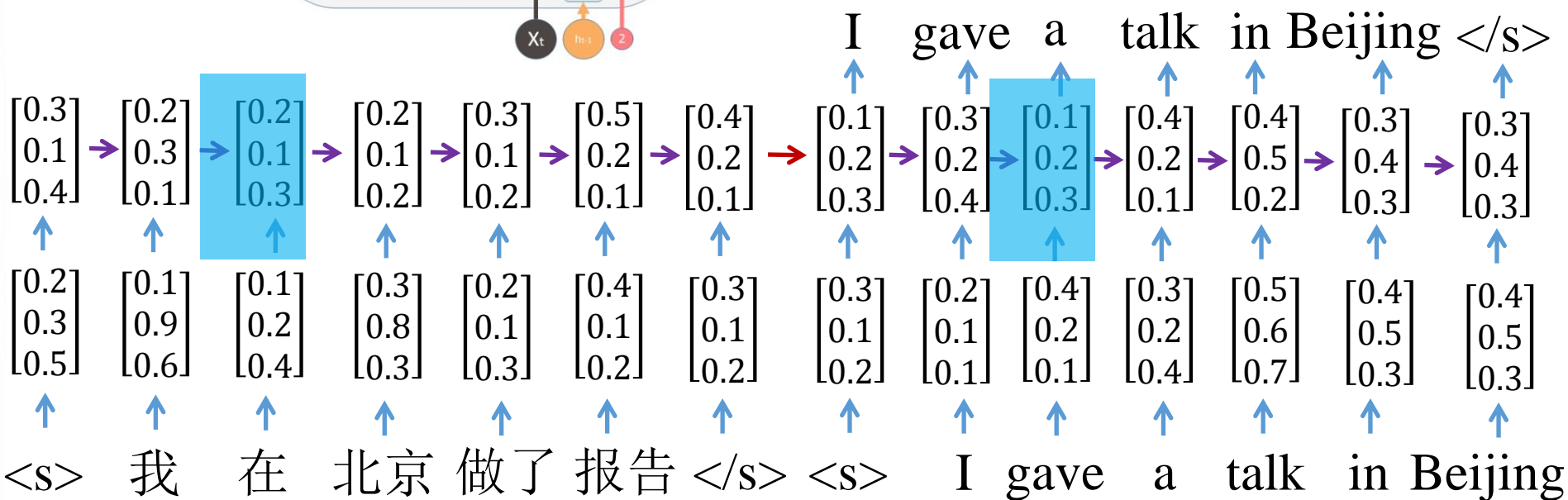
# 神经机器翻译-计算单元



## LSTM计算单元

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

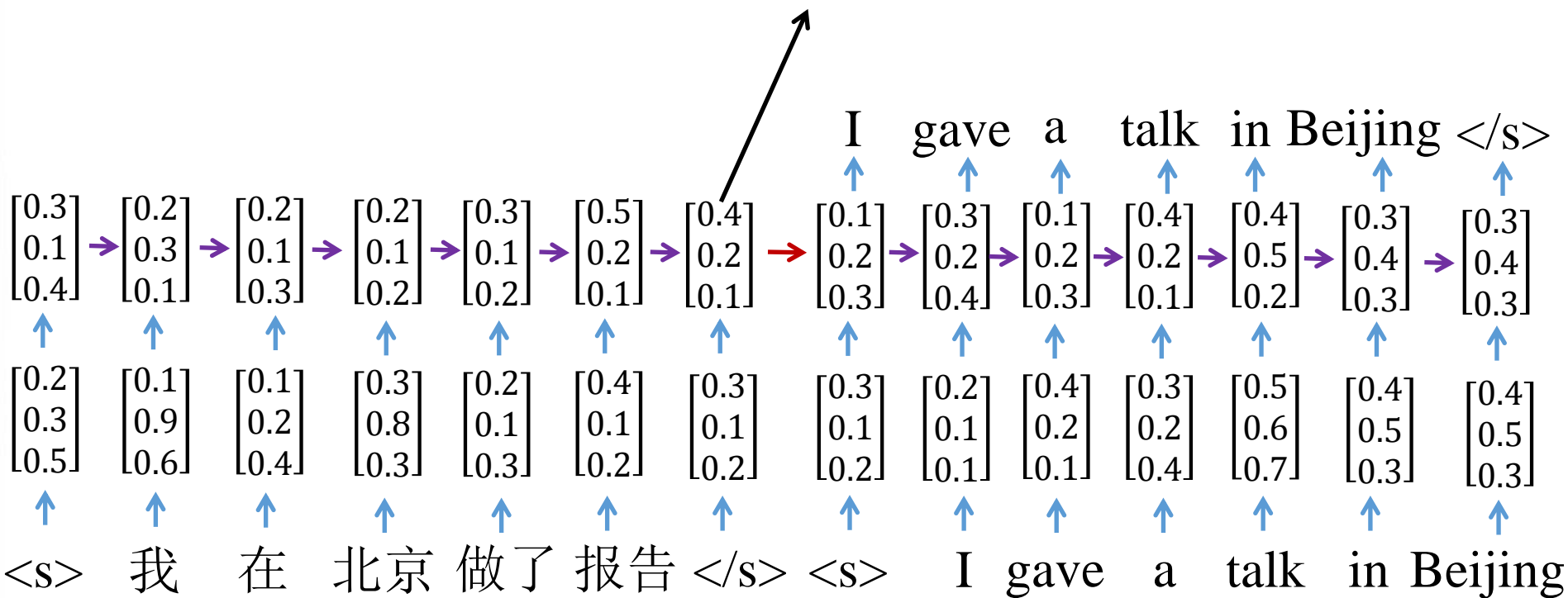
$$h_t = LSTM(w_t, h_{t-1})$$





# 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

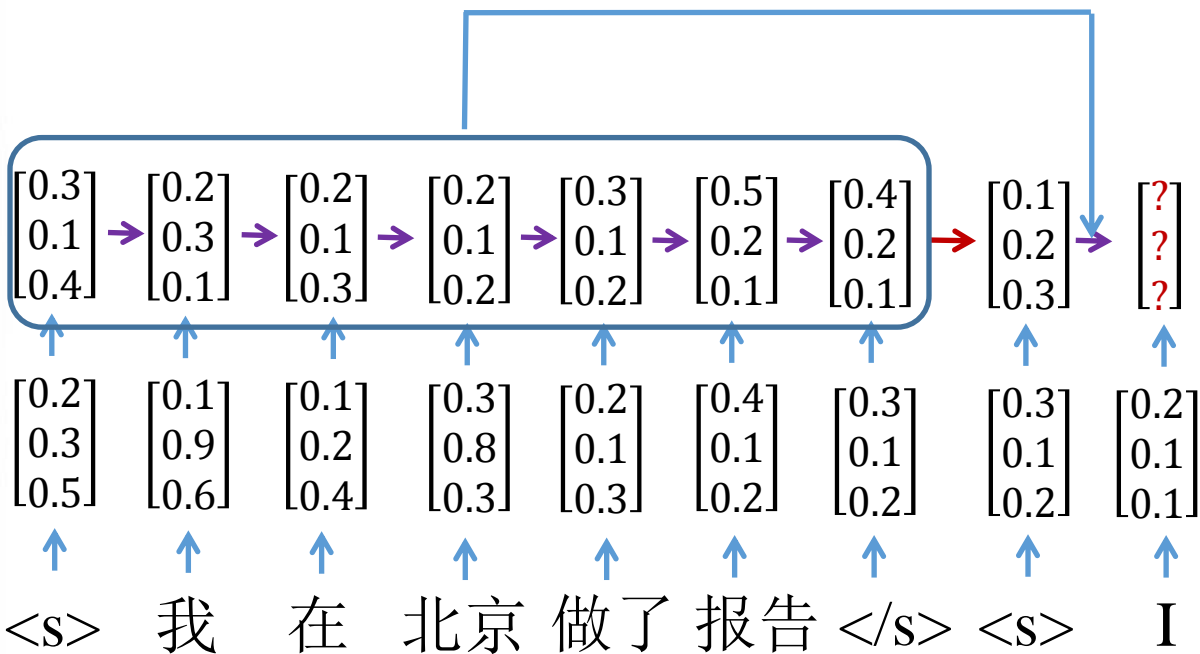
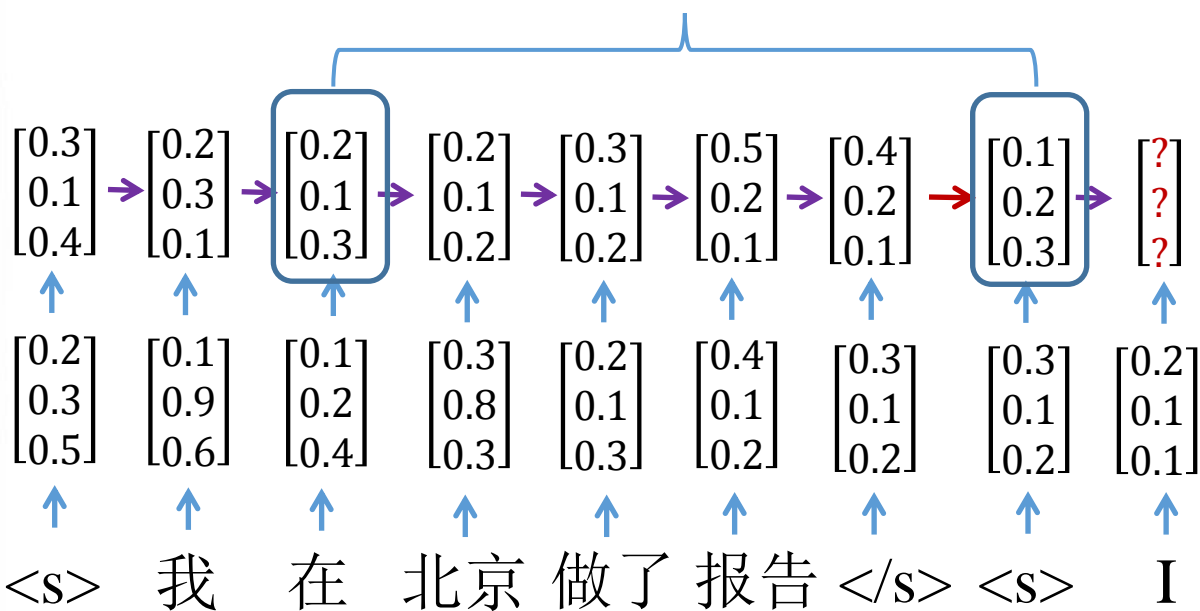


