自然语言处理

Natural Language Processing

权小军 教授

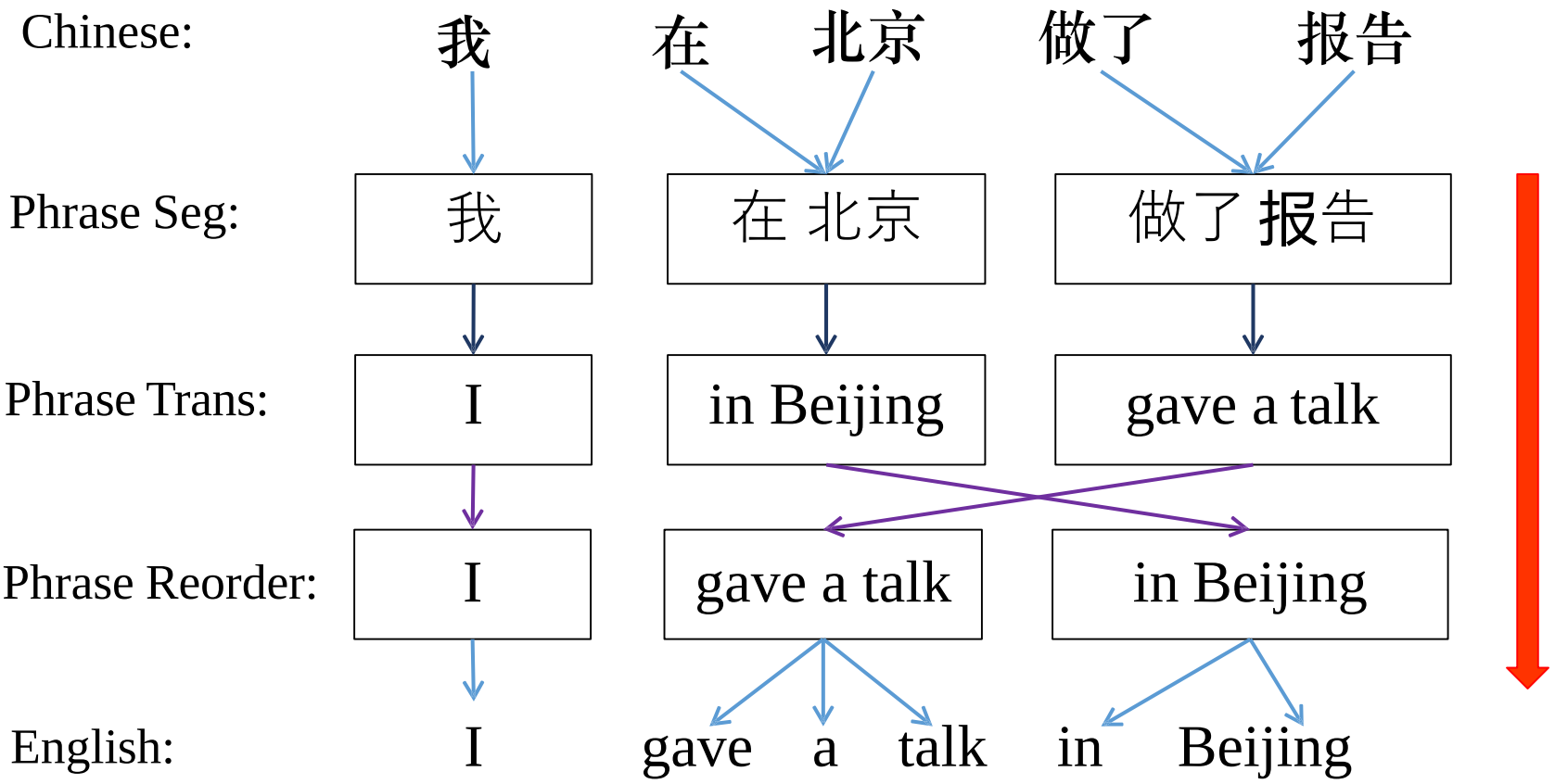
中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

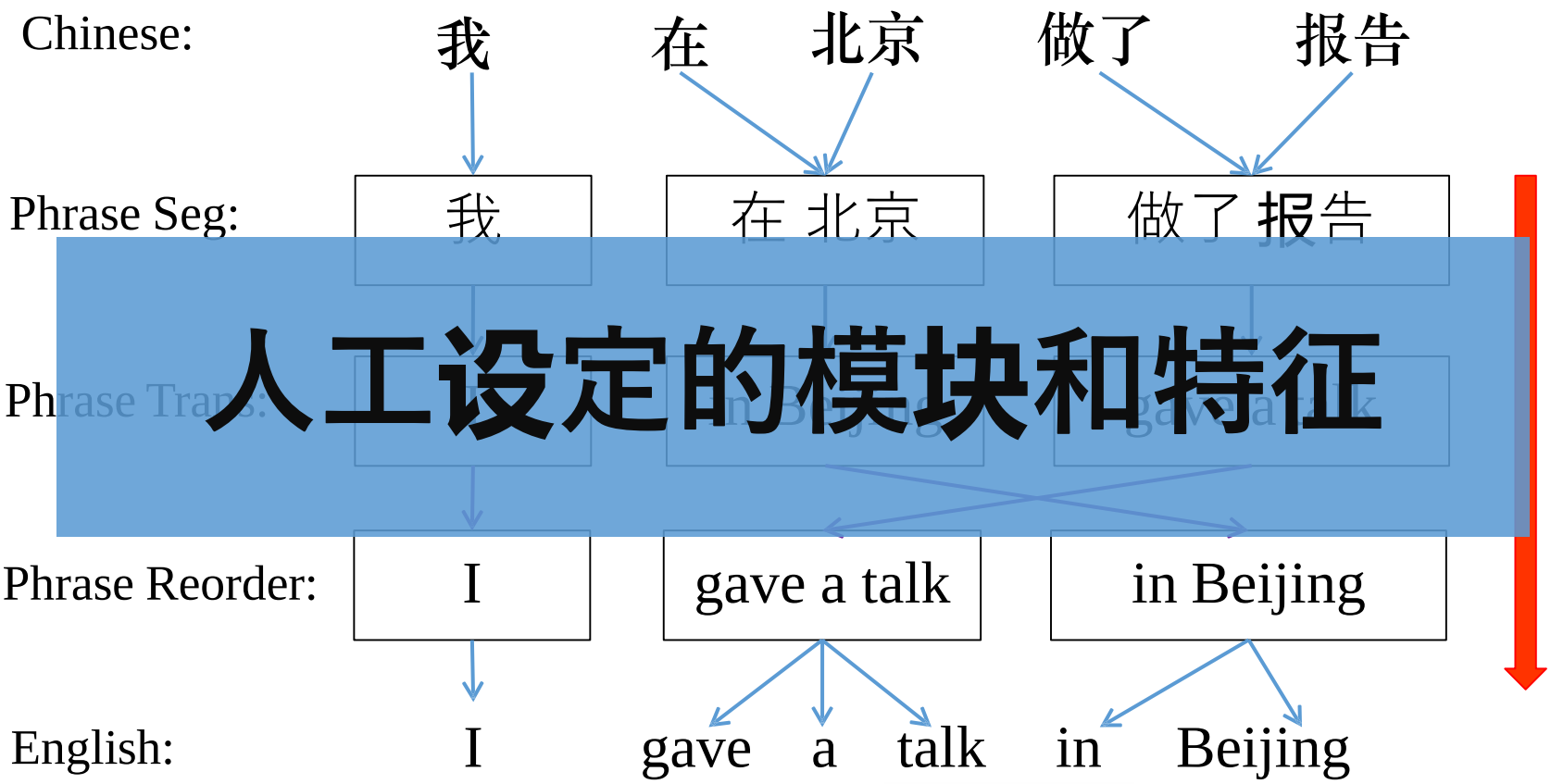
Lecture 14：机器翻译（下）

14.1: 神经机器翻译

统计机器翻译

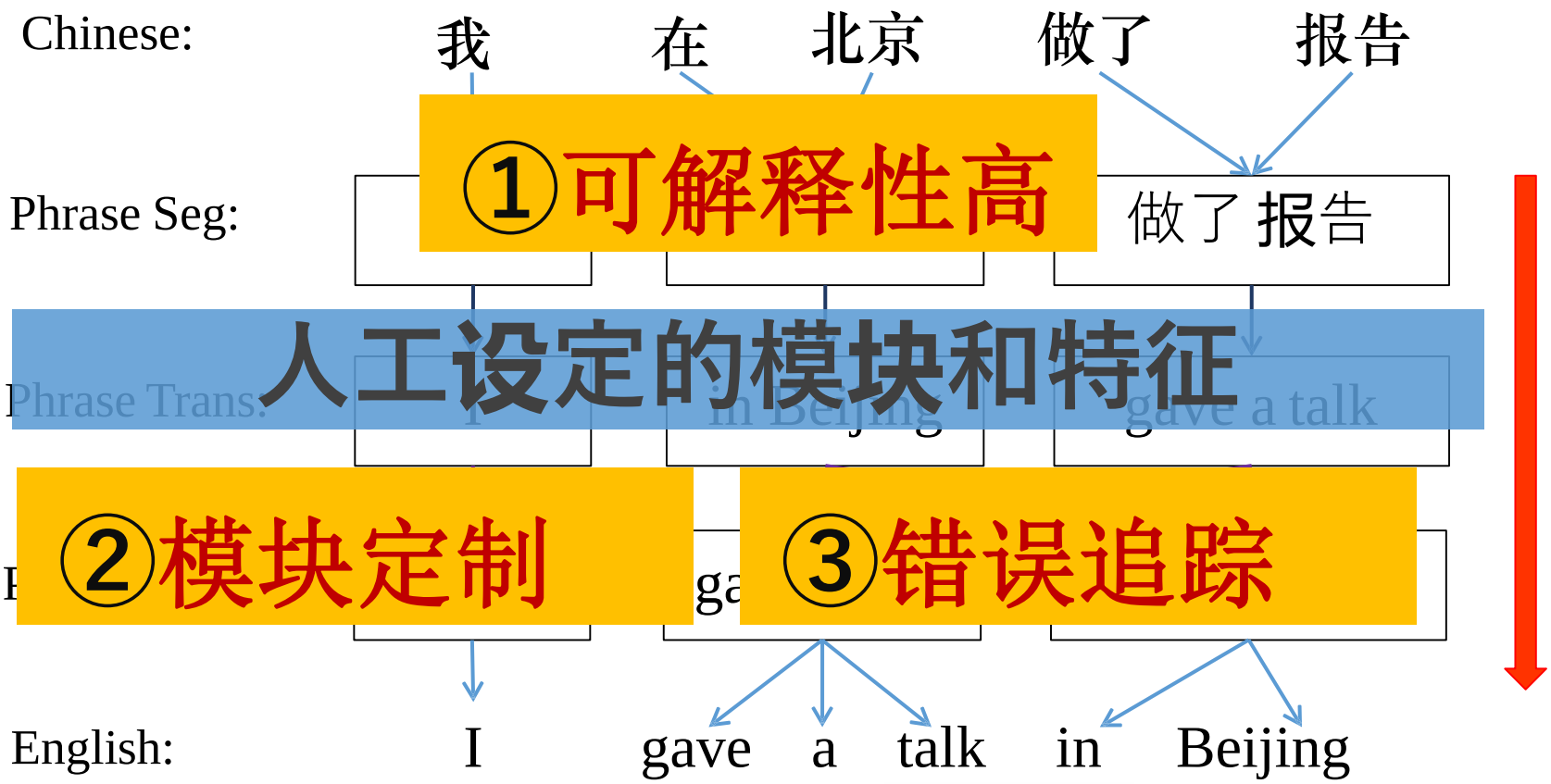

Chinese:
我
在
北京
做了
报告
Phrase Seg:
我
在 北京
做了 报告
Phrase Trans:
I
in Beijing
gave a talk
Phrase Reorder:
I
gave a talk
in Beijing
English:
I
gave
a
talk
in
Beijing