Diagram illustrating the state transition in a sequence-to-sequence model. The diagram shows a sequence of hidden states (vectors) and input tokens. The first sequence is "I in Beijing did a report" followed by a start token "<s>". The second sequence is "I" followed by a start token "<s>". The hidden states are represented as 3x1 column vectors. The first state of the first sequence is highlighted with a blue box. The last state of the first sequence is also highlighted with a blue box. A red arrow indicates the transition from the last state of the first sequence to the first state of the second sequence. A blue bracket above the states indicates a sequence of operations.

Diagram illustrating the state transition in an RNN encoder for the sentence "我在北京做了报告" (I made a report in Beijing). The diagram shows a sequence of hidden states (vectors) and input tokens. The first state is  $\begin{bmatrix} 0.3 \\ 0.1 \\ 0.4 \end{bmatrix}$ , followed by a sequence of states:  $\begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \end{bmatrix}$ ,  $\begin{bmatrix} 0.2 \\ 0.1 \\ 0.3 \end{bmatrix}$ ,  $\begin{bmatrix} 0.2 \\ 0.1 \\ 0.2 \end{bmatrix}$ ,  $\begin{bmatrix} 0.3 \\ 0.1 \\ 0.2 \end{bmatrix}$ ,  $\begin{bmatrix} 0.5 \\ 0.2 \\ 0.1 \end{bmatrix}$ ,  $\begin{bmatrix} 0.4 \\ 0.2 \\ 0.1 \end{bmatrix}$ , and finally  $\begin{bmatrix} 0.1 \\ 0.2 \\ 0.3 \end{bmatrix}$ . The last state is highlighted with a blue box. The input tokens are  $\langle s \rangle$ , 我, 在, 北, 京, 做, 了, 报, 告,  $\langle /s \rangle$ ,  $\langle s \rangle$ , and I. The inputs for the first state are  $\begin{bmatrix} 0.2 \\ 0.3 \\ 0.5 \end{bmatrix}$ , and for the last state are  $\begin{bmatrix} 0.3 \\ 0.1 \\ 0.2 \end{bmatrix}$ . The output for the last state is  $\begin{bmatrix} ? \\ ? \\ ? \end{bmatrix}$ .

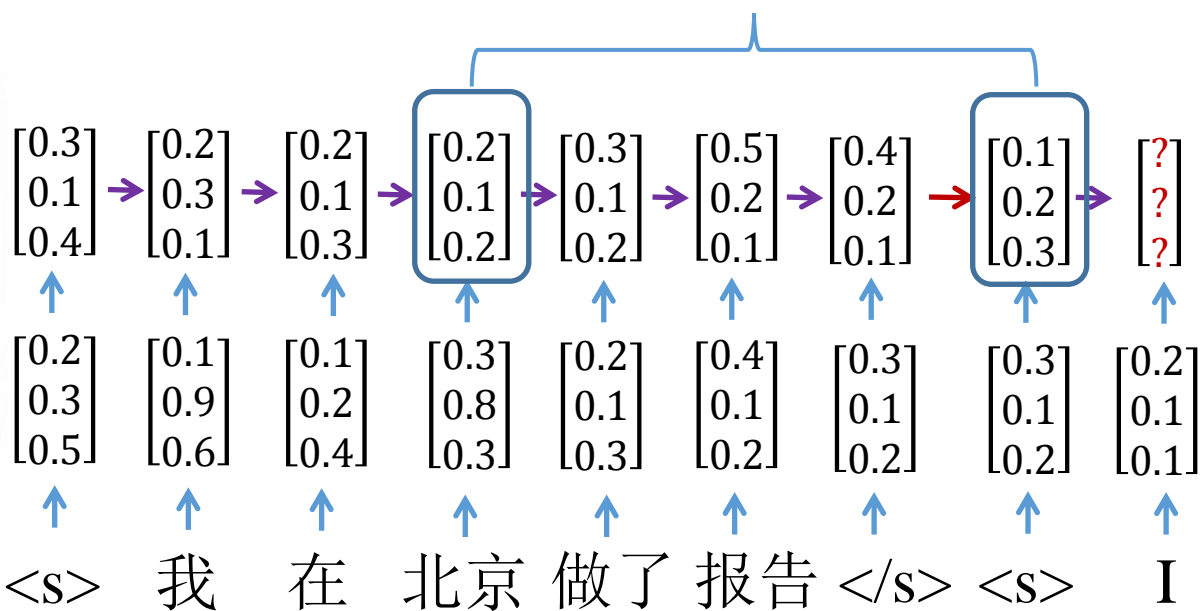
# 神经机器翻译-注意机制

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



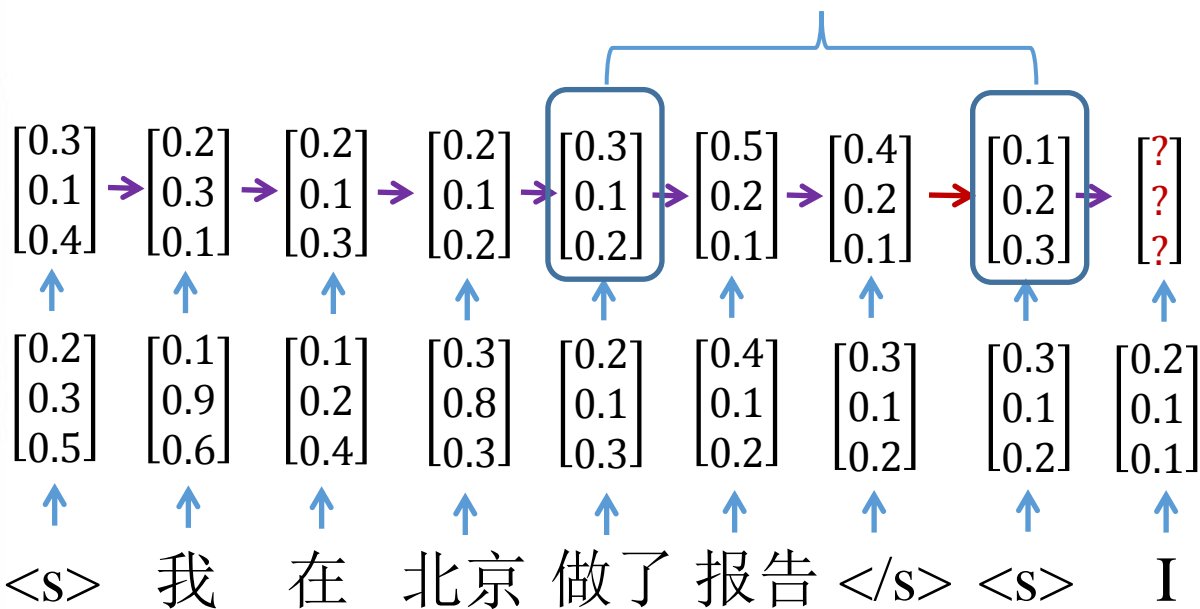
# 神经机器翻译-注意机制

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



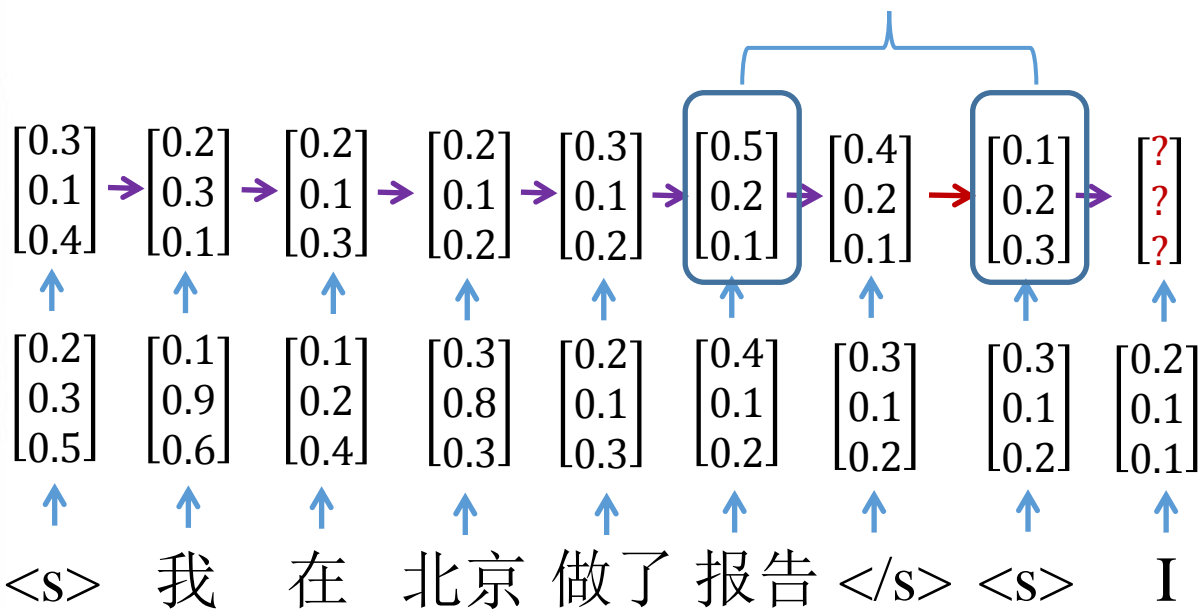
# 神经机器翻译-注意机制

$$score(h_s, h_t) = 4$$



# 神经机器翻译-注意机制

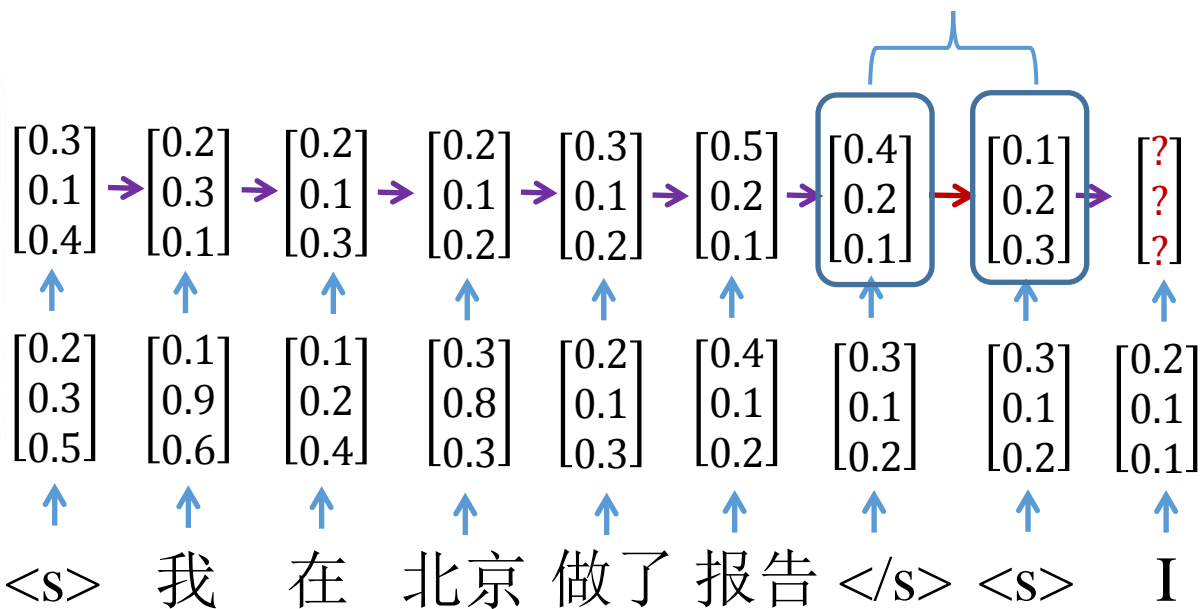
$$score(h_s, h_t) = 2$$





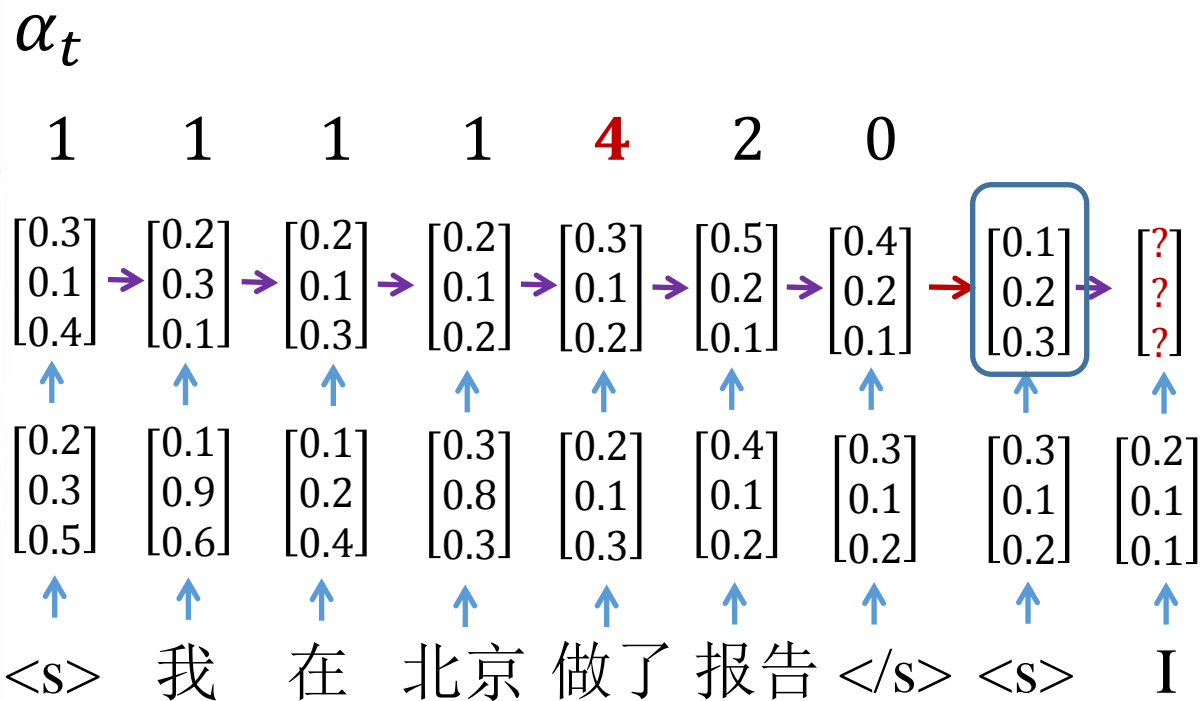
# 神经机器翻译-注意机制

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

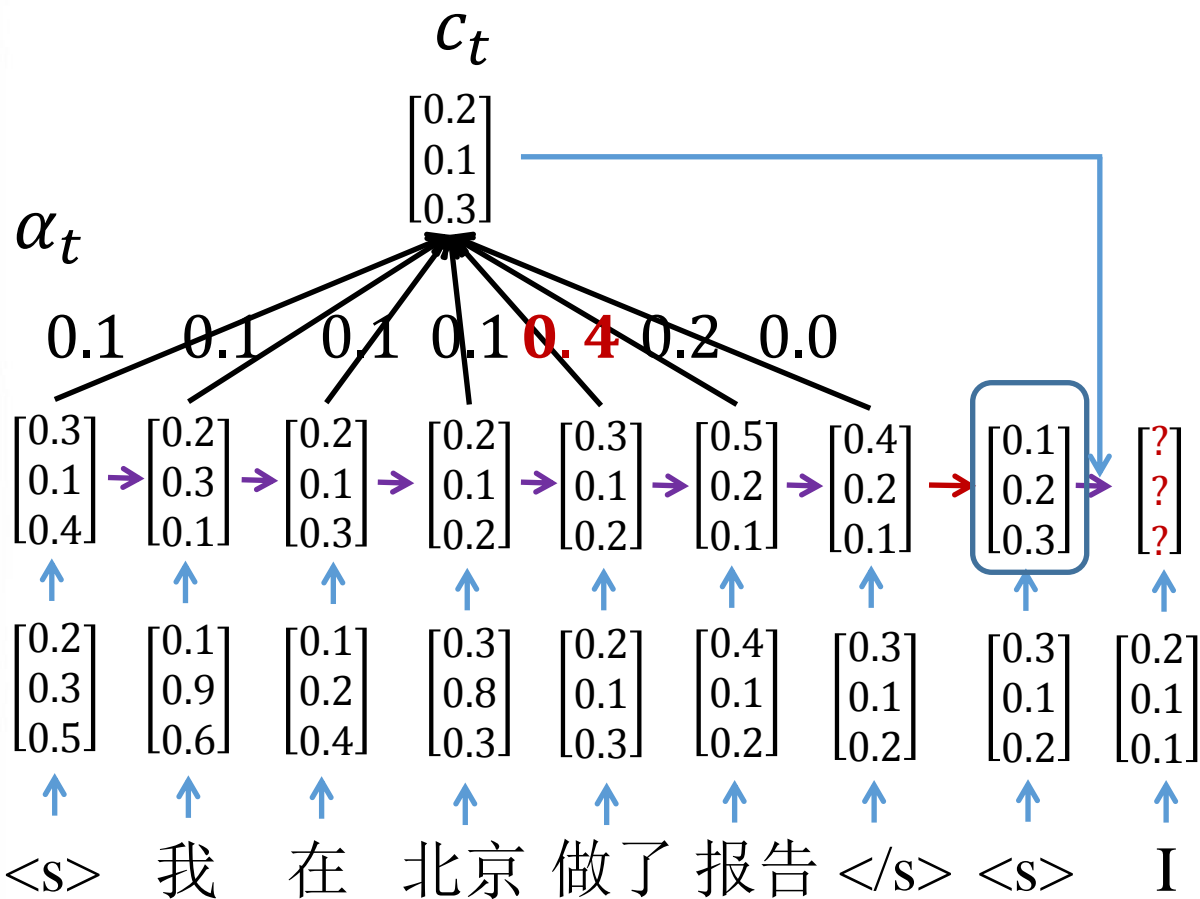


# 神经机器翻译-注意机制

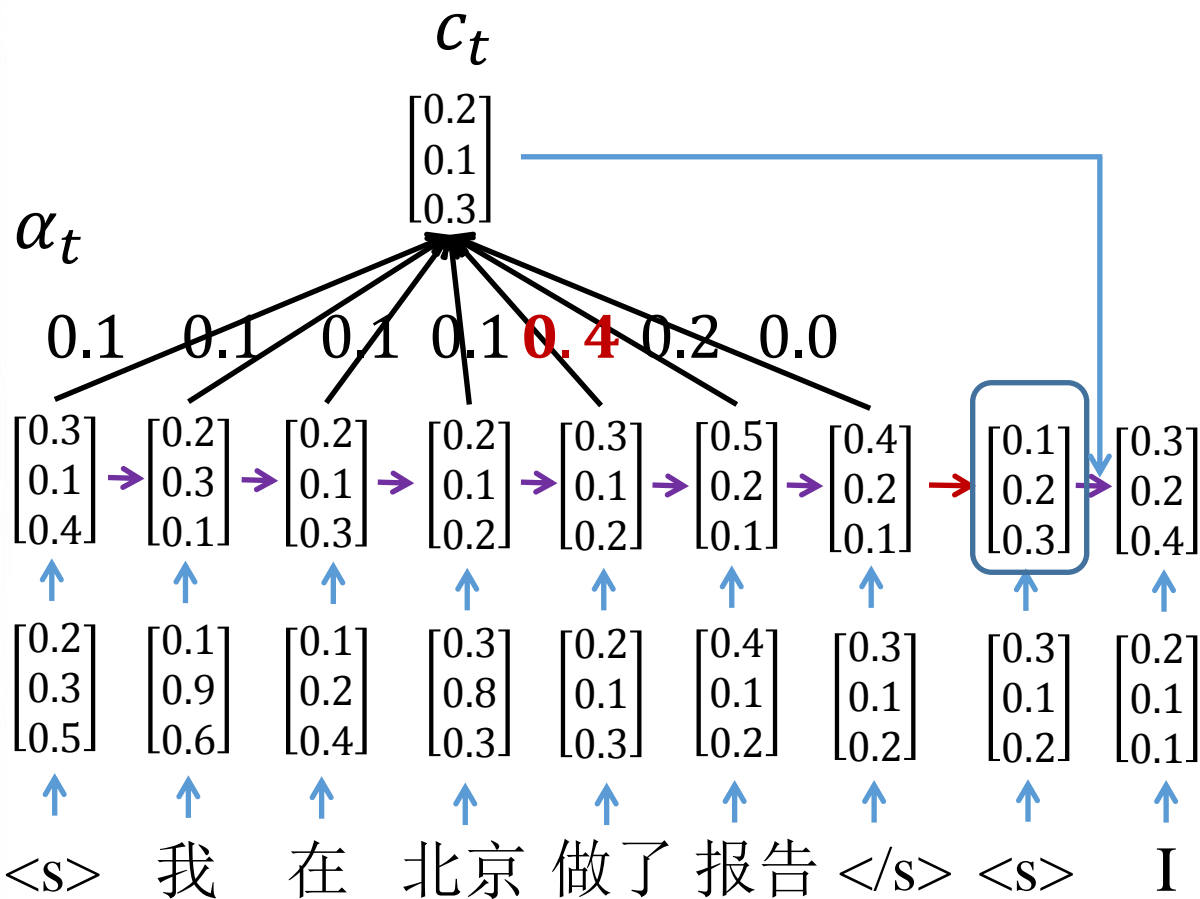
$$\text{score}(h_s, h_t)$$



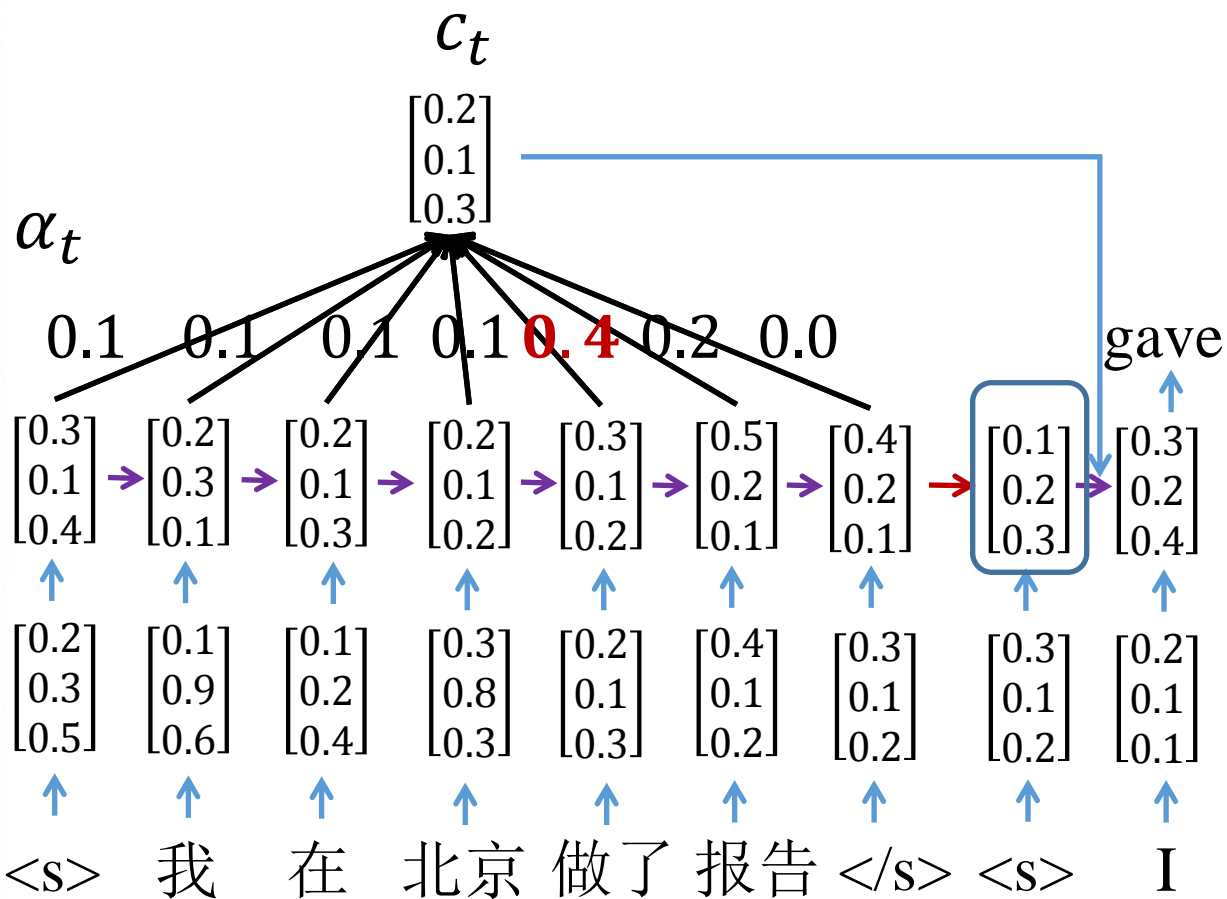
# 神经机器翻译-注意机制



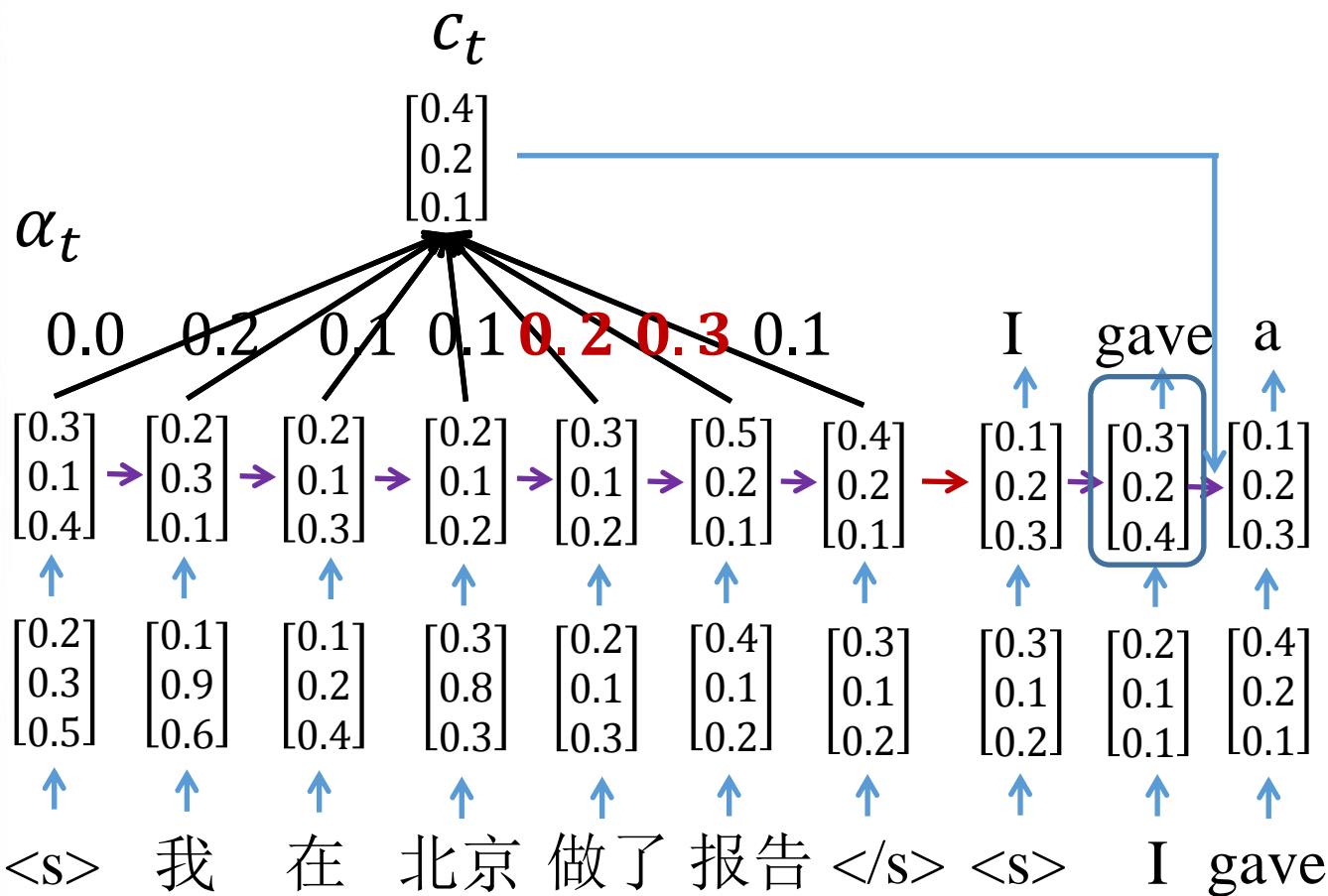
# 神经机器翻译-注意机制



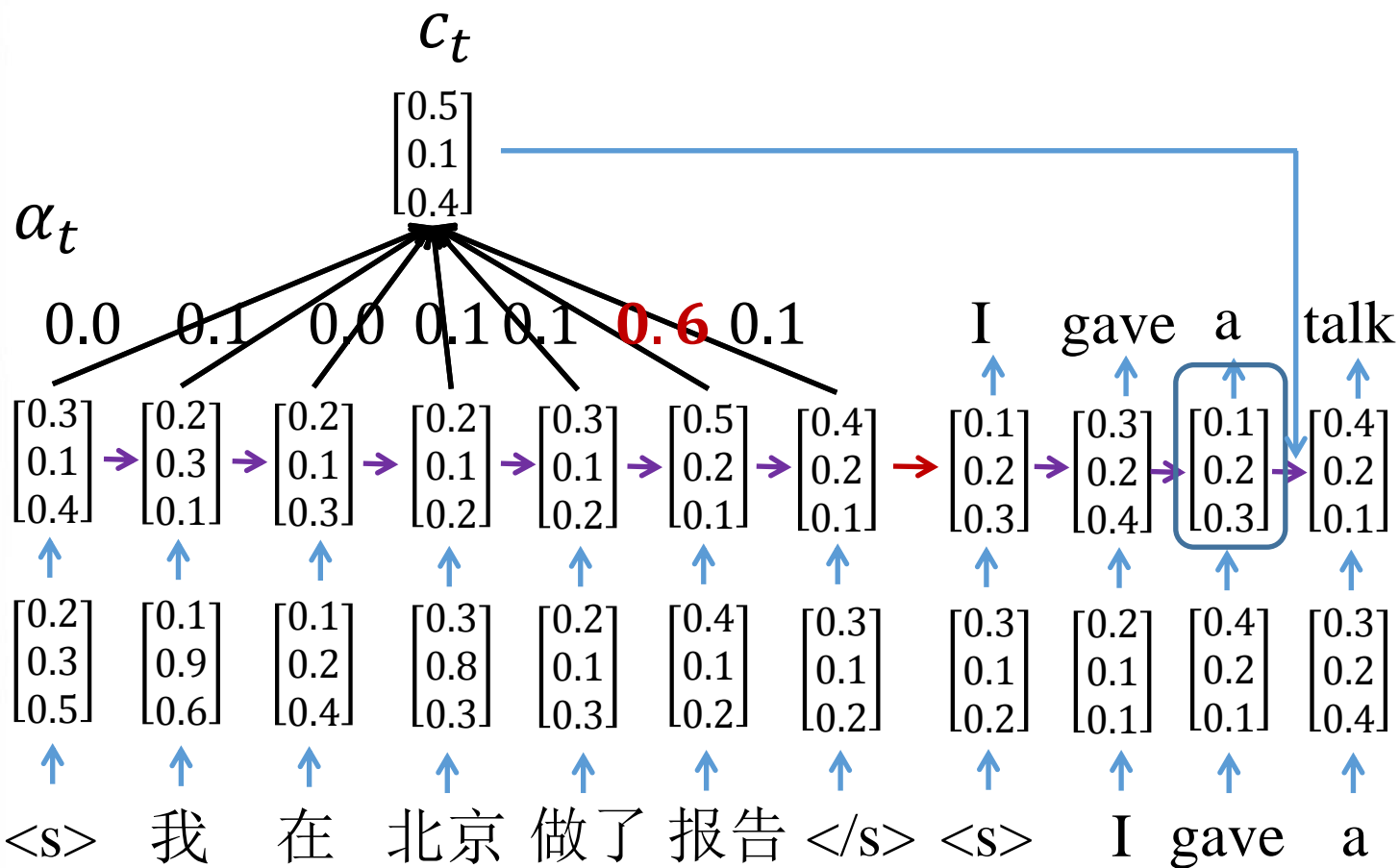
# 神经机器翻译-注意机制



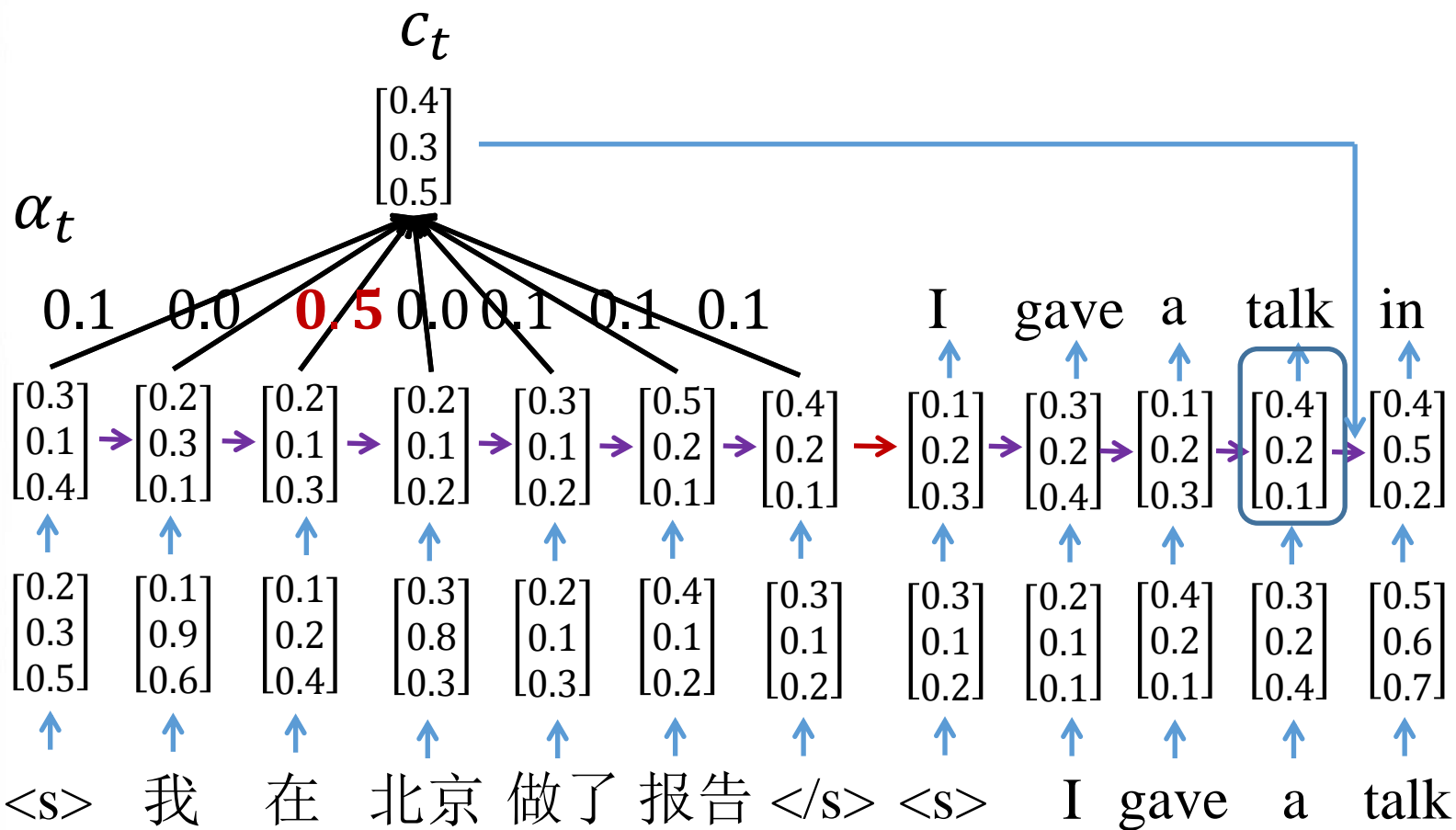
# 神经机器翻译-注意机制



# 神经机器翻译-注意机制

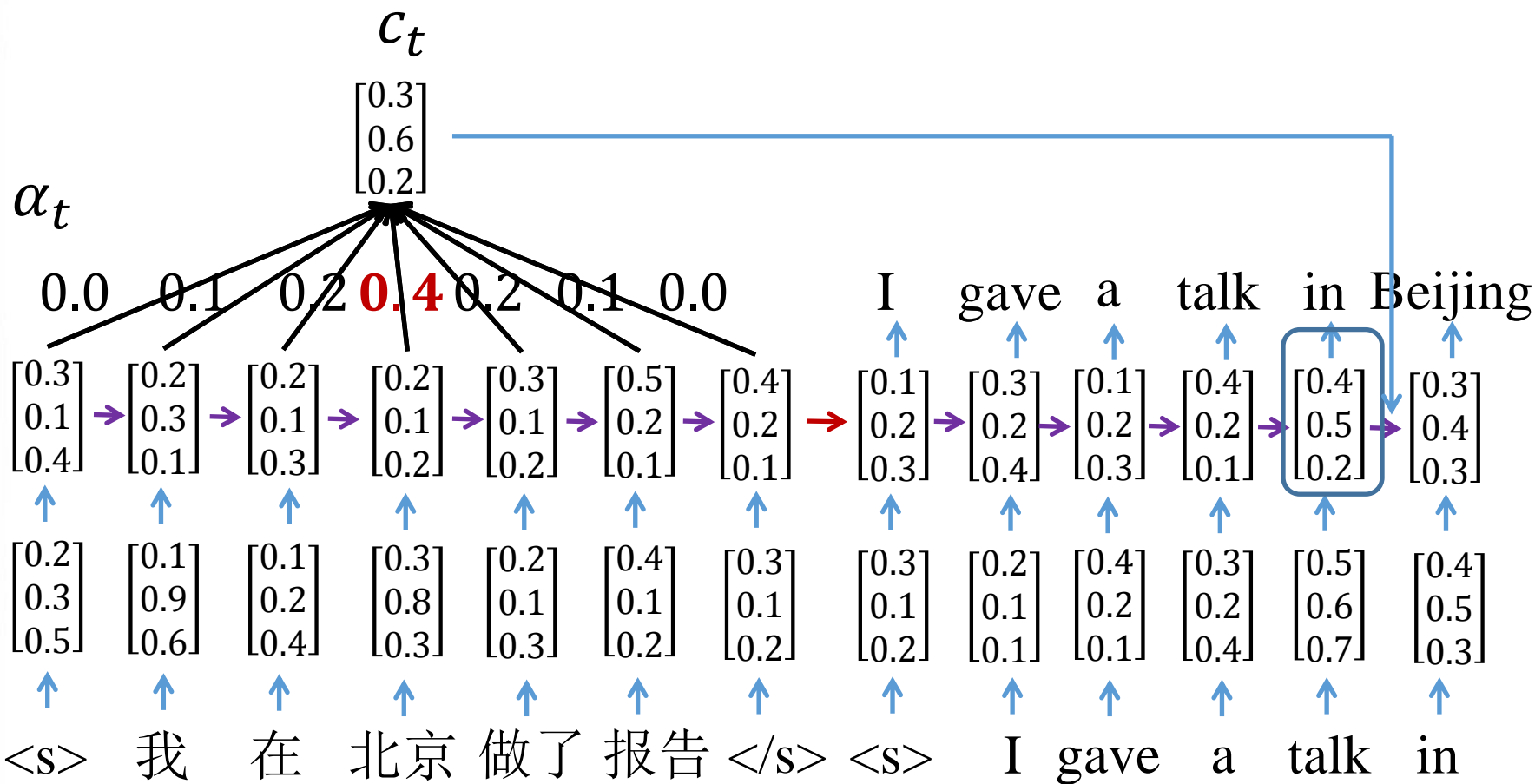