统计机器翻译

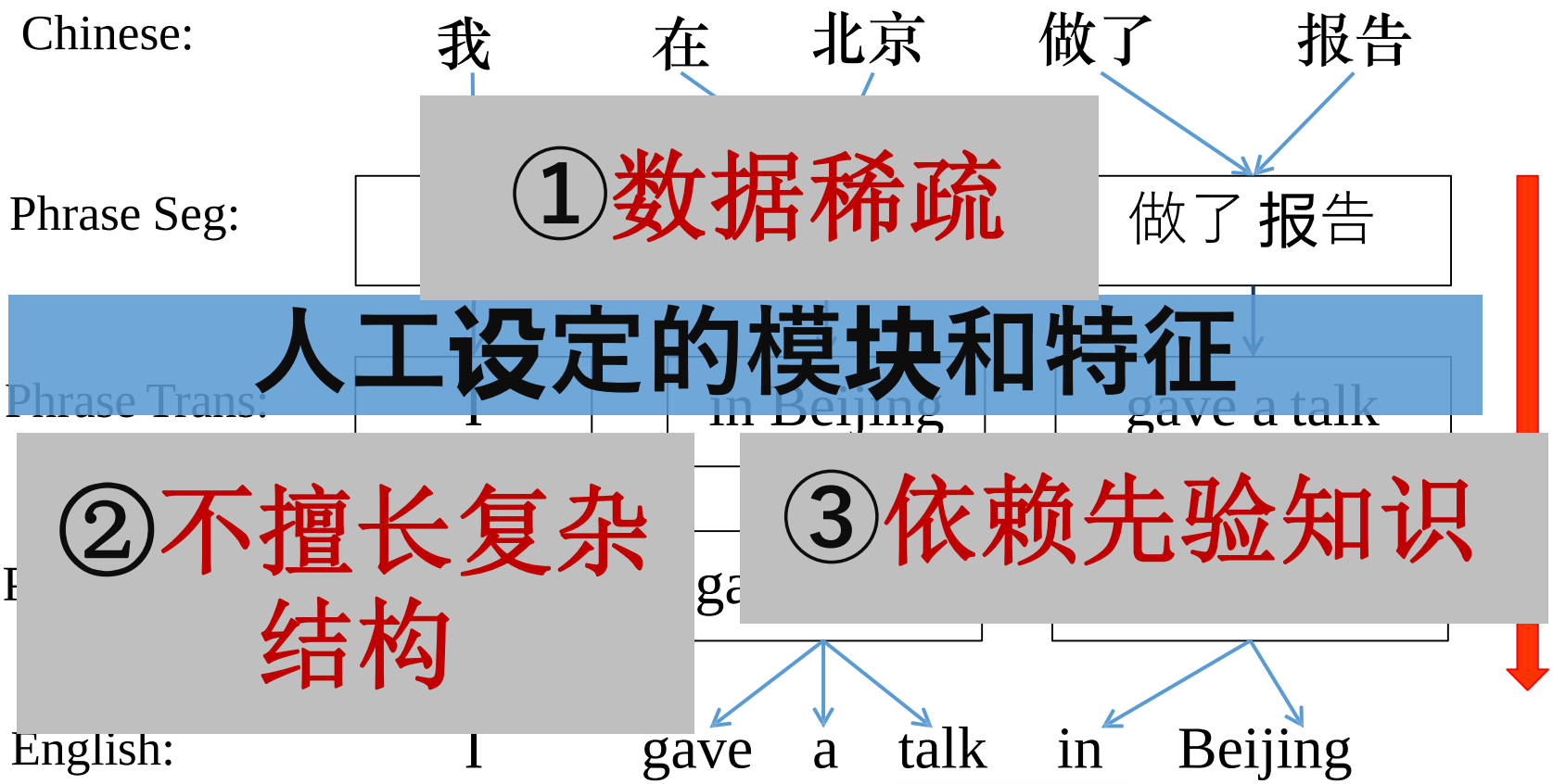


人工设定的模块和特征

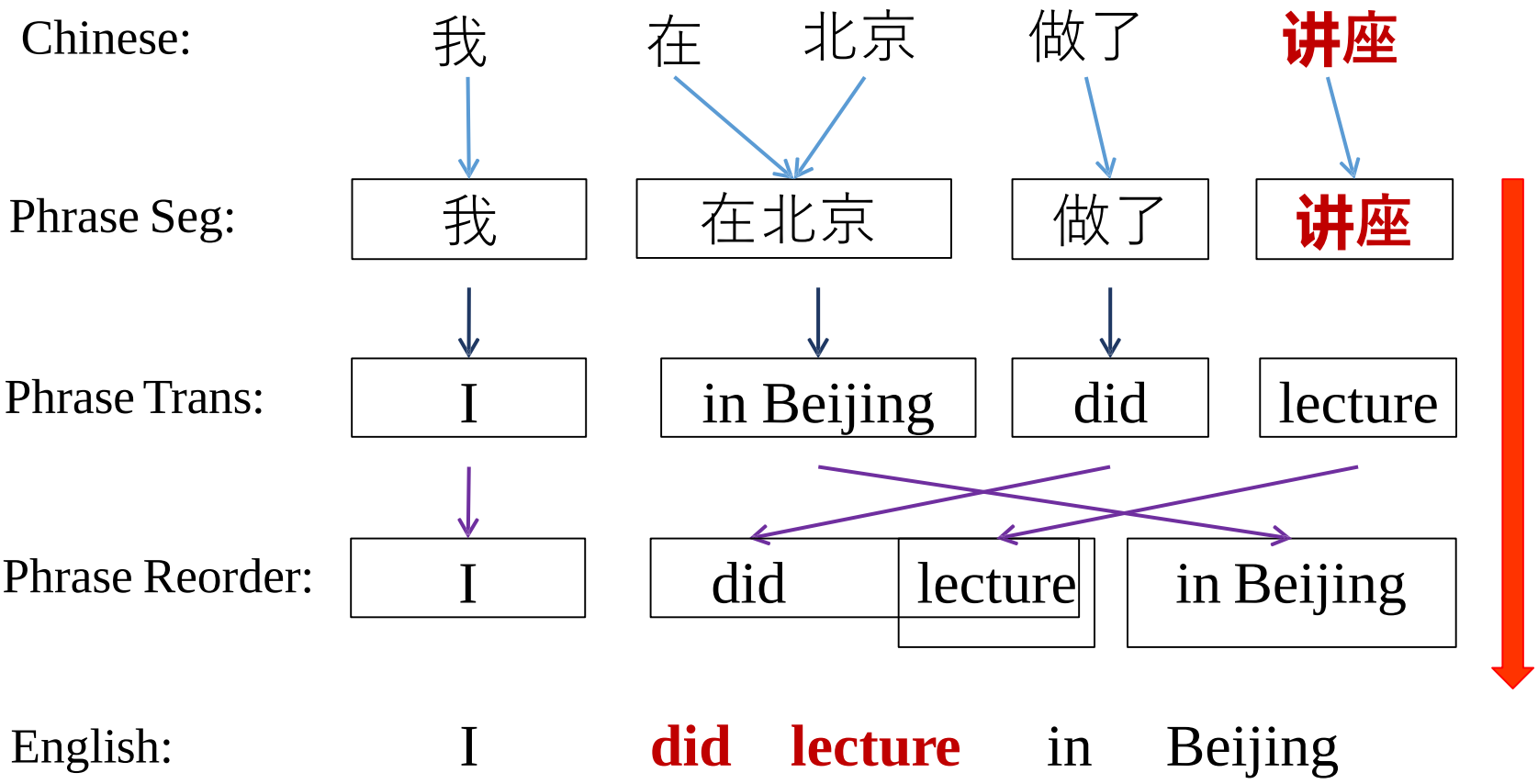
统计机器翻译


Chinese:
我 在 北京 做了 报告
Phrase Seg:
做了 报告
Phrase Trans:
English:
I gave a talk in Beijing
①可解释性高
人工设定的模块和特征
②模块定制
③错误追踪

统计机器翻译


Chinese:
我 在 北京 做了 报告
Phrase Seg:
做了 报告
English:
I gave a talk in Beijing
①数据稀疏
人工设定的模块和特征
②不擅长复杂结构
③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不擅长复杂结构