# 神经机器翻译-注意机制

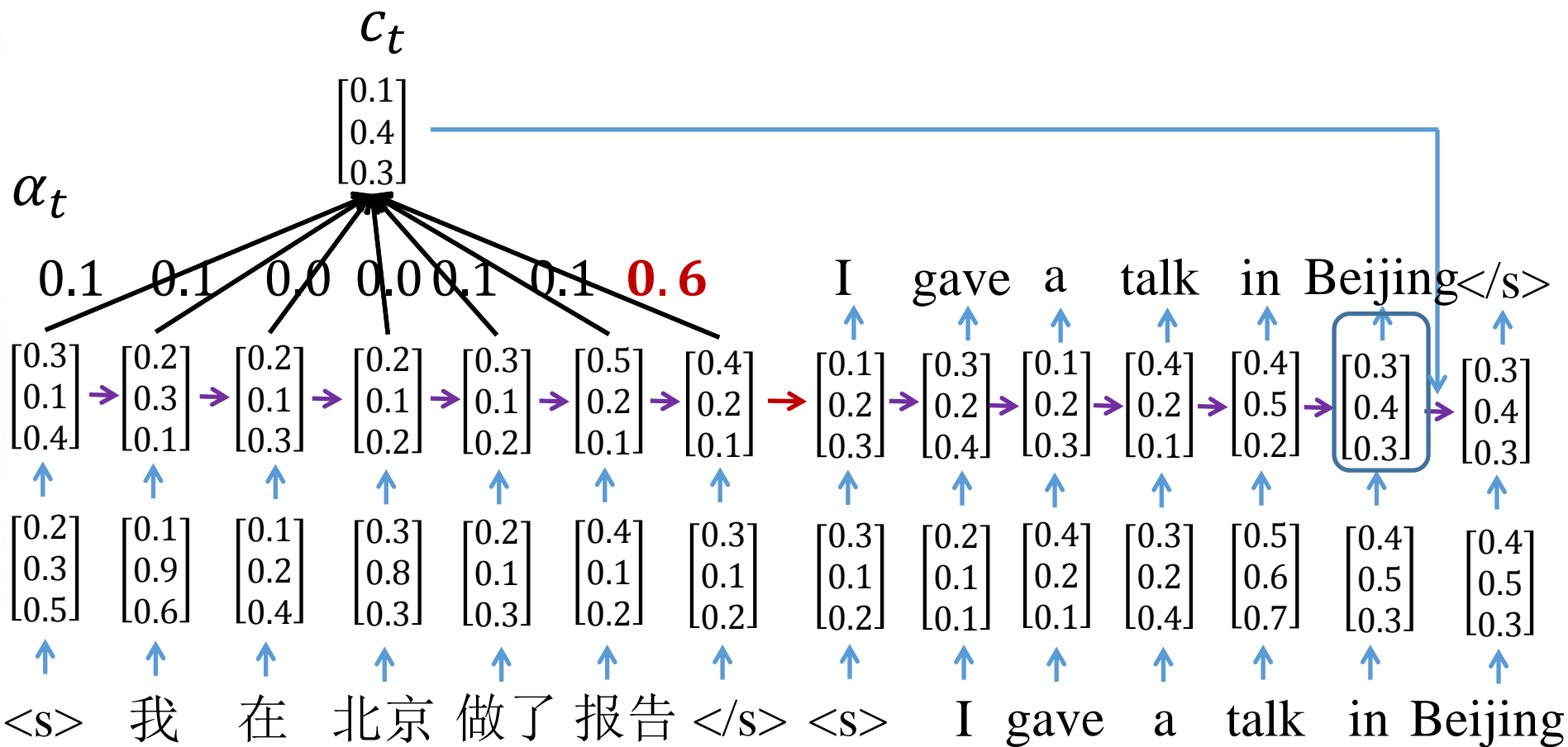




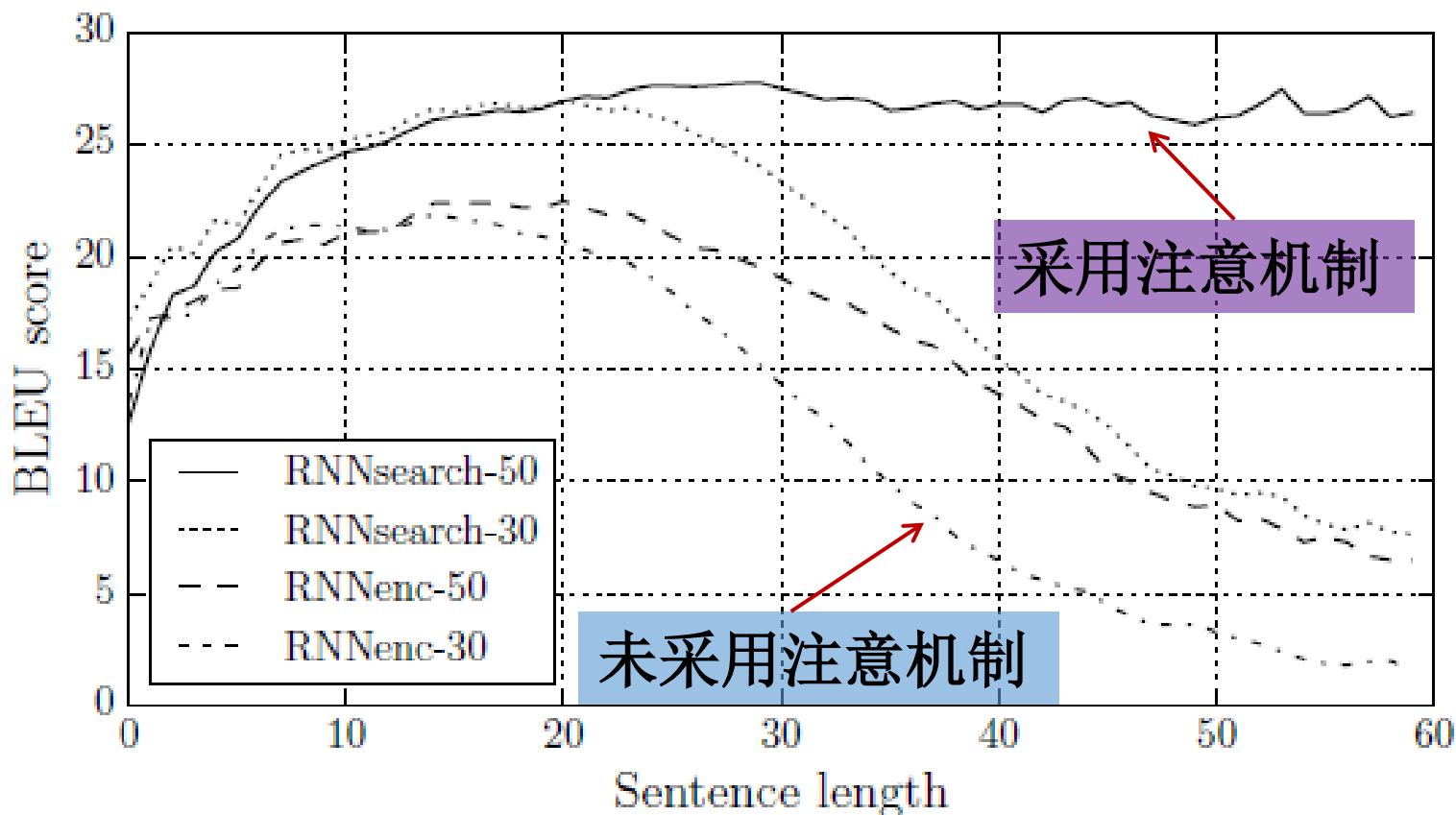
# 神经机器翻译-注意机制



# 神经机器翻译-注意机制



# 神经机器翻译-注意机制



**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
- ...

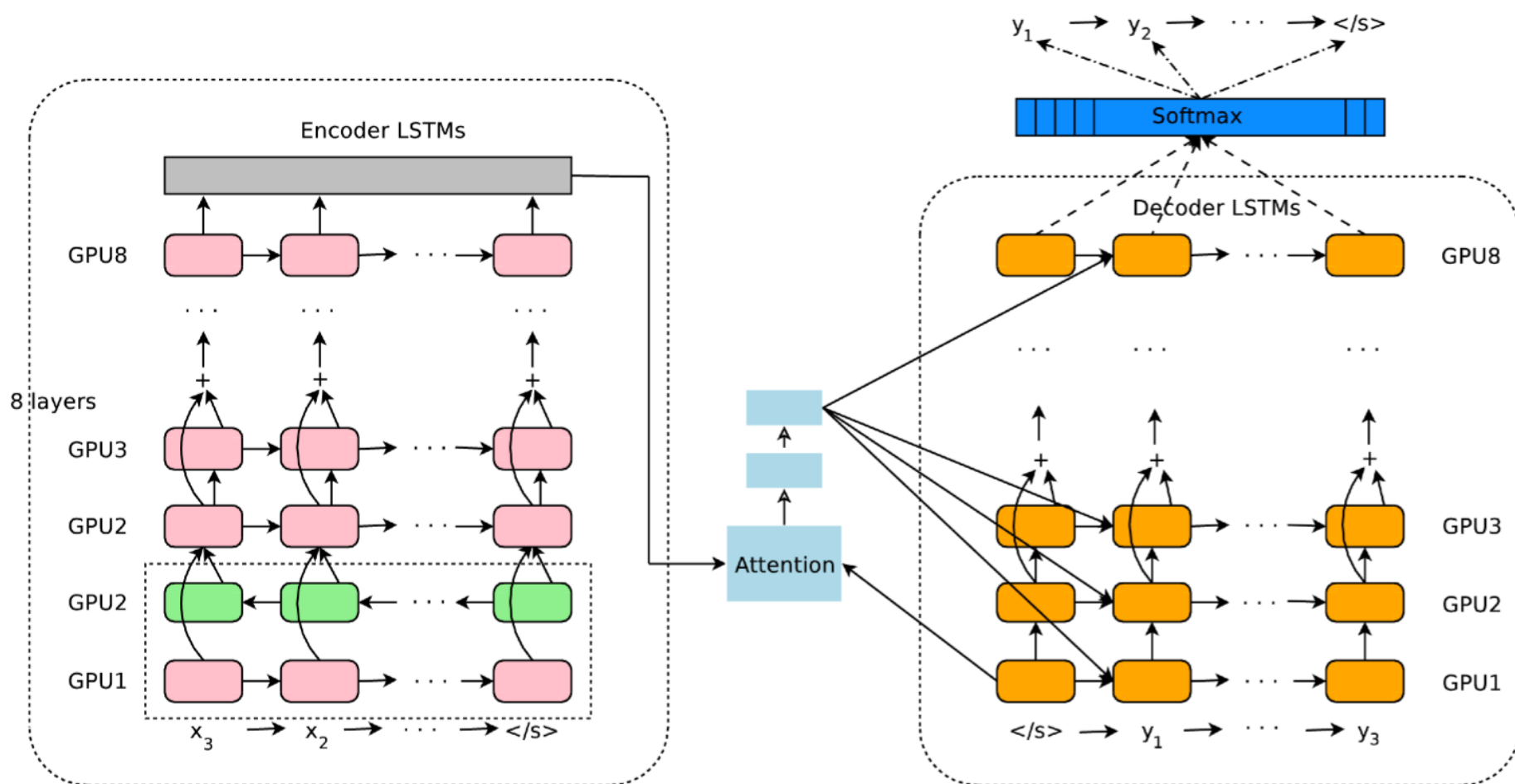
# 工业界线上产品



# 工业界线上产品



# 工业界线上产品



**GNMT:** 谷歌神经翻译系统



# 工业界线上产品

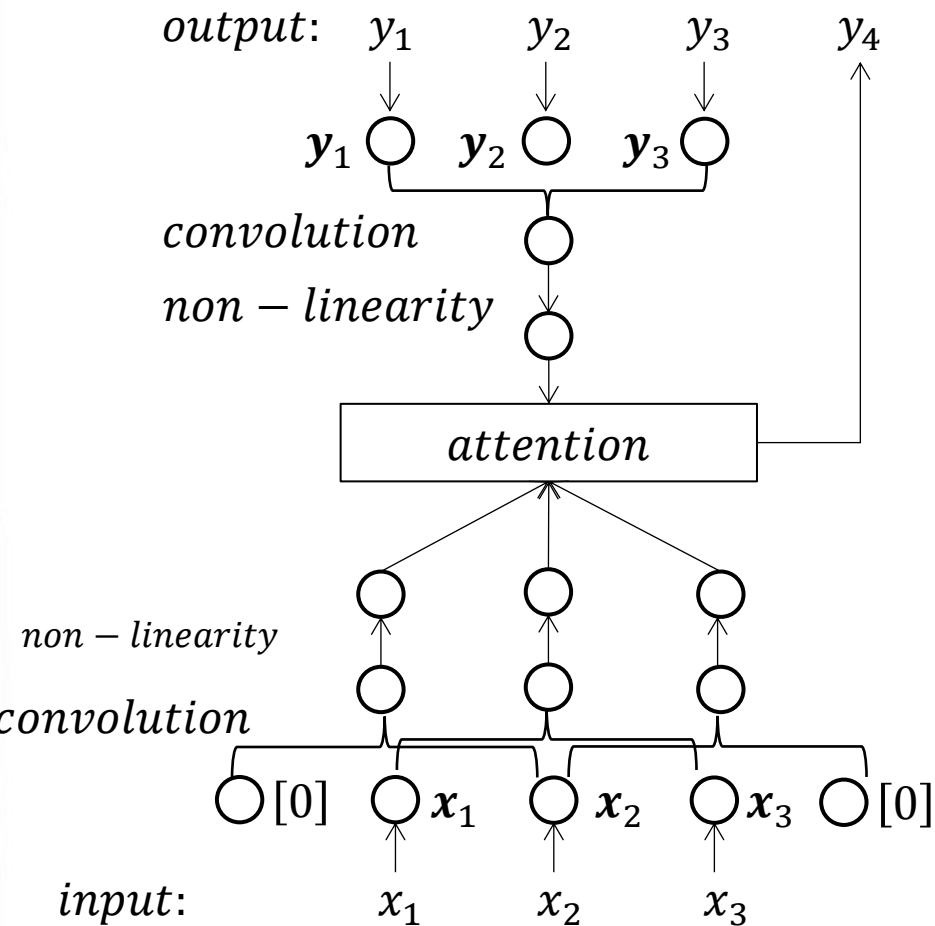
	PBMT	GNMT	Human	Relative Improvement
English → Spanish	4.885	5.428	5.550	87%
English → French	4.932	5.295	5.496	64%
English → Chinese	4.825	5.194	4.987	58%
Spanish → English	4.872	5.187	5.372	63%
French → English	5.046	5.343	5.404	83%
Chinese → English	3.694	4.263	4.636	60%

人工评测提升显著!

**GNMT:** 谷歌神经翻译系统

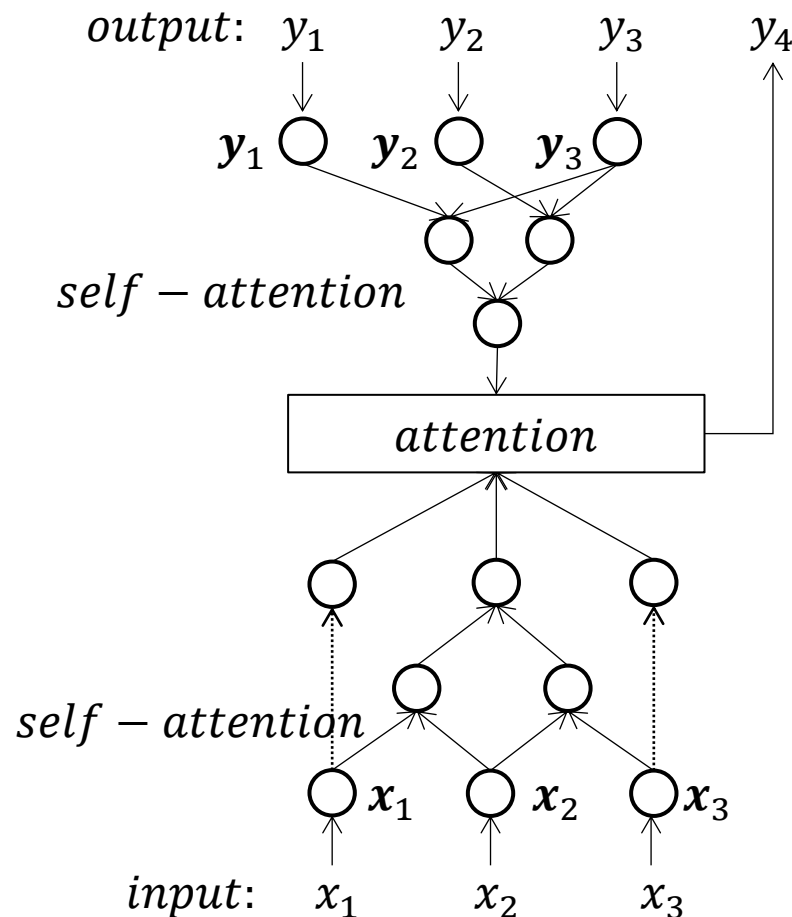
[Wu et al, 2016]<sup>61</sup>

# 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

# 参考文献

1. Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio, 2015. [Neural Translation by Jointly Learning to Align and Translate](http://arxiv.org/pdf/1409.0473.pdf). *http://arxiv.org/pdf/1409.0473.pdf*