统计机器翻译→神经机器翻译

离散符号表示方法

讲座 ⊗ **报告** = 0

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

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

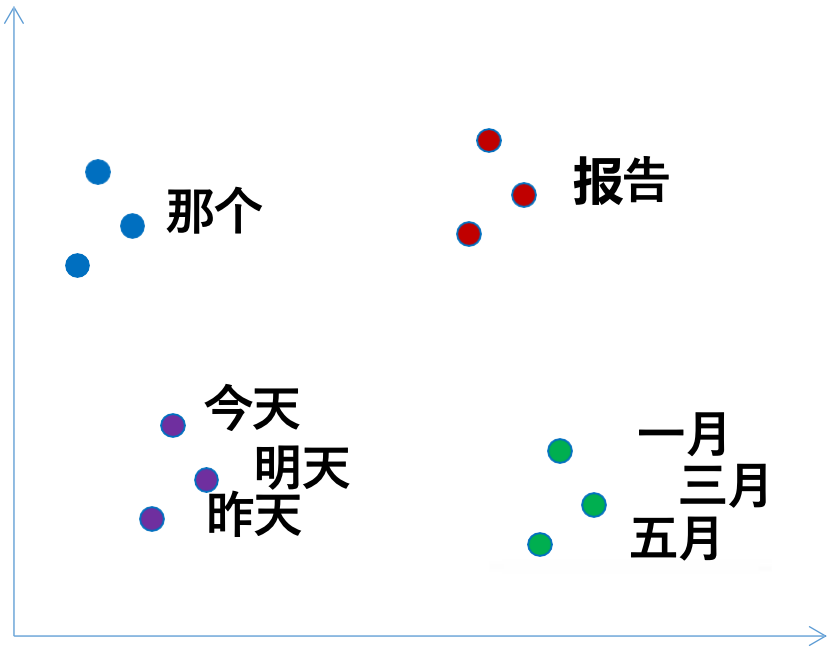
统计机器翻译→神经机器翻译

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

统计机器翻译→神经机器翻译

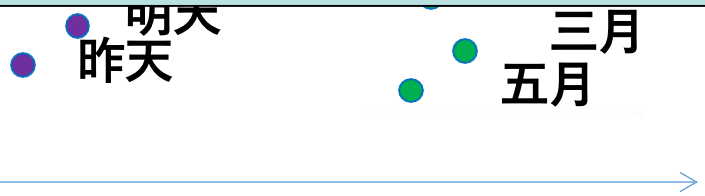
离散符号表示方法 ⇨ 连续分布式表示方法

讲座 ⊗ 报告 = 0

讲座 报告

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

分布式的语义表示是统计机器翻译到神经机器翻译的核心

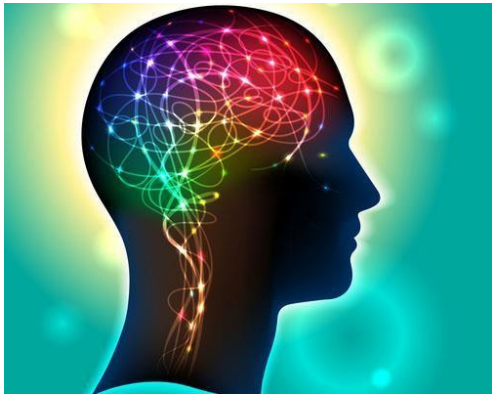


低维、稠密的连续实数空间

神经机器翻译

Chinese: 我 在 北京 做了 报告

编码网络



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

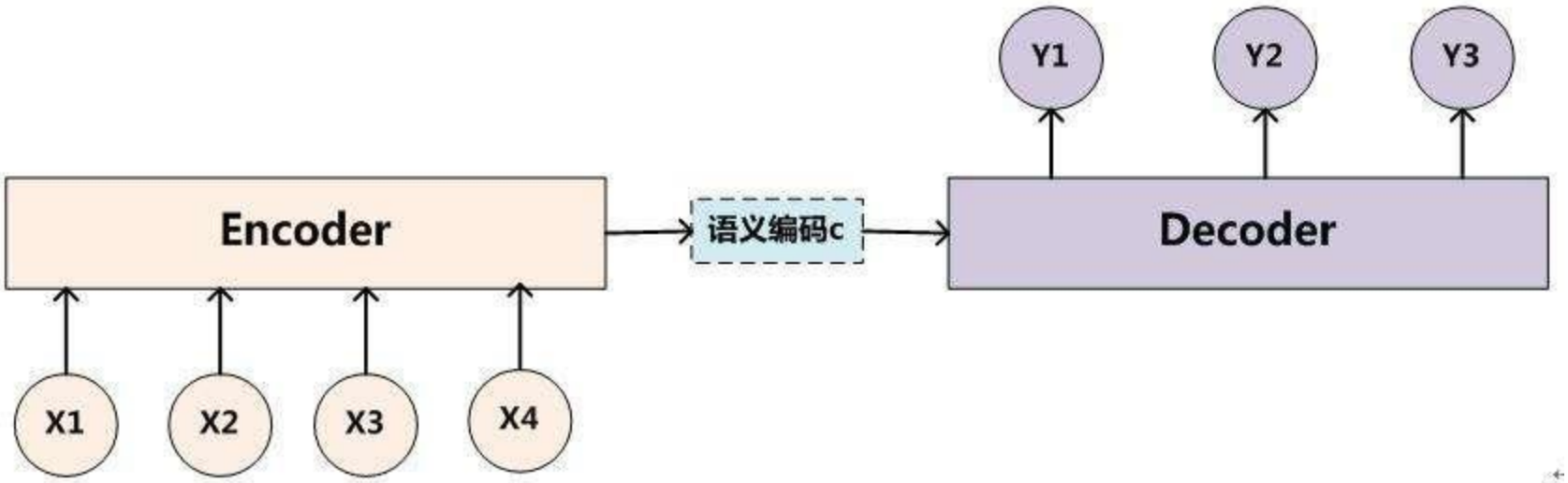
编码网络

仅需要两个神经网络

解码网络

English: I gave a talk in Beijing

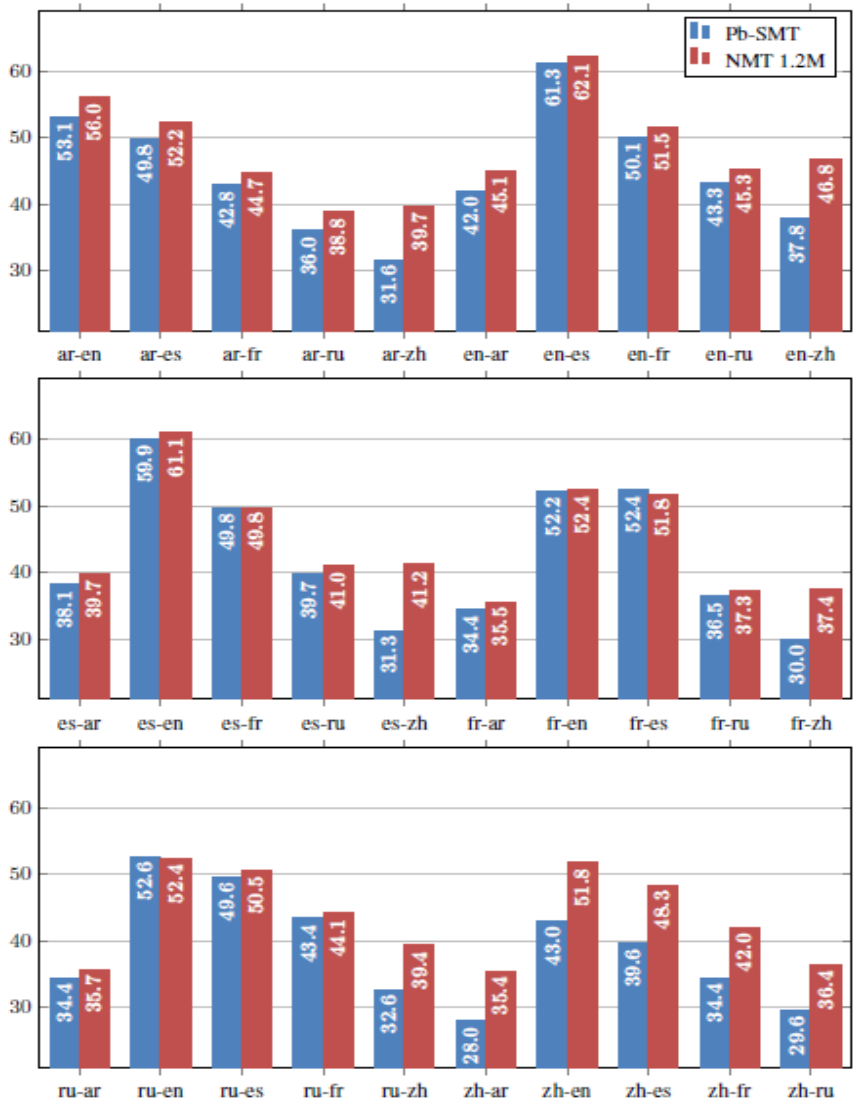
神经机器翻译



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

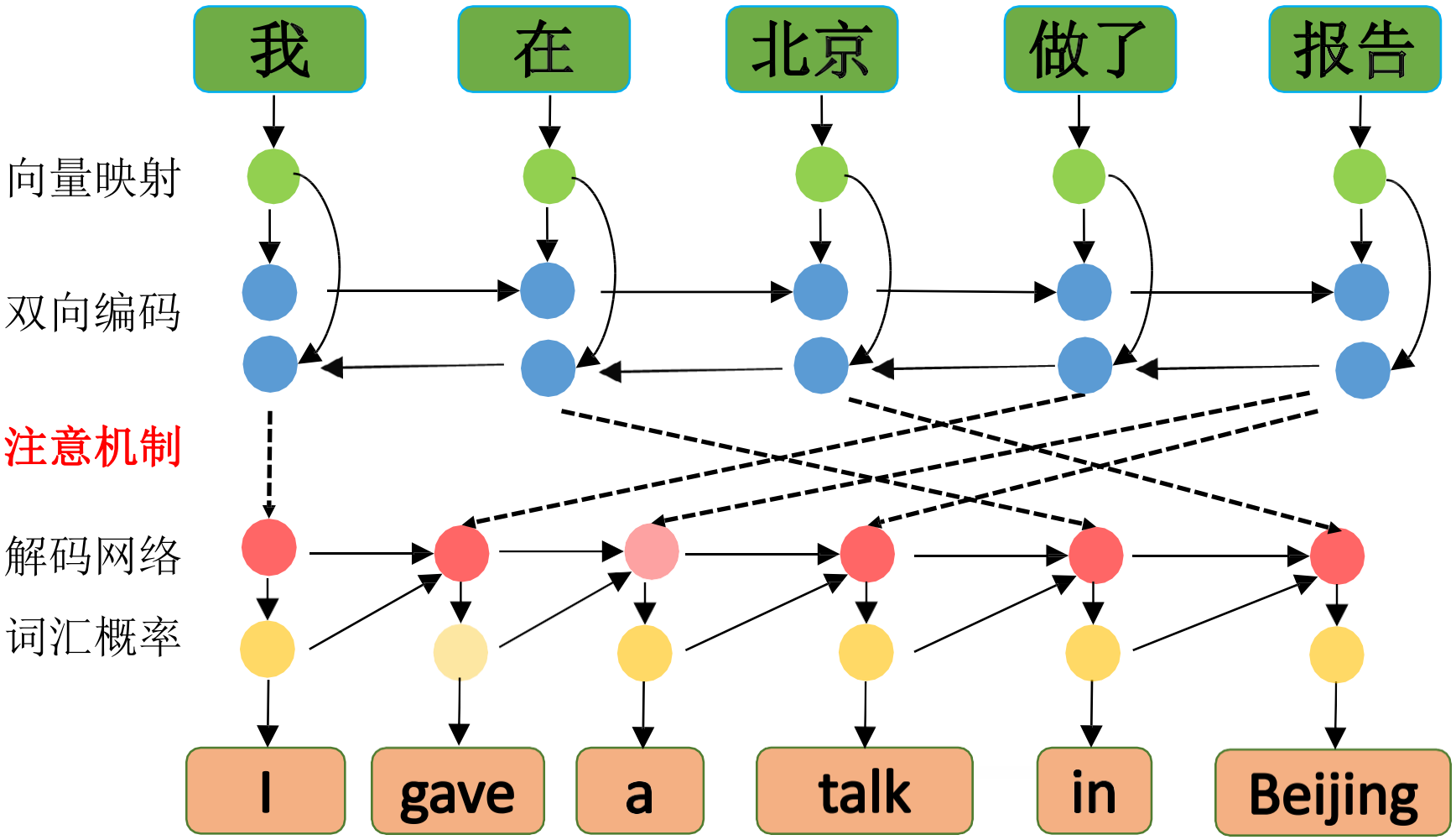
离散符号表示方法 ⇨ 连续分布式表示方法

讲座 报告

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

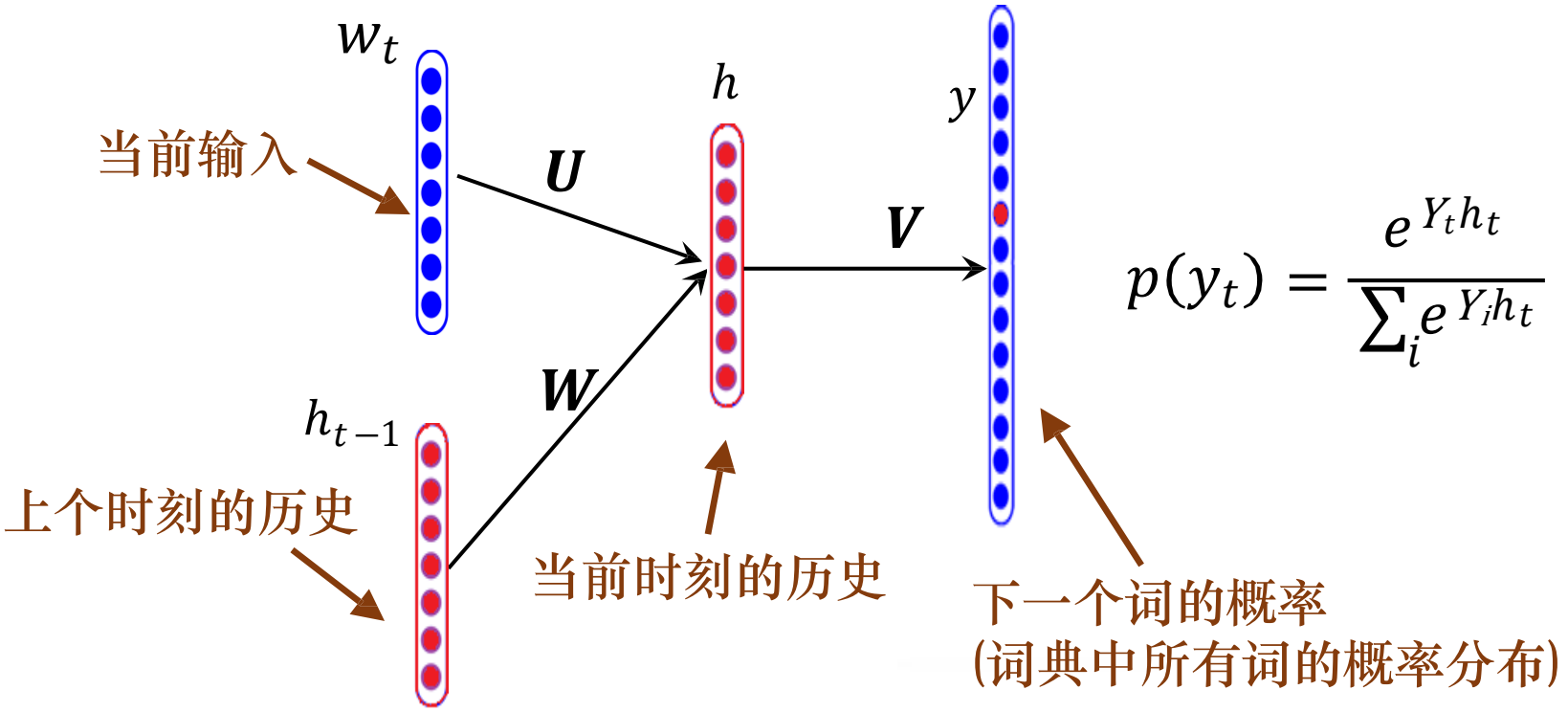
表示是核心运算是关键

神经机器翻译



循环神经网络

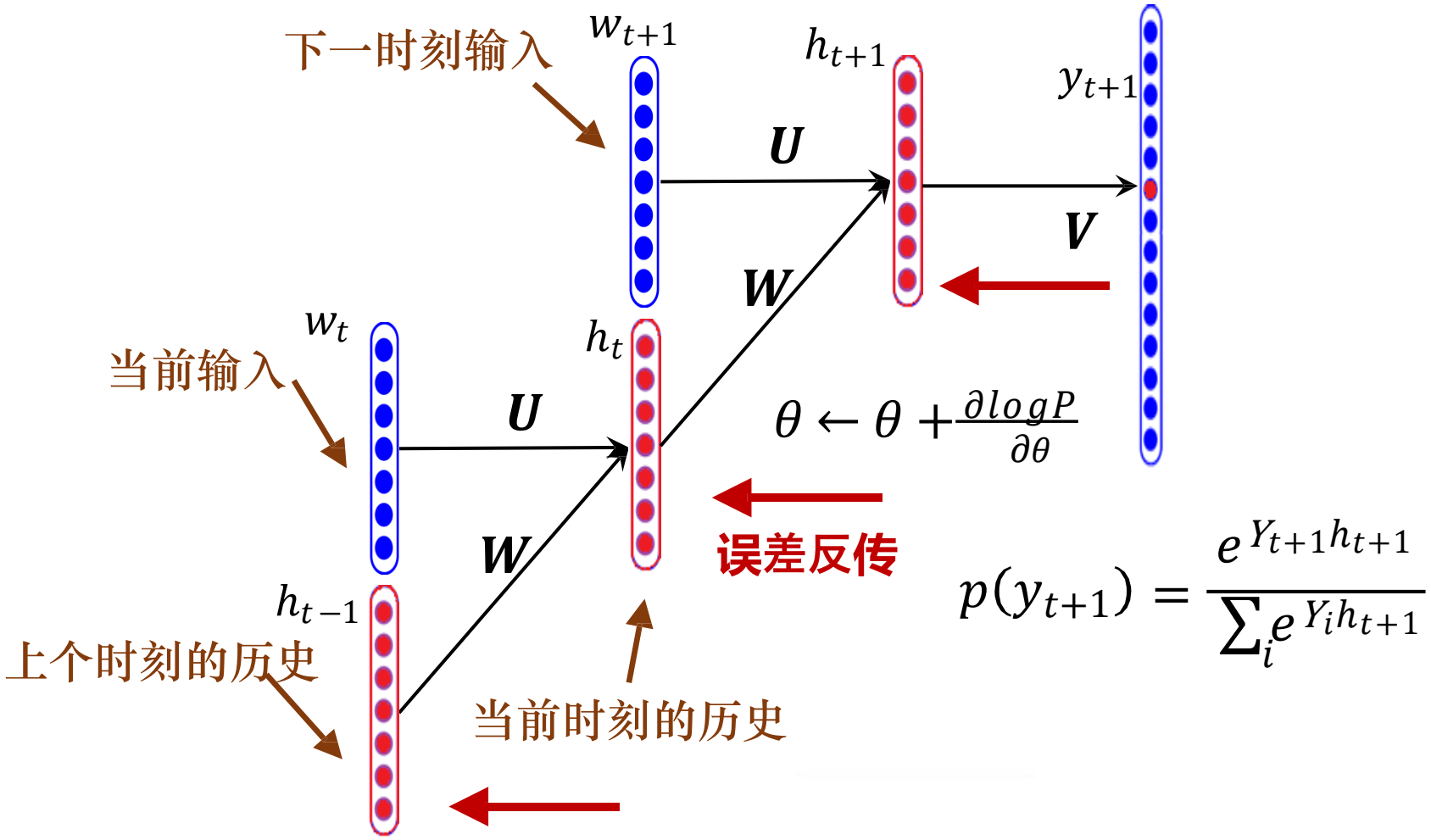
- 输入：$t-1$时刻历史h_{t-1} 与 t时刻输入w_t
- 输出：t时刻历史h_t 与 下个时刻$t+1$输入y_t 的概率



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

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

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
当前输入
h_t
U
$\theta \leftarrow \theta + \frac{\partial log P}{\partial \theta}$
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

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

$w_s \longrightarrow \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in R^3$ 我 $\longrightarrow \begin{bmatrix} 0.1 \\ 0.9 \\ 0.6 \end{bmatrix}$ 随机初始化/预训练

神经机器翻译

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

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

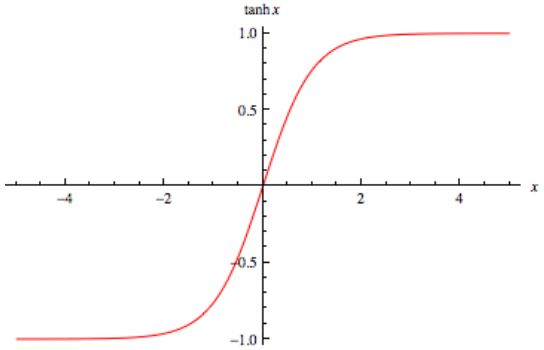
$$U = \begin{bmatrix} 0.1 & 0.2 & 0.1 \\ 0.3 & 0.2 & 0.0 \\ 0.4 & 0.0 & 0.2 \end{bmatrix} \in R^{3\times 3} \qquad W = \begin{bmatrix} 0.0 & 0.3 & 0.2 \\ 0.1 & 0.1 & 0.3 \\ 0.0 & 0.4 & 0.1 \end{bmatrix} \in R^{3\times 3}$$

神经机器翻译

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

$$z = Uw_s + Wh_{s-1} \in R^3$$

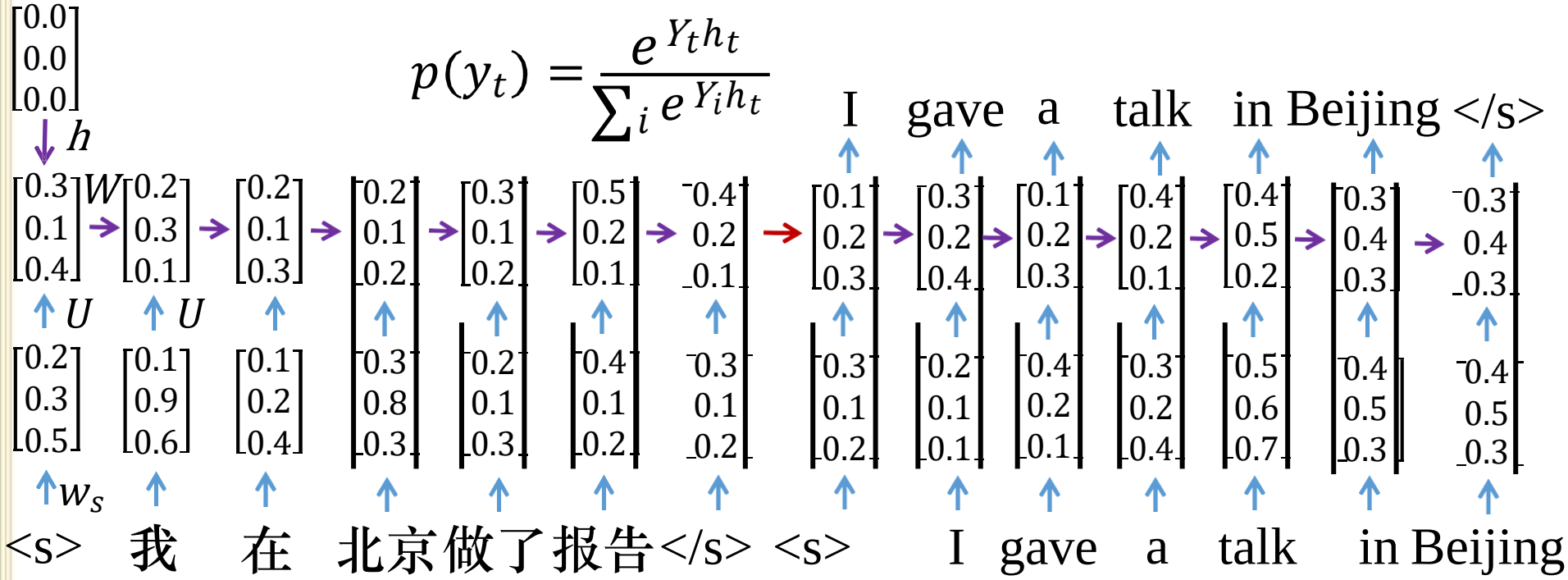
$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$$



神经机器翻译

$$h_s = tanh(Uw_s + Wh_{s-1}) \quad h_t = tanh\,(Uw_t + Wh_{t-1}\,)$$

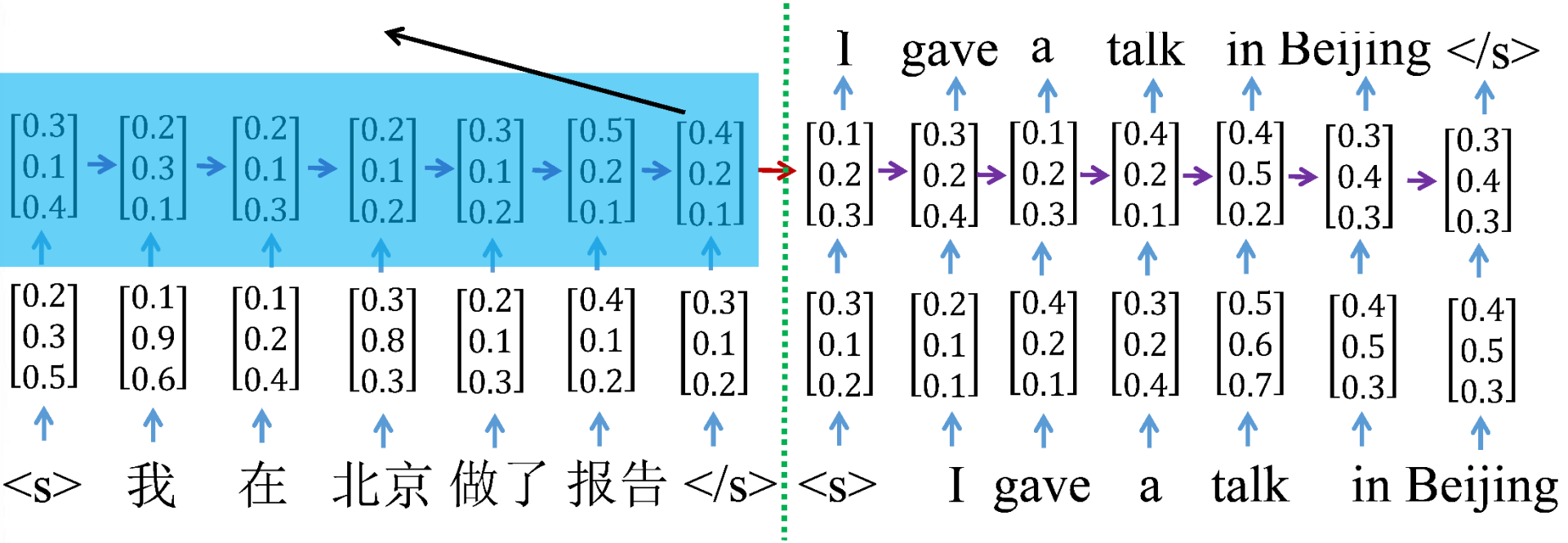
$$p(y_t) = \frac{e^{Y_t h_t}}{\sum_i e^{Y_i h_t}}$$



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

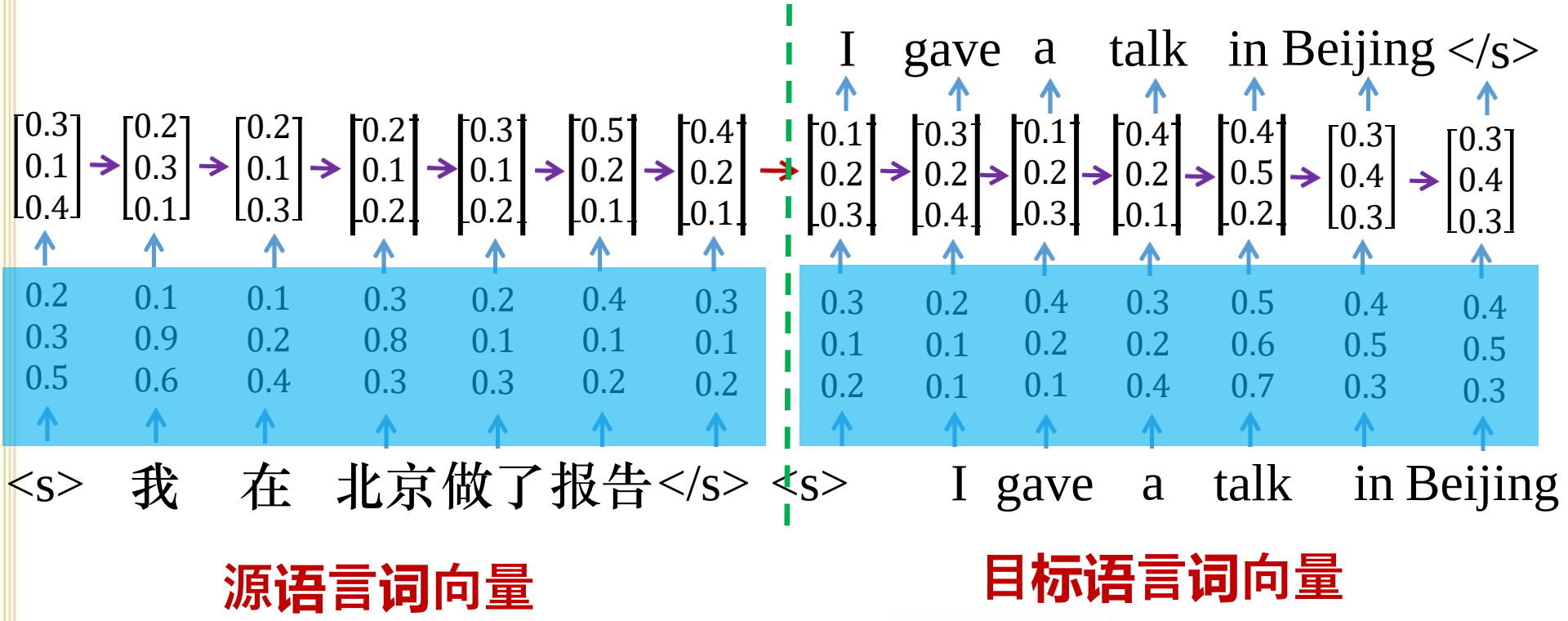


编码器

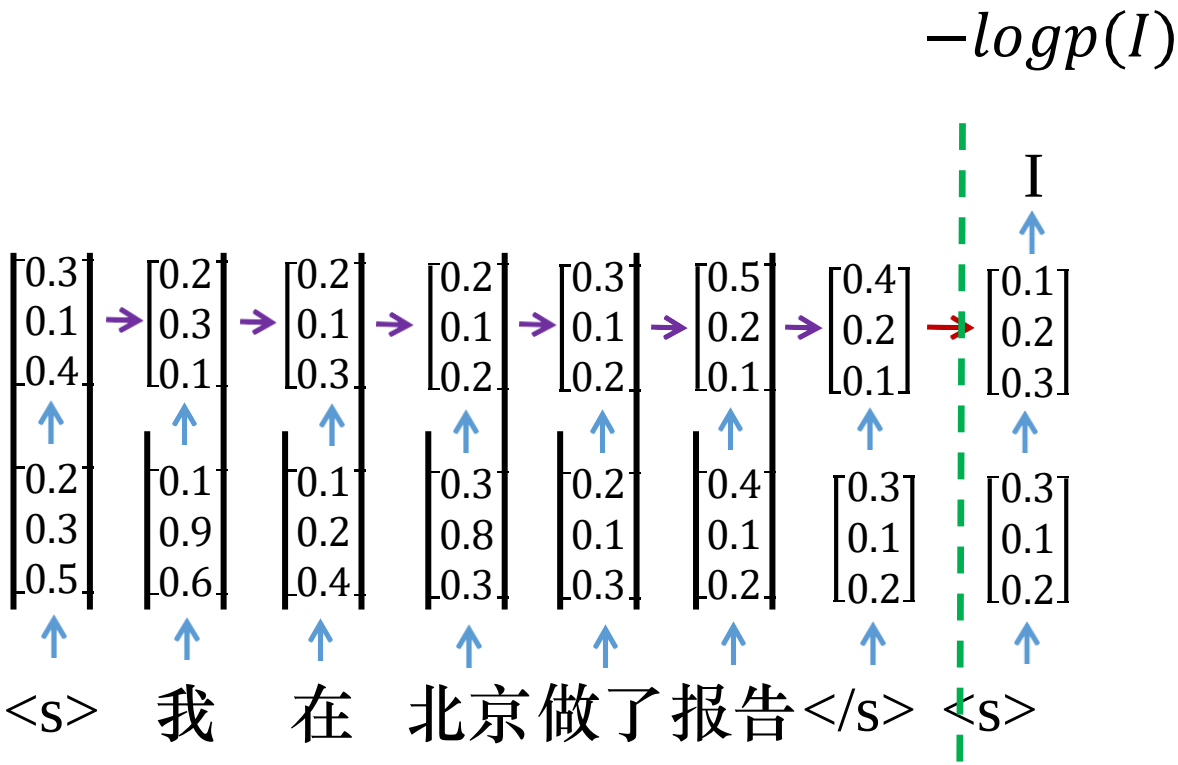
解码器

神经机器翻译

词向量可以随机初始化，在训练过程中进行优化！

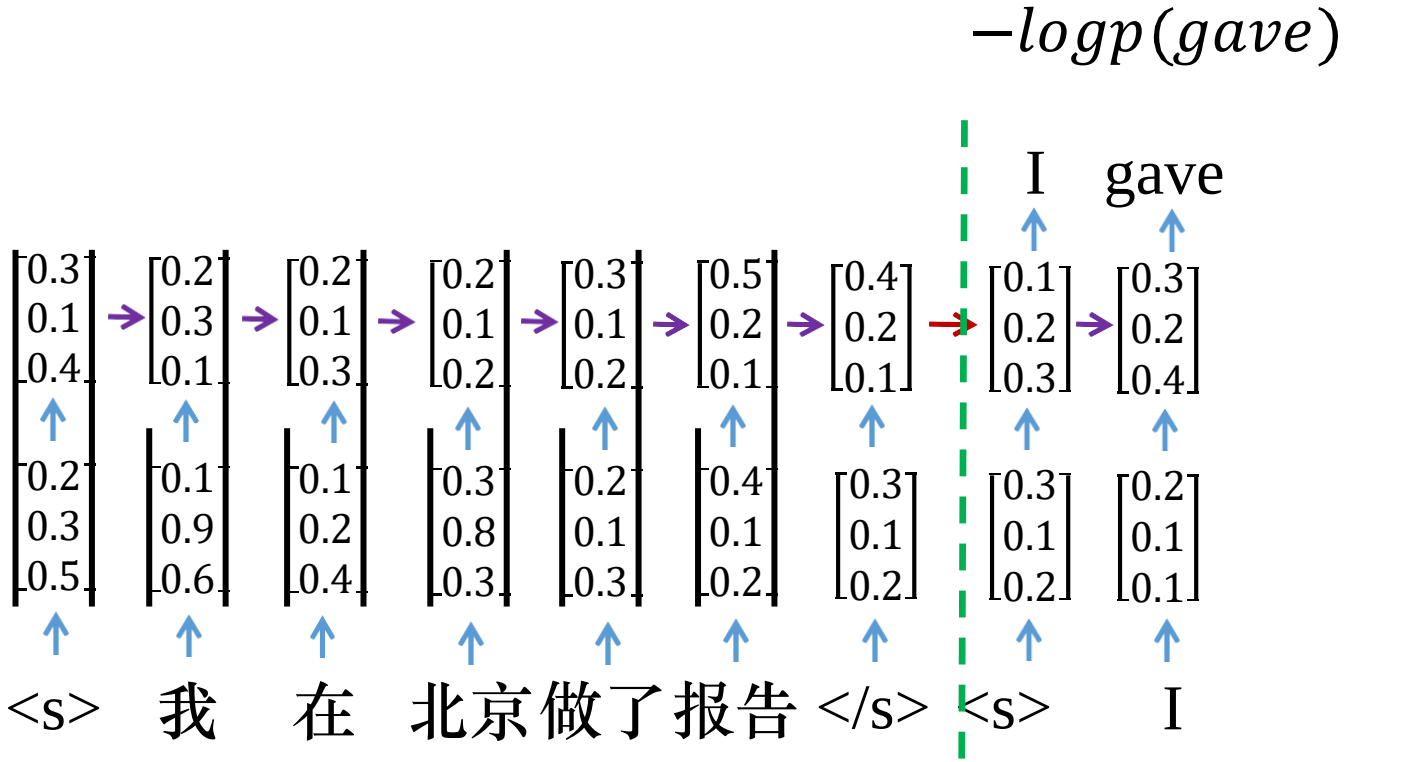


神经机器翻译



最大化 $P(target|source)$

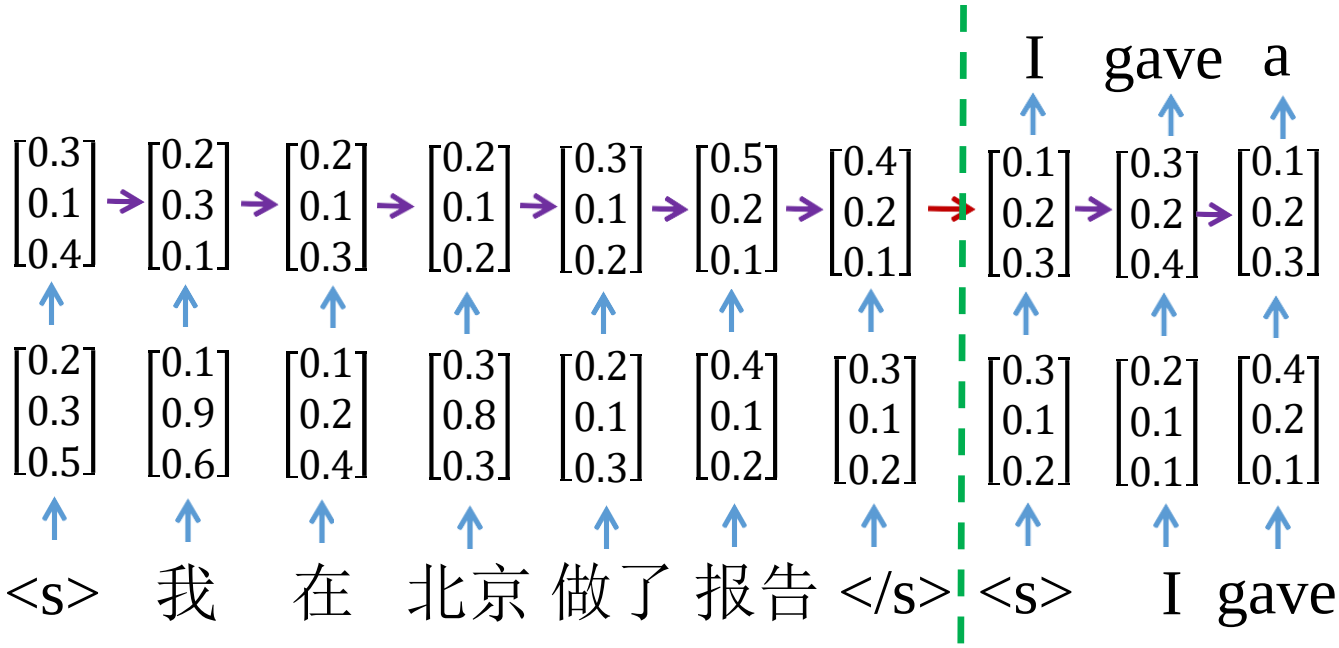
神经机器翻译



最大化 $P(target|source)$

神经机器翻译

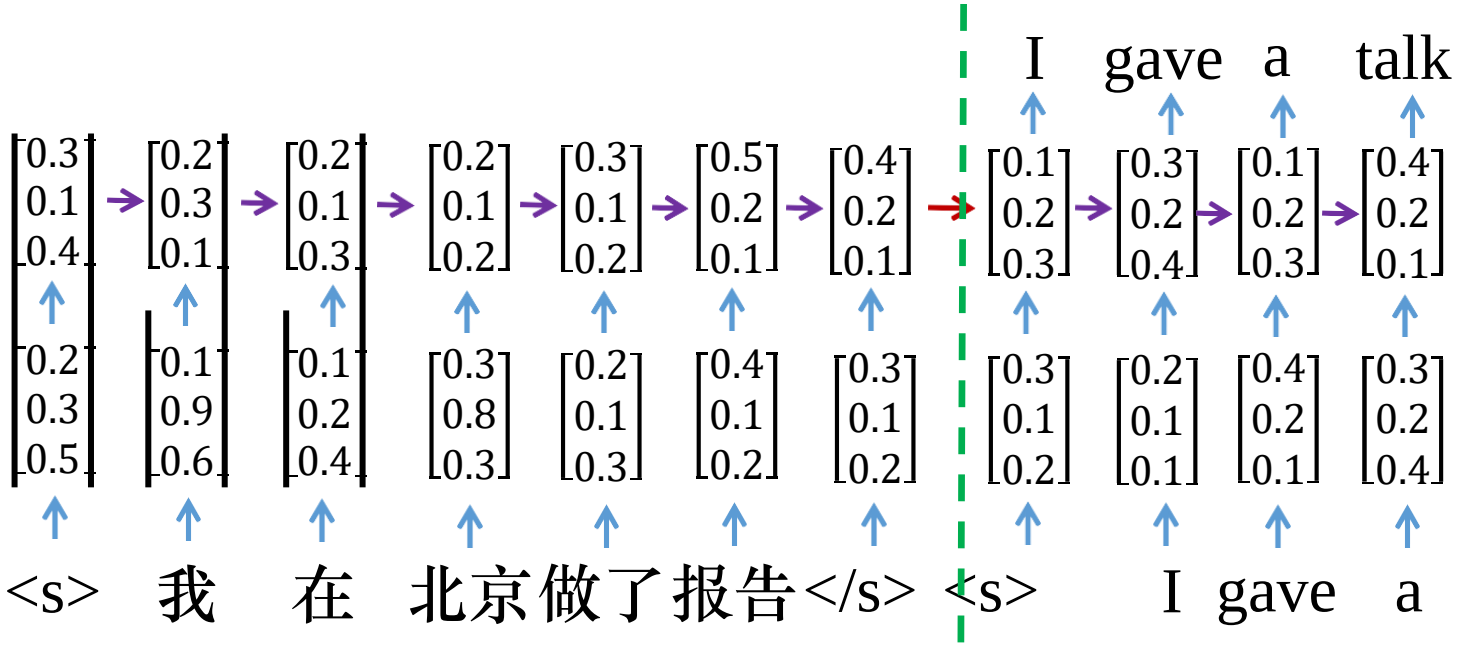
$-logp(a)$



最大化 $P(target|source)$

神经机器翻译

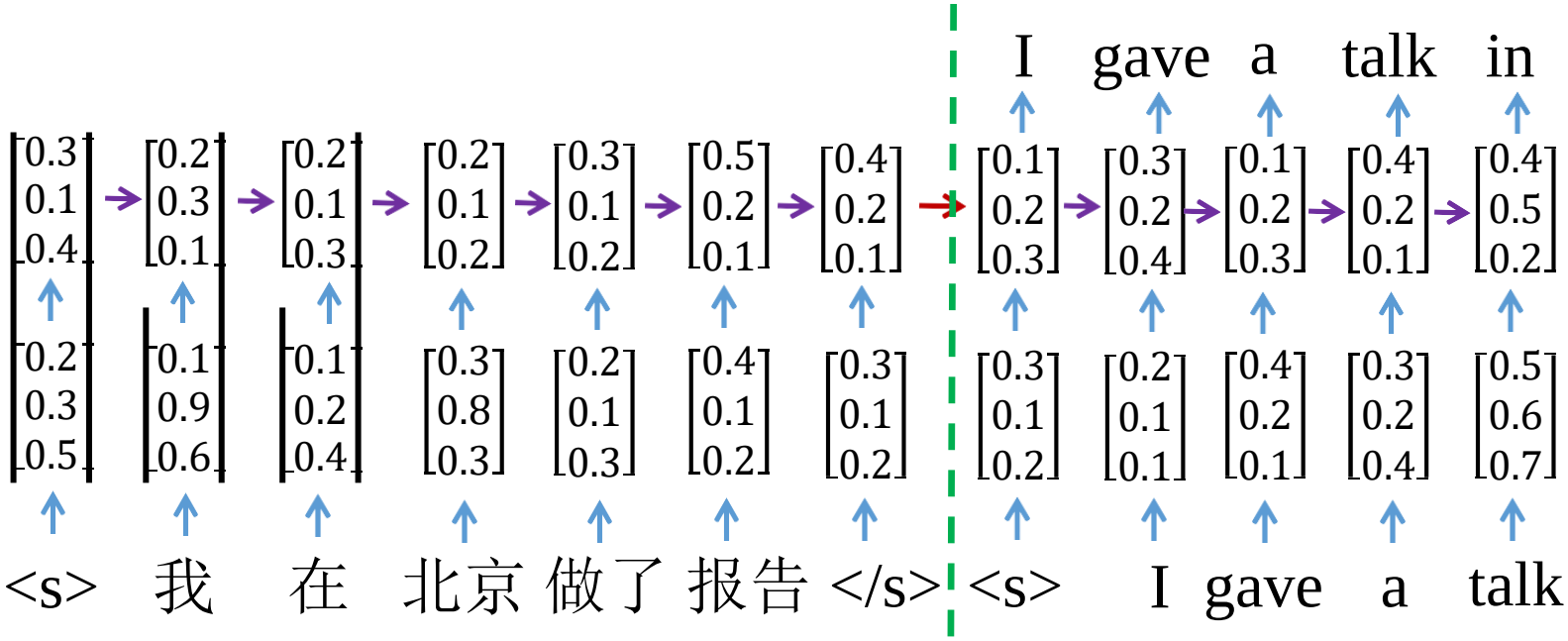
$-logp(talk)$



最大化 $P(target|source)$

神经机器翻译

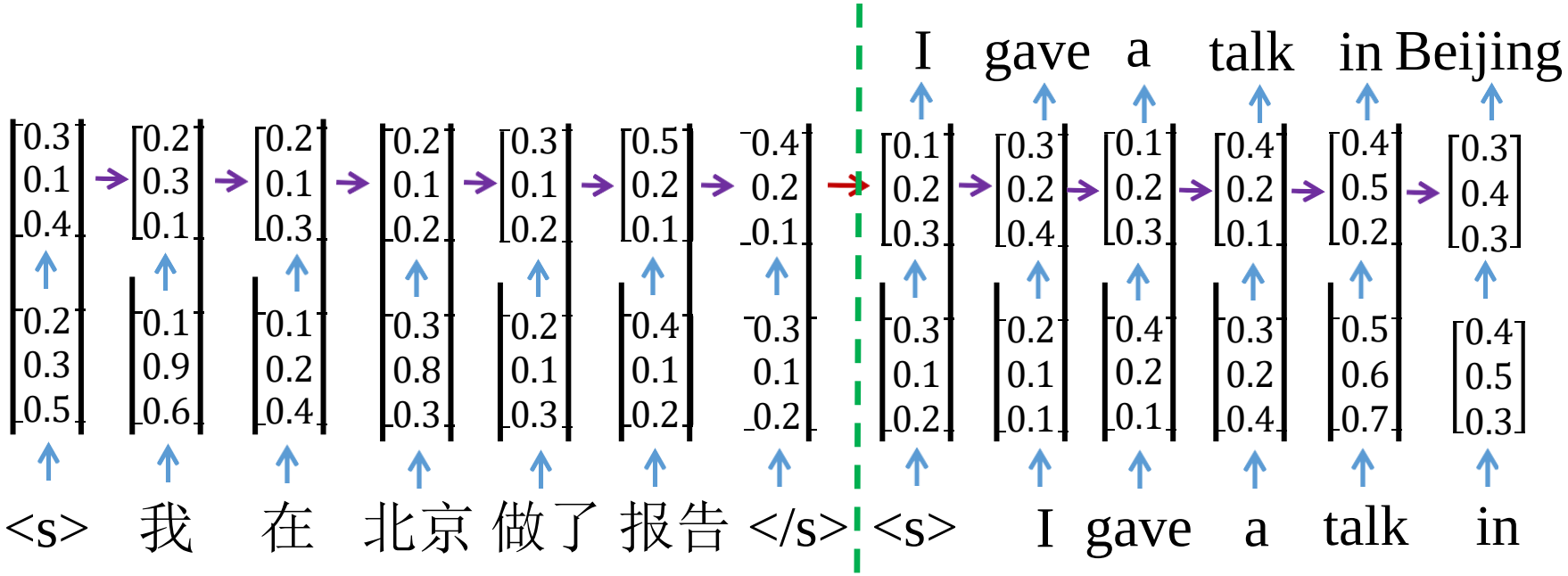