2. Yong Cheng, Wei Xu, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. [Semi-Supervised Learning for Neural Machine Translation](#). *In Proc. of ACL 2016*.
3. Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio, 2014. [Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation](#). *In Proc. of EMNLP 2014*.
4. Daxiang Dong, HuaWu, Wei He, Dianhai Yu and Haifeng Wang, 2015. [Multi-Task Learning for Multiple Language Translation](#). *In Proc. of ACL 2015*.
5. Xiaoqing Li, Jiajun Zhang and Chengqing Zong, 2016a. [Towards Zero Unknown Word in Neural Machine Translation](#). *In Proc. of IJCAI 2016*.
6. Xiaoqing Li, Jiajun Zhang and Chengqing Zong, 2016b. [One Sentence One Model for Neural Machine Translation](https://arxiv.org/pdf/1609.06490.pdf). *https://arxiv.org/pdf/1609.06490.pdf*
7. Minh-Thang Luong, Ilya Sutskever, Quoc Le, Oriol Vinyals and Wojciech Zaremba, 2015a. [Addressing the Rare Word Problem in Neural Machine Translation](#). *In Proc. of ACL 2015*.
8. Minh-Thang Luong, Hieu Pham and Christopher D. Manning, 2015b. [Effective Approaches to Attention - based Neural Machine Translation](#). *In Proc. of EMNLP 2015*.

# 参考文献

9. Minh-Thang Luong and Christopher D. Manning, 2016. [Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models](#). *In Proc. of ACL 2016*.
10. Sébastien Jean, Kyunghyun Cho, Roland Memisevic and Yoshua Bengio, 2015. [On Using Very Large Target Vocabulary for Neural Machine Translation](#). *In Proc. of ACL 2016*.
11. 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>
12. Nal Kalchbrenner and Phil Blunsom, 2013. [Recurrent Continuous Translation Models](#). *In Proc. of EMNLP 2013*.