$-logp(in)$



最大化 $P(target|source)$

神经机器翻译

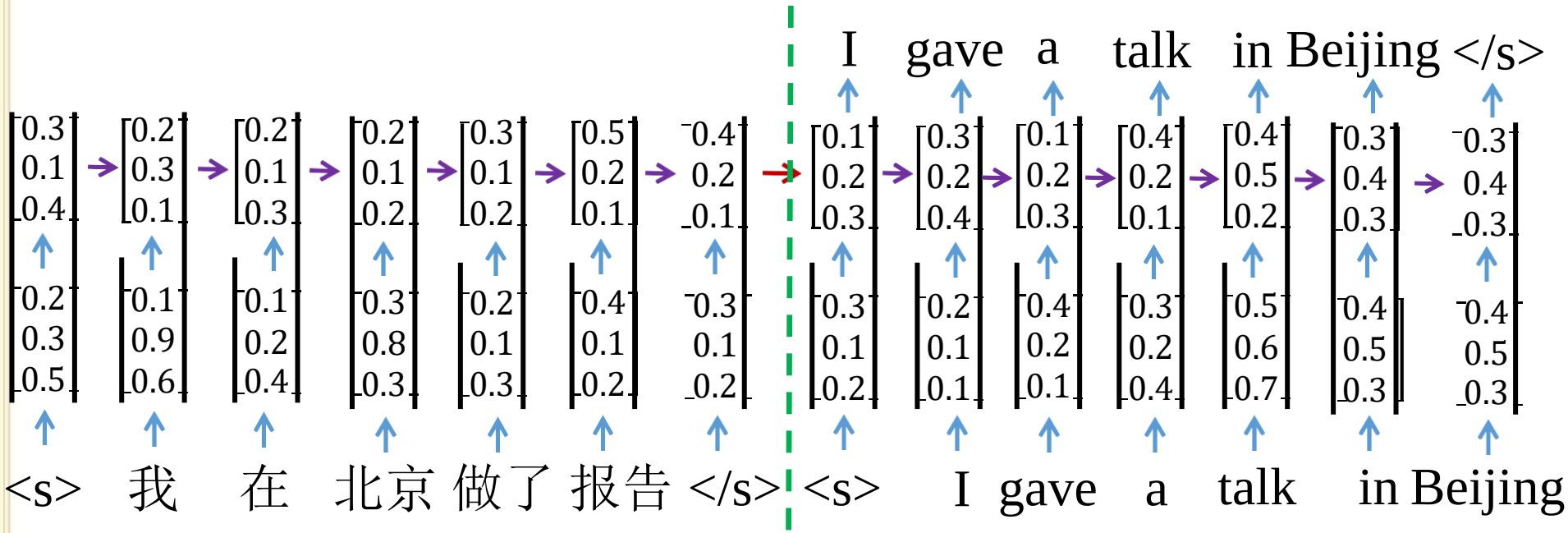
$-logp(Beijing)$



最大化 $P(target|source)$

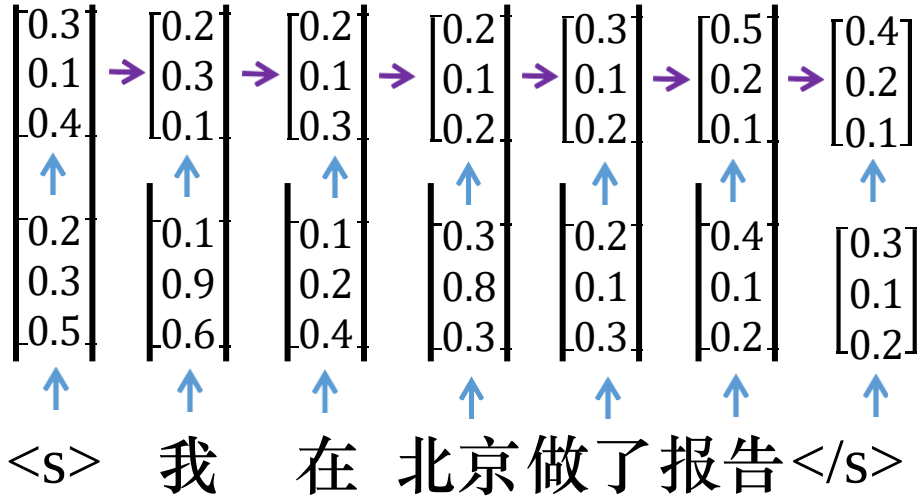
神经机器翻译

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



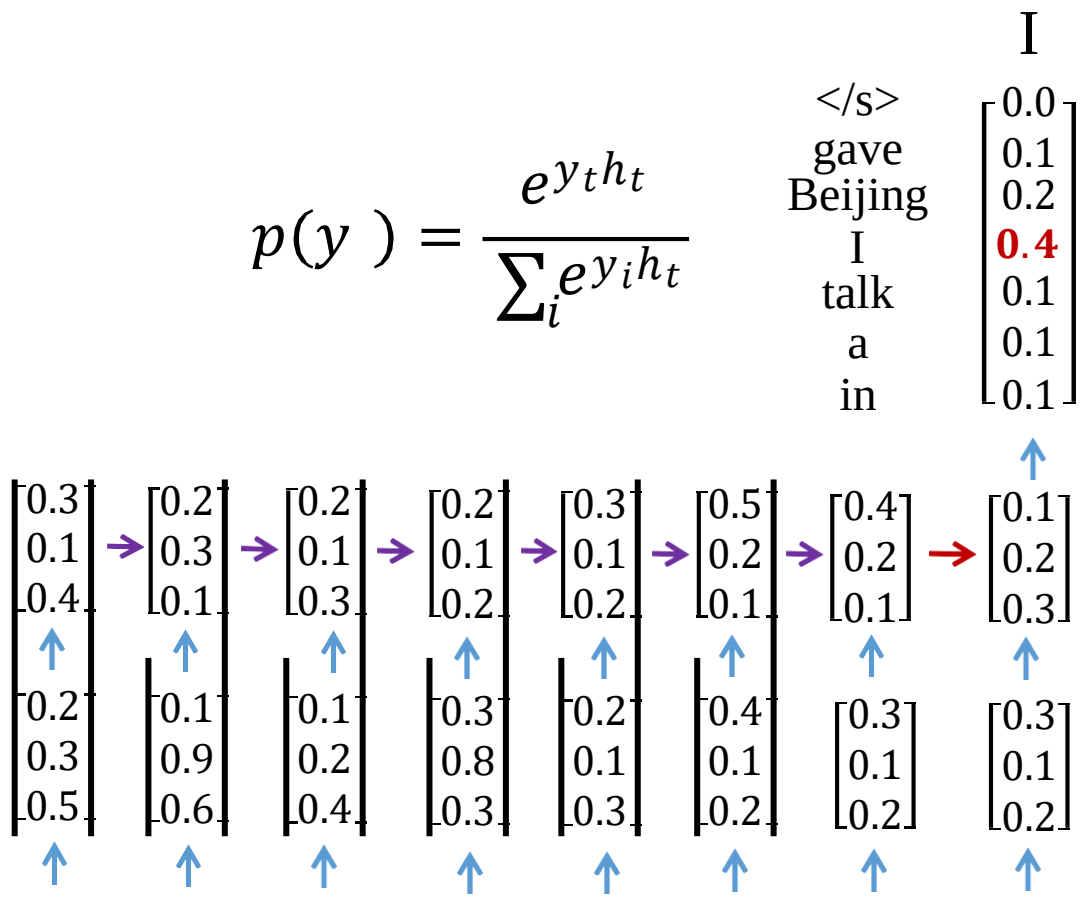
最大化 $P(target|source)$

神经机器翻译-测试

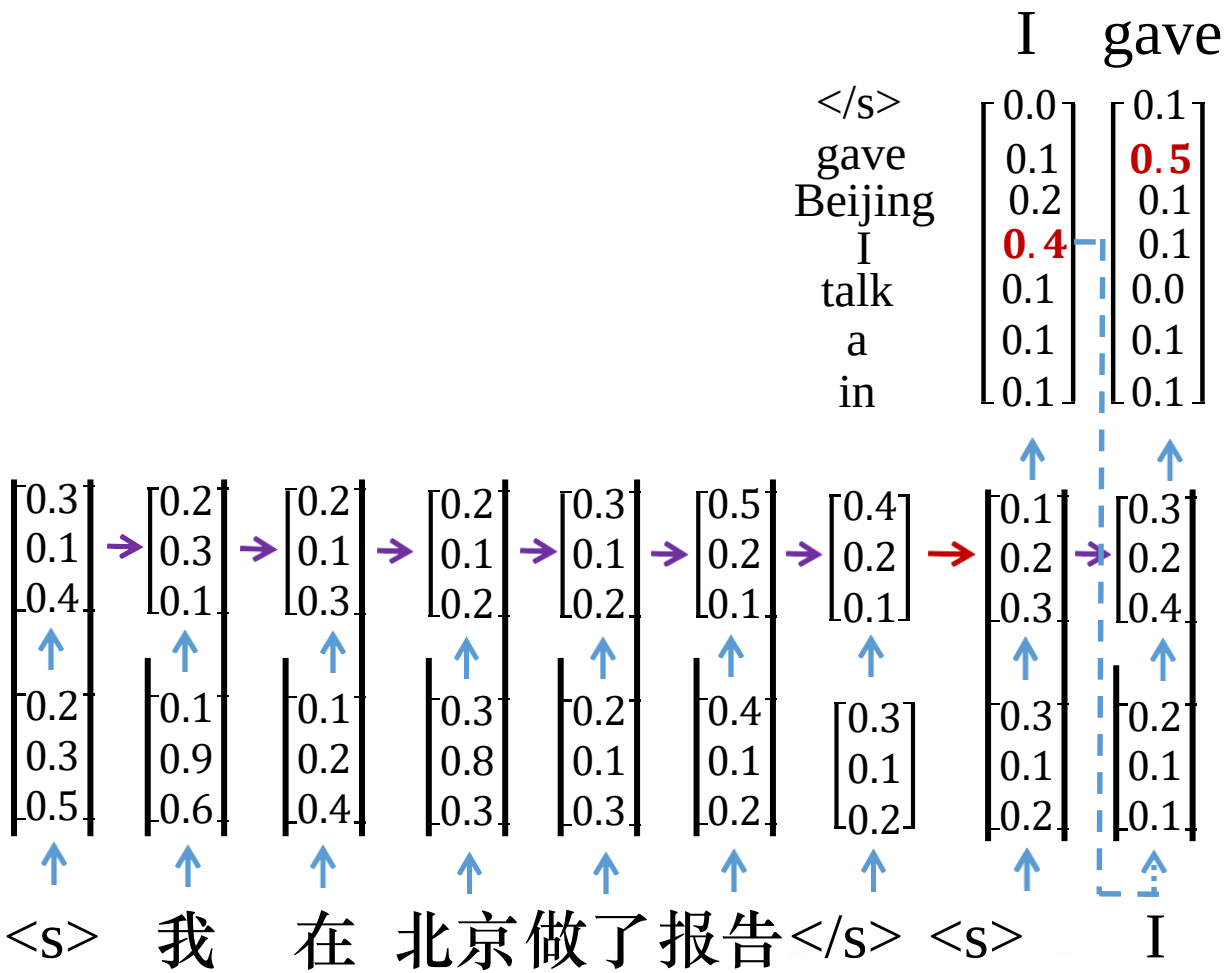

0.3 0.1 0.4
0.2 0.3 0.5
0.2 0.3 0.1
0.1 0.9 0.6
0.2 0.1 0.3
0.1 0.2 0.4
0.2 0.1 0.2
0.3 0.8 0.3
0.3 0.1 0.2
0.2 0.1 0.3
0.5 0.2 0.1
0.4 0.1 0.2
0.4 0.2 0.1
0.3 0.1 0.2
<s> 我 在 北京做了报告</s>

神经机器翻译-测试

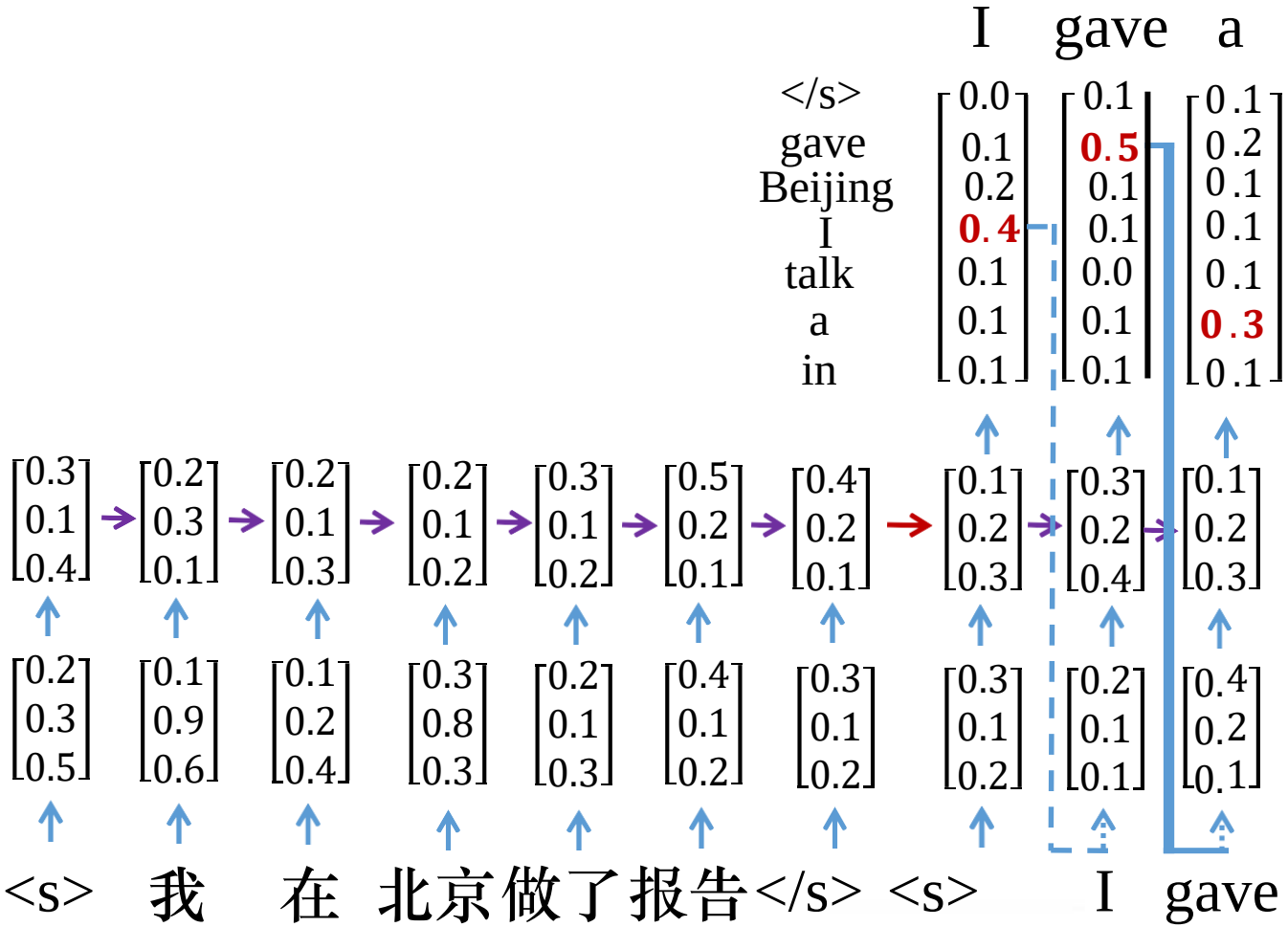
$$p(y\ ) = \frac{e^{y_t h_t}}{\sum_i e^{y_i h_t}}$$