13. Shiqi Shen, Yong Cheng, Zhongjun He, Wei He, Hua Wu, Maosong Sun, and Yang Liu. 2016. [Minimum Risk Training for Neural Machine Translation](#). *In Proc. of ACL 2016*.
14. Rico Sennrich, Barry Haddow and Alexandra Birch, 2016. [Improving Neural Machine Translation Models with Monolingual Data](#). *In Proc. of ACL 2016*.
15. Ilya Sutskever, Oriol Vinyals and Quoc V. Le, 2014. [Sequence to Sequence Learning with Neural Networks](#). *In Proc. of NIPS 2014*.
16. Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu and Hang Li, 2016. [Modeling Coverage for Neural Machine Translation](#). *In Proc. of ACL 2016*.



# 参考文献

17. Mingxuan Wang, Zhengdong Lu, Hang Li and Qun Liu. [Memory-enhanced Decoder for Neural Machine Translation](#). *In Proc. of EMNLP 2016*.
18. Wu et al., 2016. [Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation](#). <https://arxiv.org/pdf/1609.08144.pdf>
19. Biao Zhang, Deyi Xiong, Jingsong Su, Hong Duan and Min Zhang, 2016. [Variational Neural Machine Translation](#). *In Proc. of EMNLP 2016*.
20. Jiajun Zhang and Chengqing Zong. 2016a. [Exploiting Source-side Monolingual Data in Neural Machine Translation](#). *In Proc. of EMNLP 2016*.
21. Jiajun Zhang and Chengqing Zong. 2016b. [Bridging Neural Machine Translation and Bilingual Dictionaries](#). <https://arxiv.org/pdf/1610.07272.pdf>
22. Barret Zoph and Kevin Knight, 2016. [Multi-source Neural Machine Translation](#). *In Proc. of NAACL 2016*.
23. Long Zhou, Wenpeng Hu, Jiajun Zhang and Chengqing Zong, 2017. Neural System Combination for Machine Translation. *In Proc. of ACL 2017*.
24. Jonas Gehring, Michael Auli, David Grangier, Denis Yarats and Yann N. Dauphin. Convolutional Sequence to Sequence Learning. *In Proc. of ICML 2017*.
25. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. Attention is All You Need. *In Proc. of NIPS 2017*.

N L P R



谢谢!  
Q&A