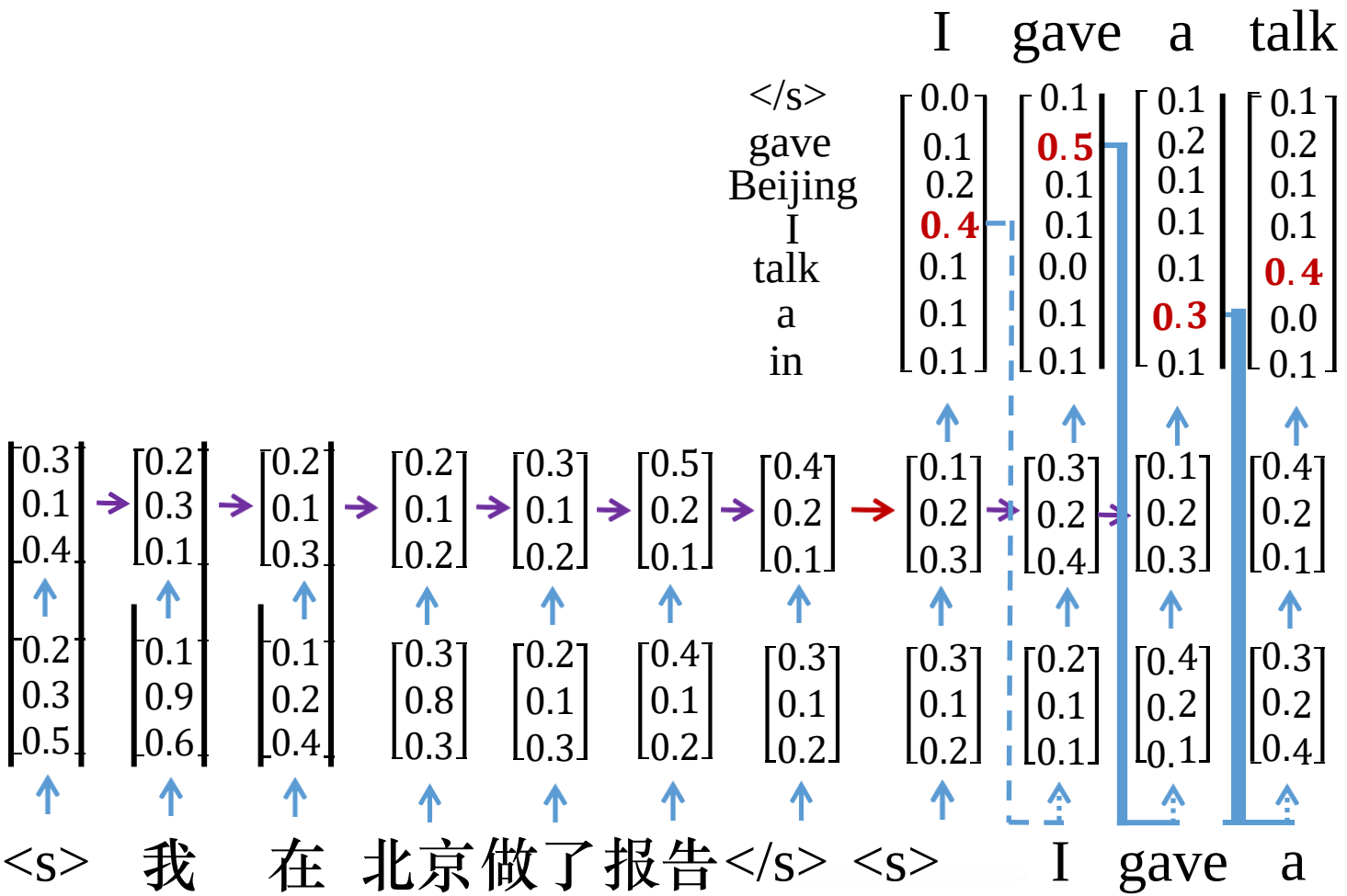
神经机器翻译-测试


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

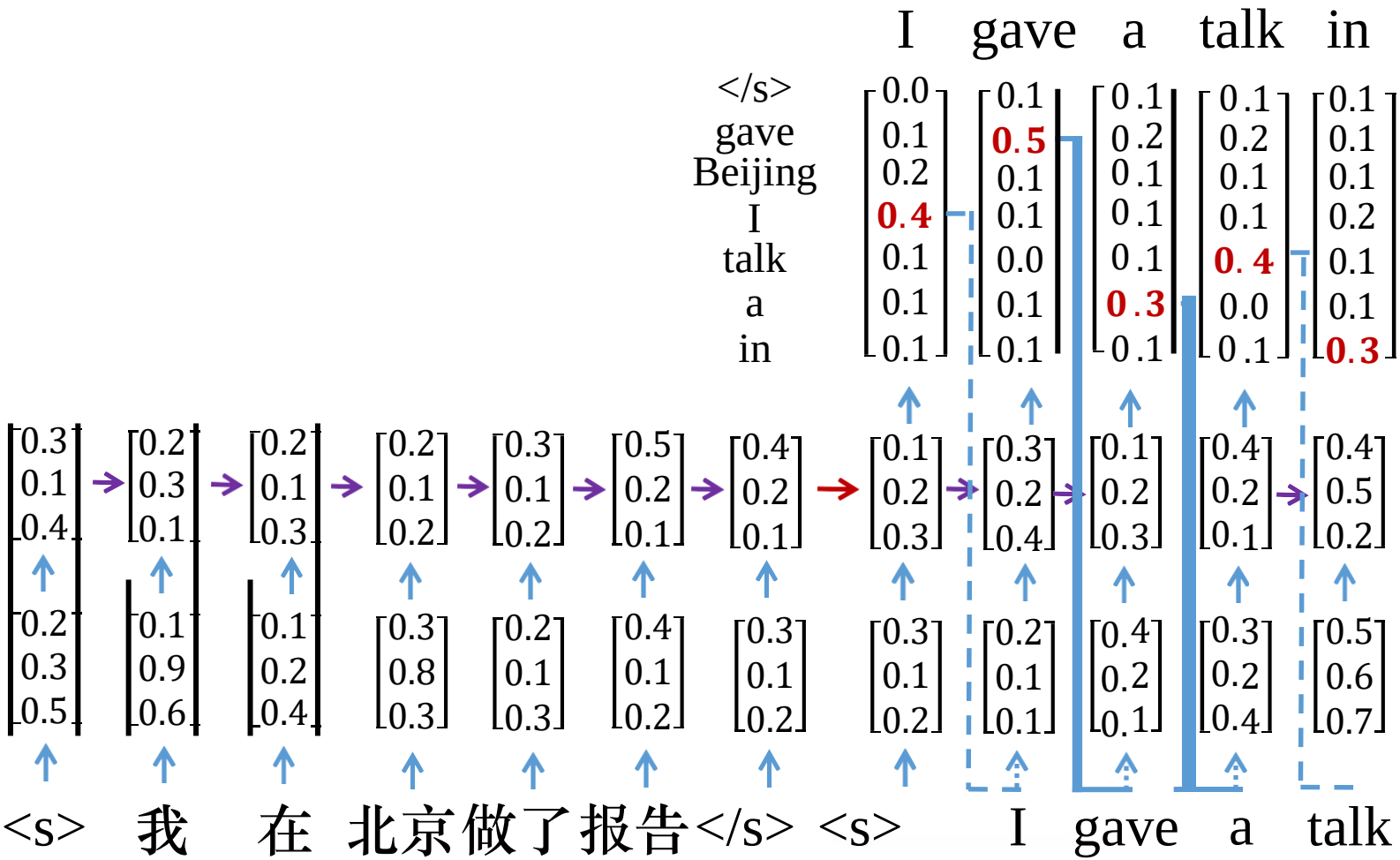
神经机器翻译-测试



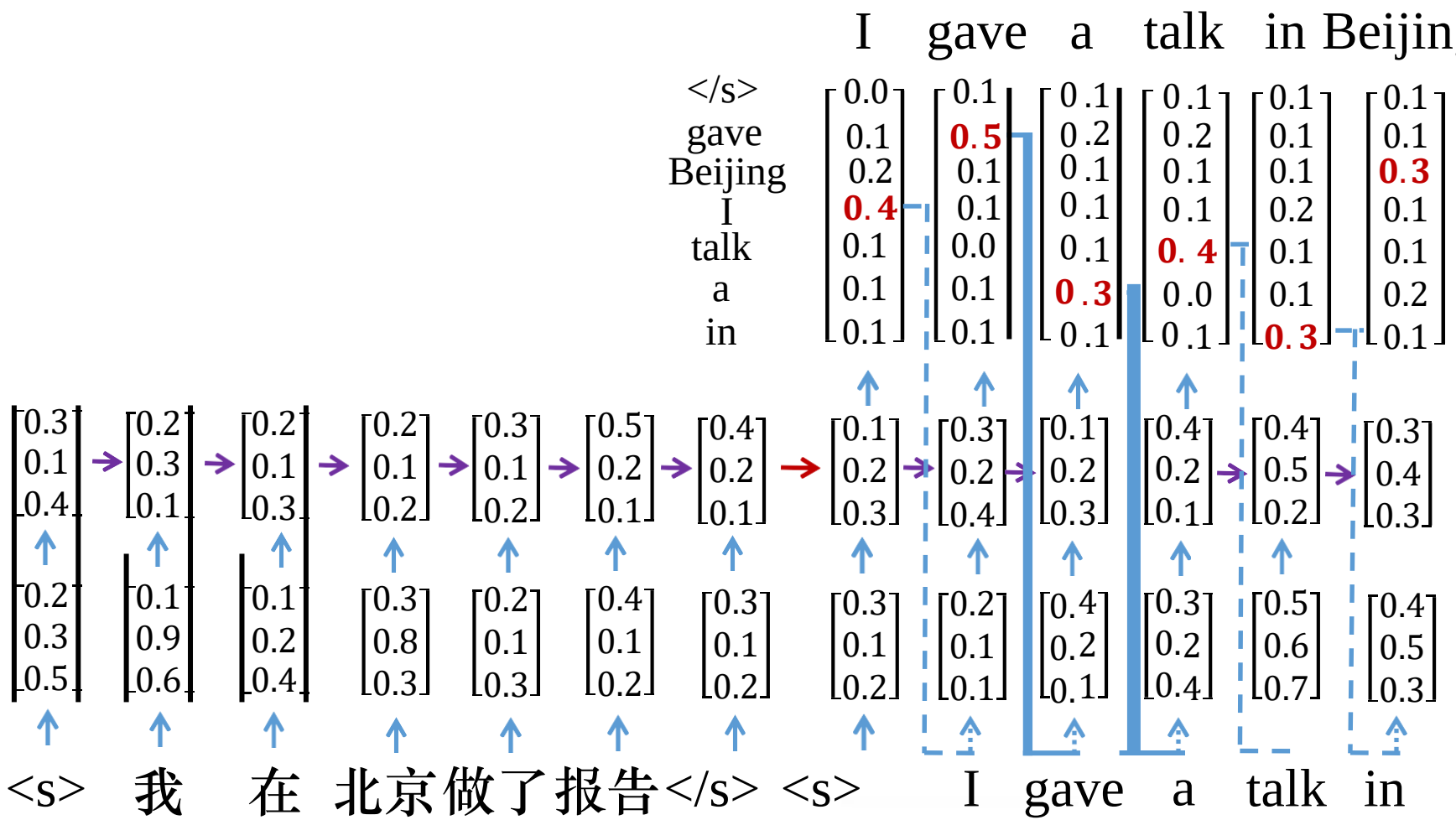
神经机器翻译-测试



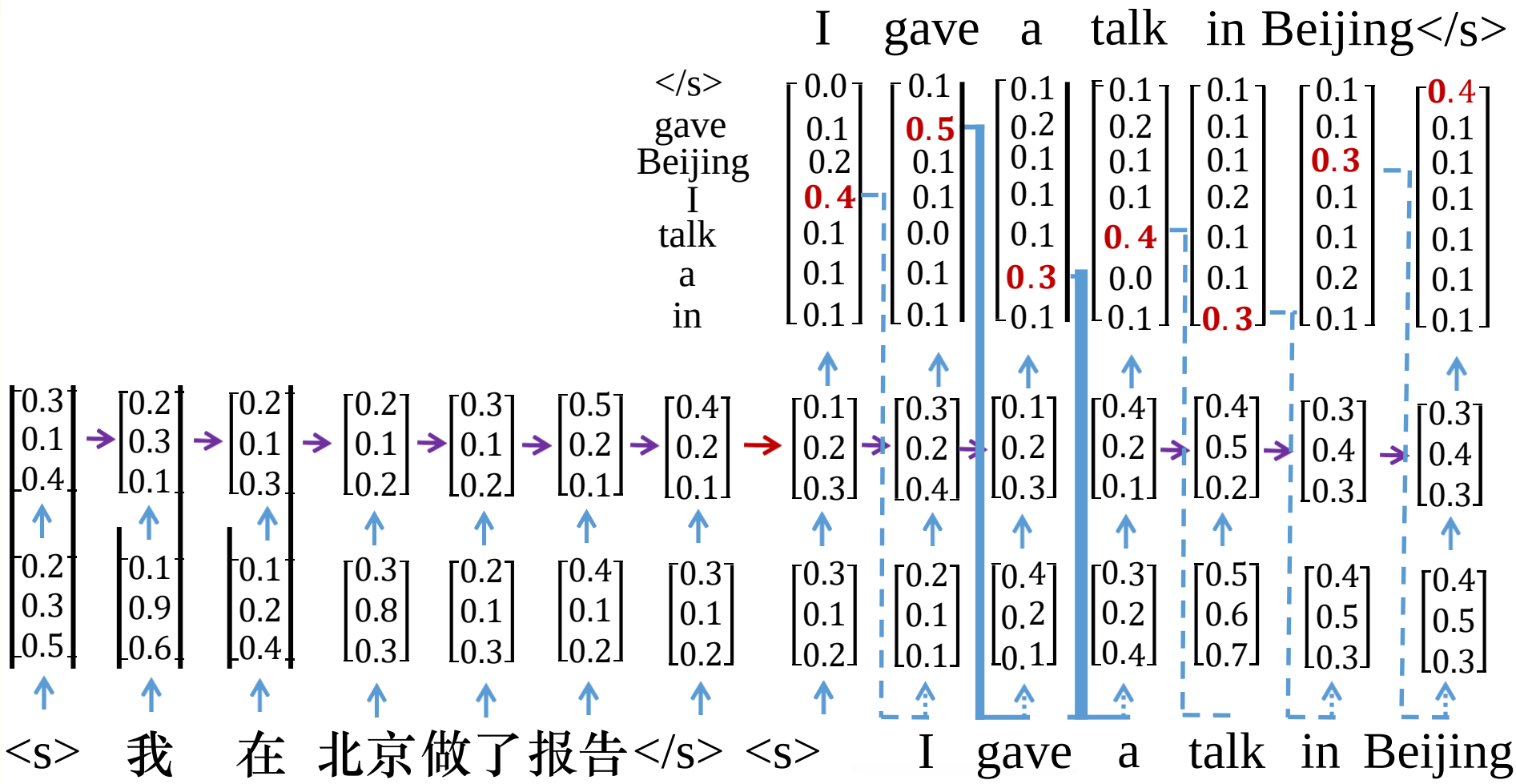
神经机器翻译-测试


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

14.2: 注意力机制

Need for Attention

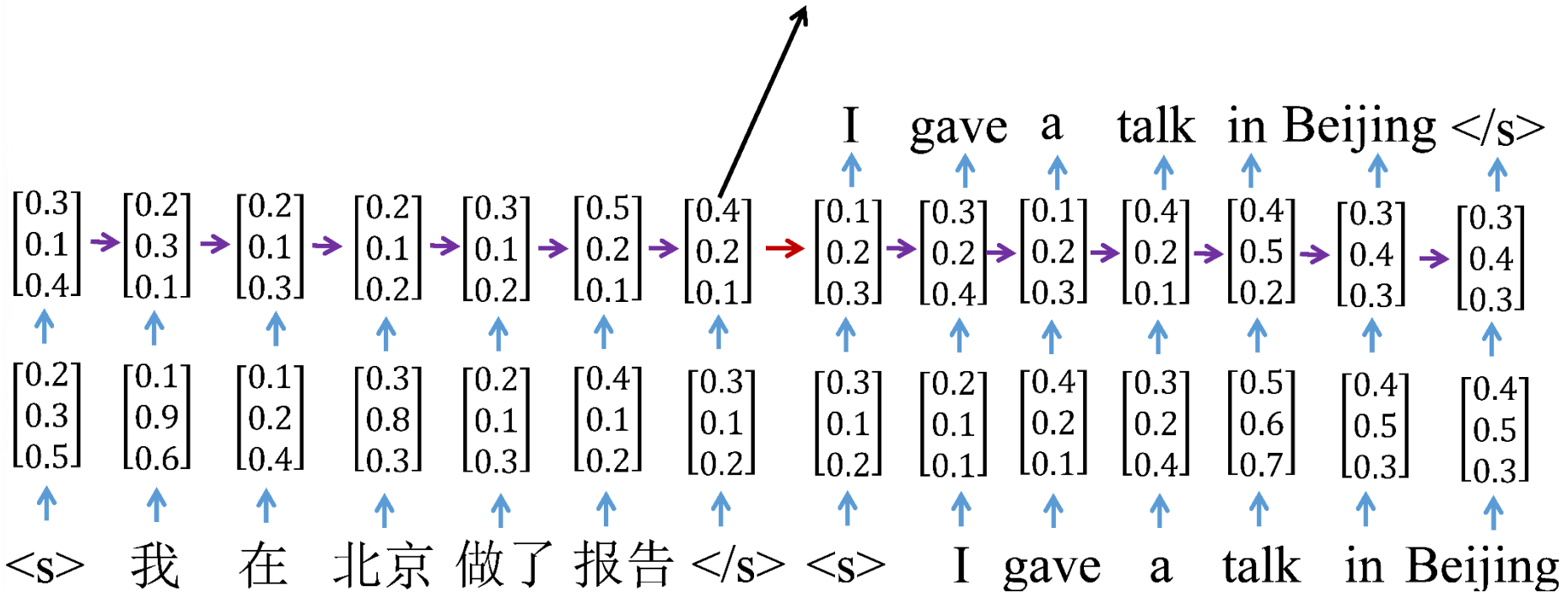
Attention removes bottleneck of Encoder-Decoder model

- Pay selective attention to relevant parts of input
- Utilize all intermediate states while decoding

  (Don’t encode the whole input into one vector losing information)
- Do Alignment while translating

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



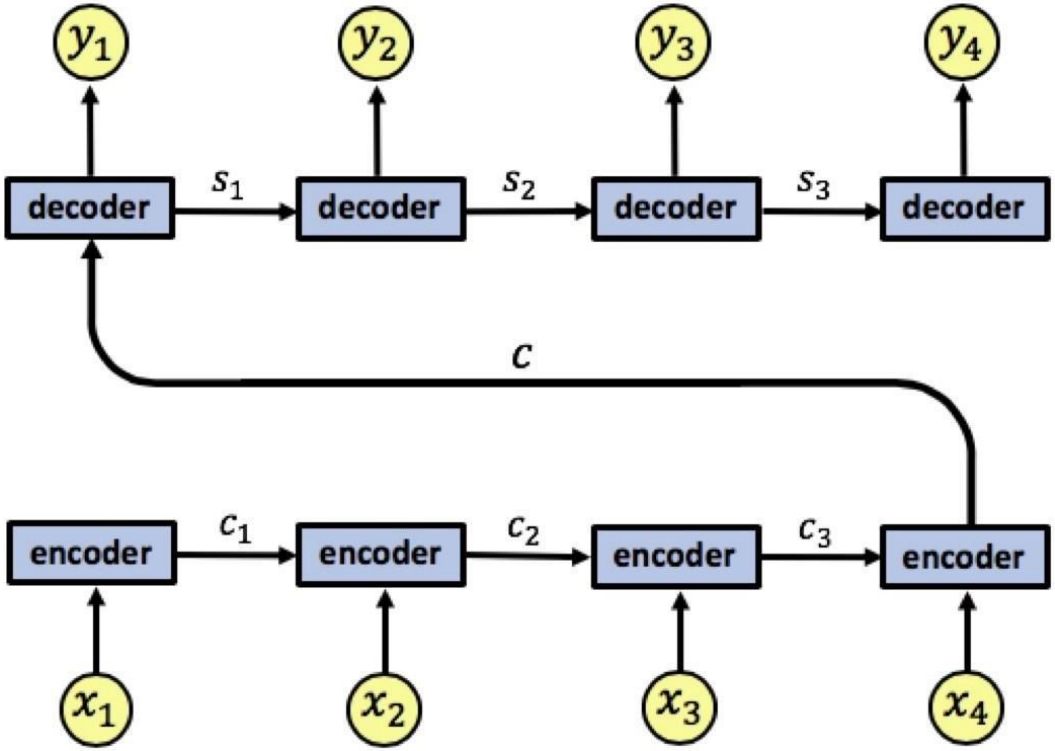
Pay selective attention

First introduced in Image Recognition

A woman is throwing a frisbee in a park.

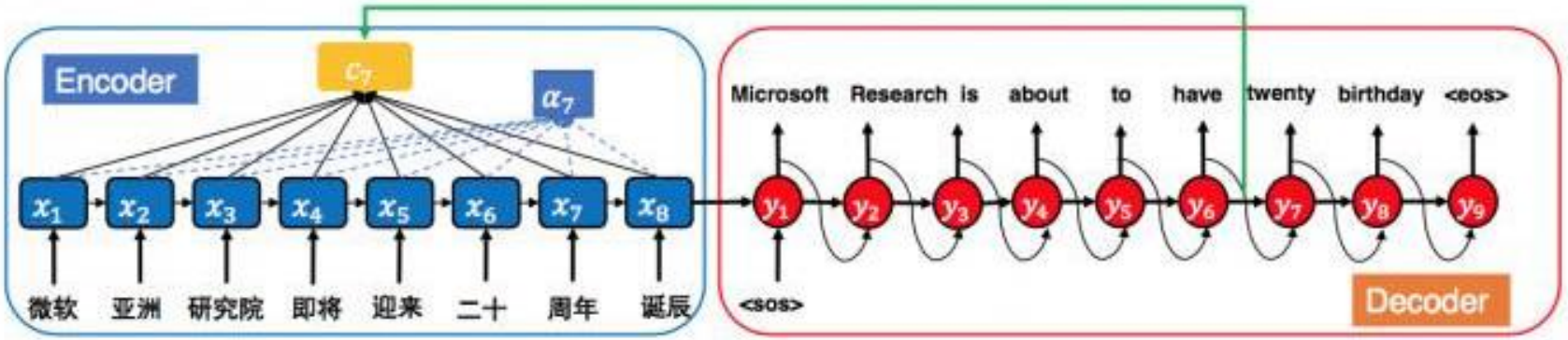
Without attention

Decode using only final state



统计机器翻译→神经机器翻译

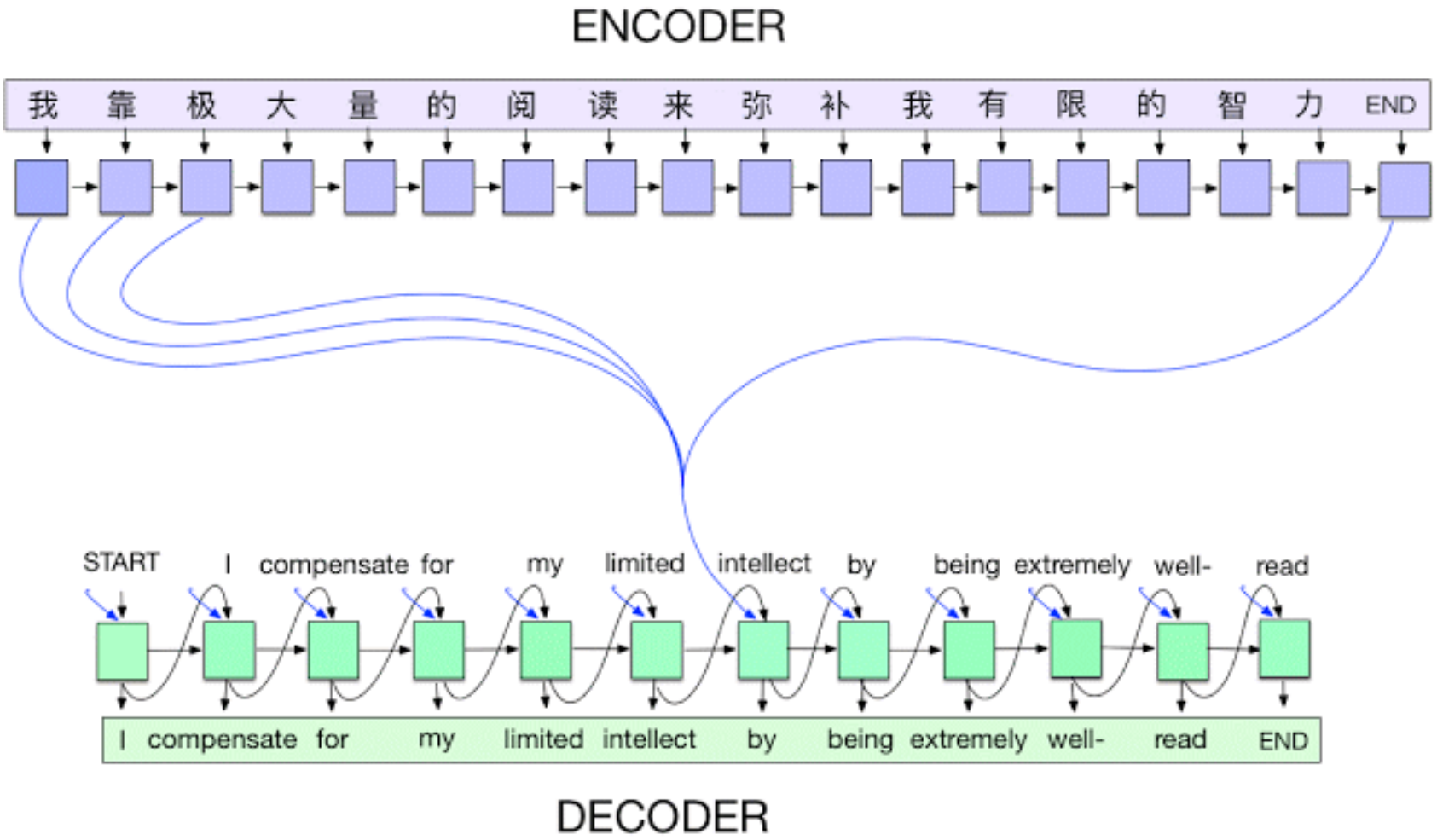
Attention 注意力机制



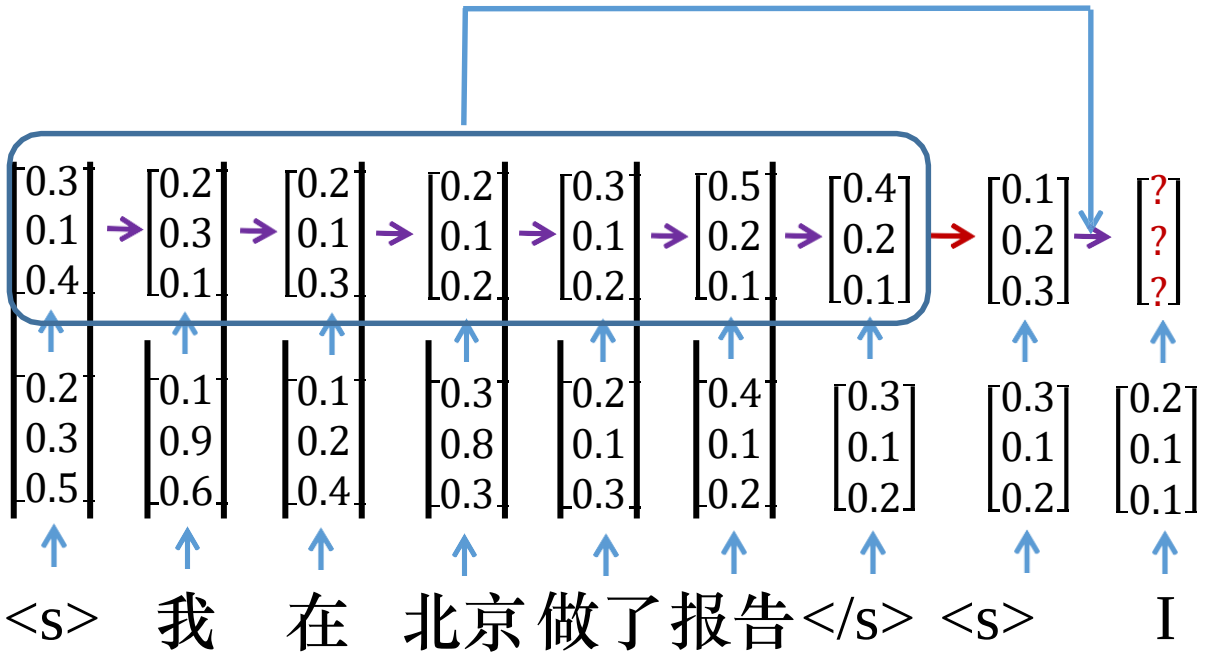
- $c_j = \sum_{i=1}^{T_x} \alpha_{ji} h_i$
- $\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^{T_x} \exp(e_{jk})}$
- $e_{ji} = A(s_{j-1}, h_i)$

$s_j = f(y_{j-1}, s_{j-1}; c_j)$: LSTM/GRU

神经机器翻译-注意机制

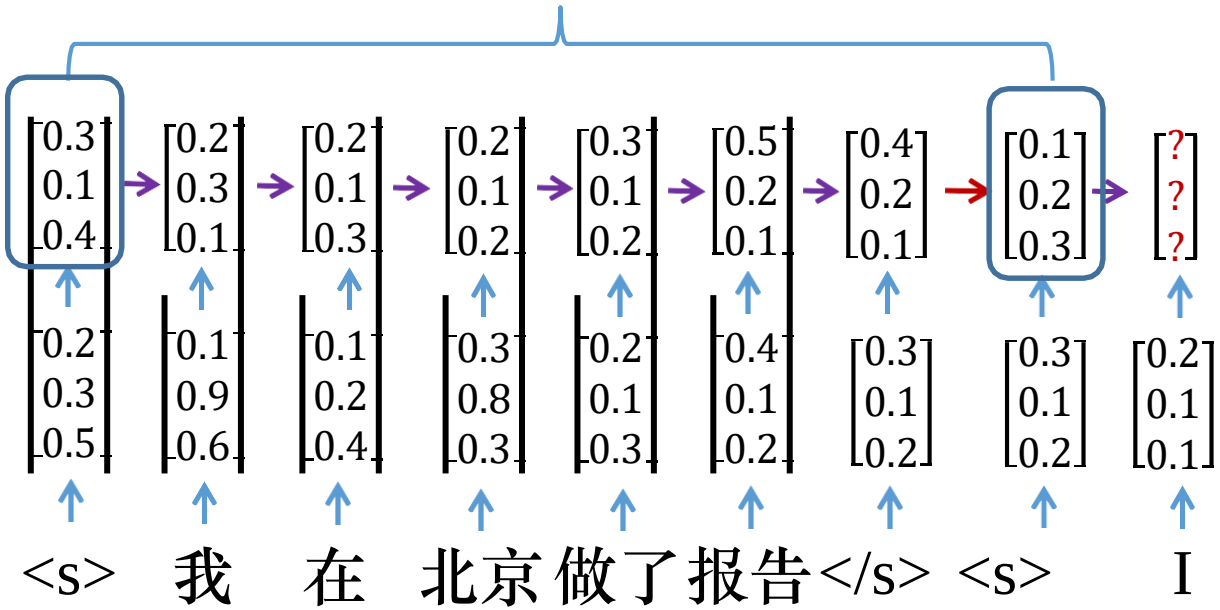

ENCODER
我 靠 极 大 量 的 阅 读 来 弥 补 我 有 限 的 智 力 END
START I compensate for my limited intellect by being extremely well- read
I compensate for my limited intellect by being extremely well- read END
DECODER

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

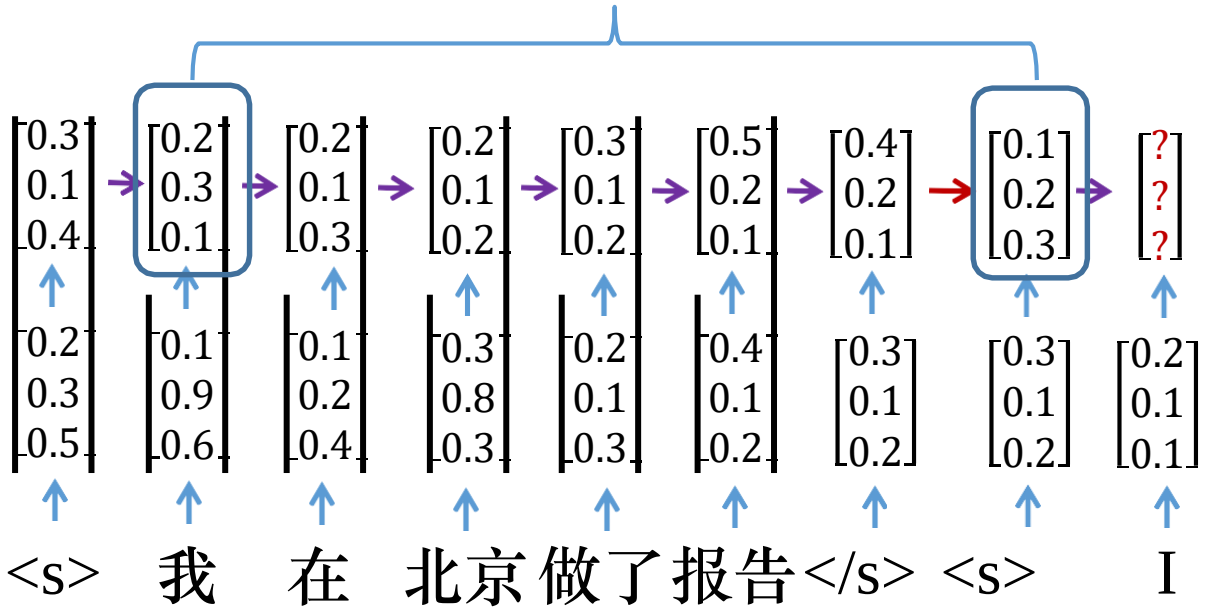
神经机器翻译-注意机制

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

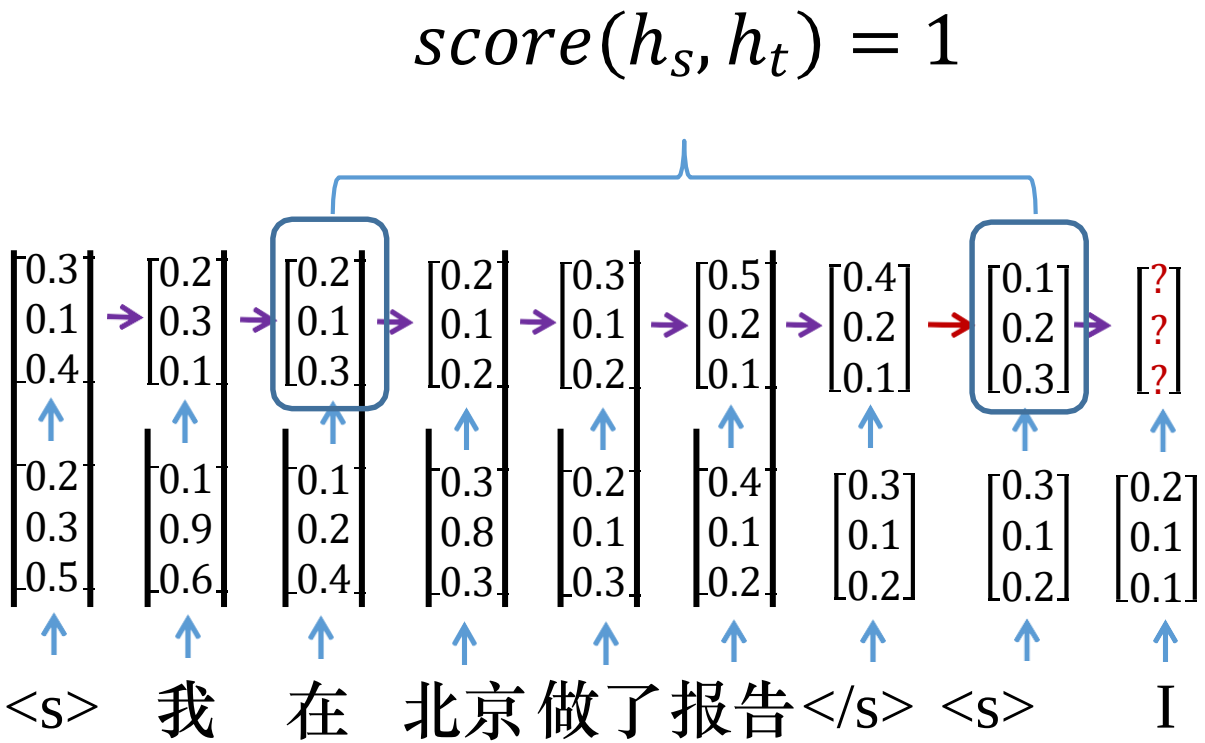


神经机器翻译-注意机制

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

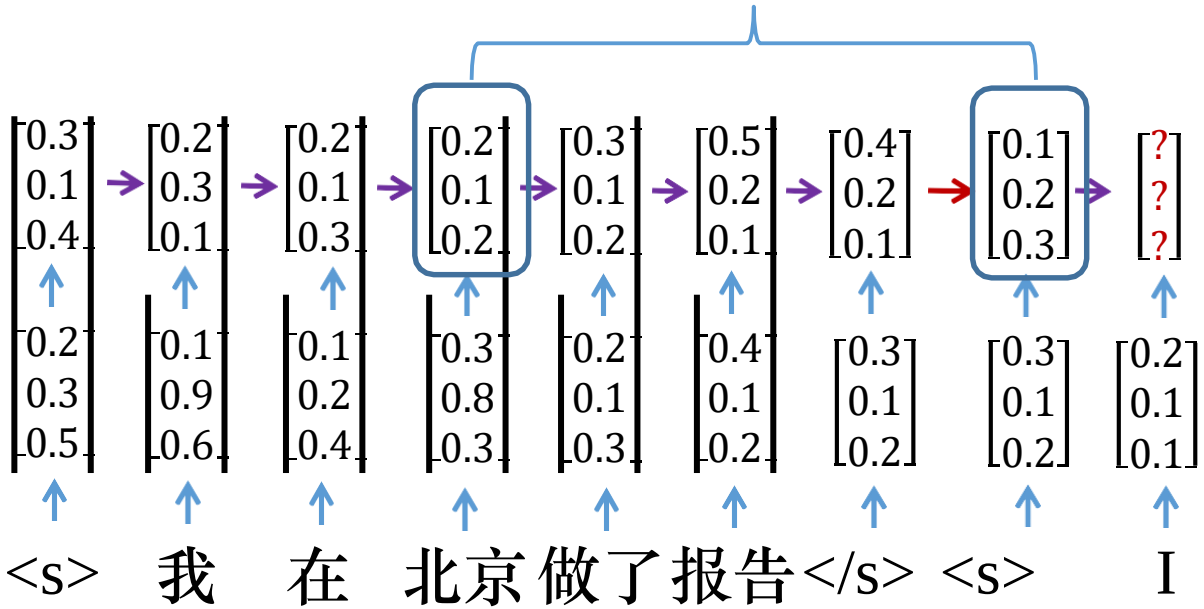


神经机器翻译-注意机制


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

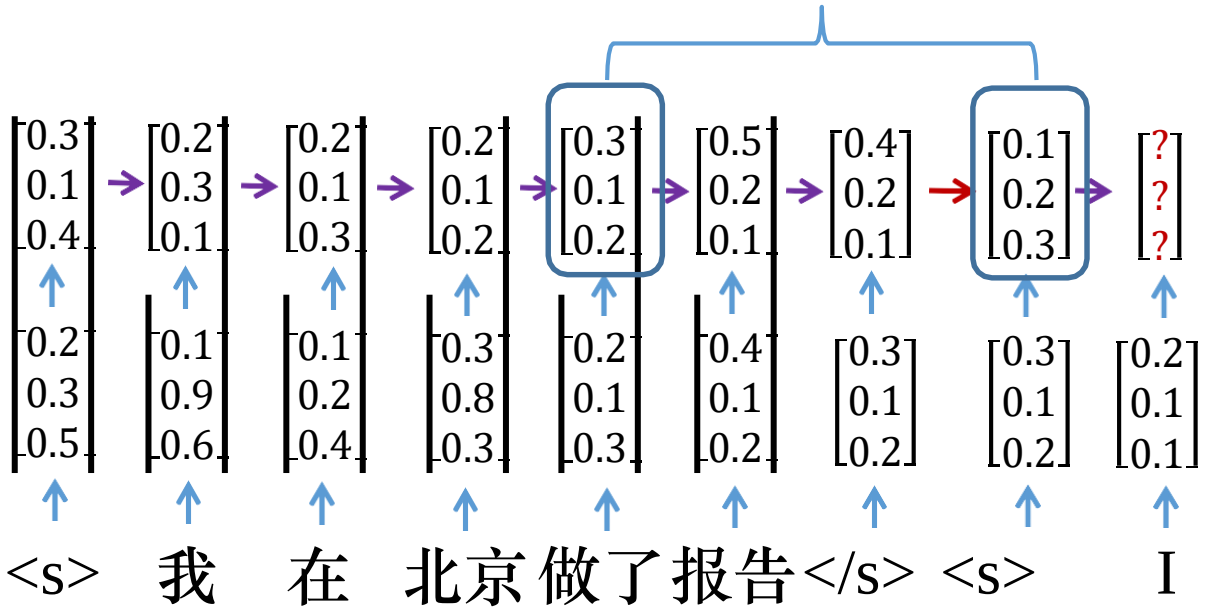
神经机器翻译-注意机制

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



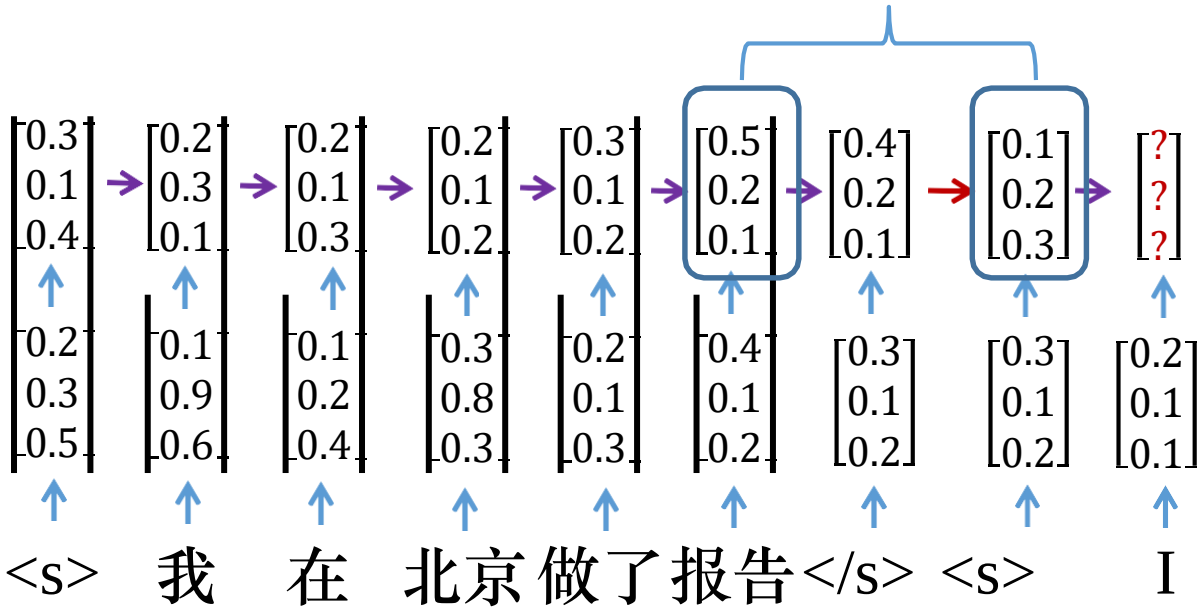
神经机器翻译-注意机制

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



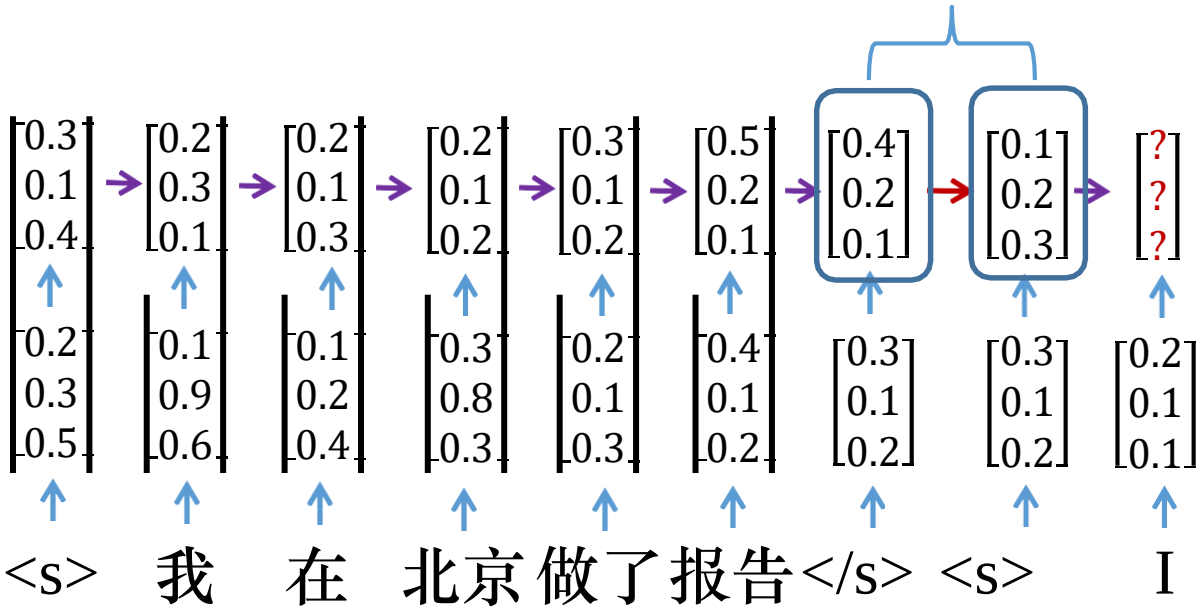
神经机器翻译-注意机制

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

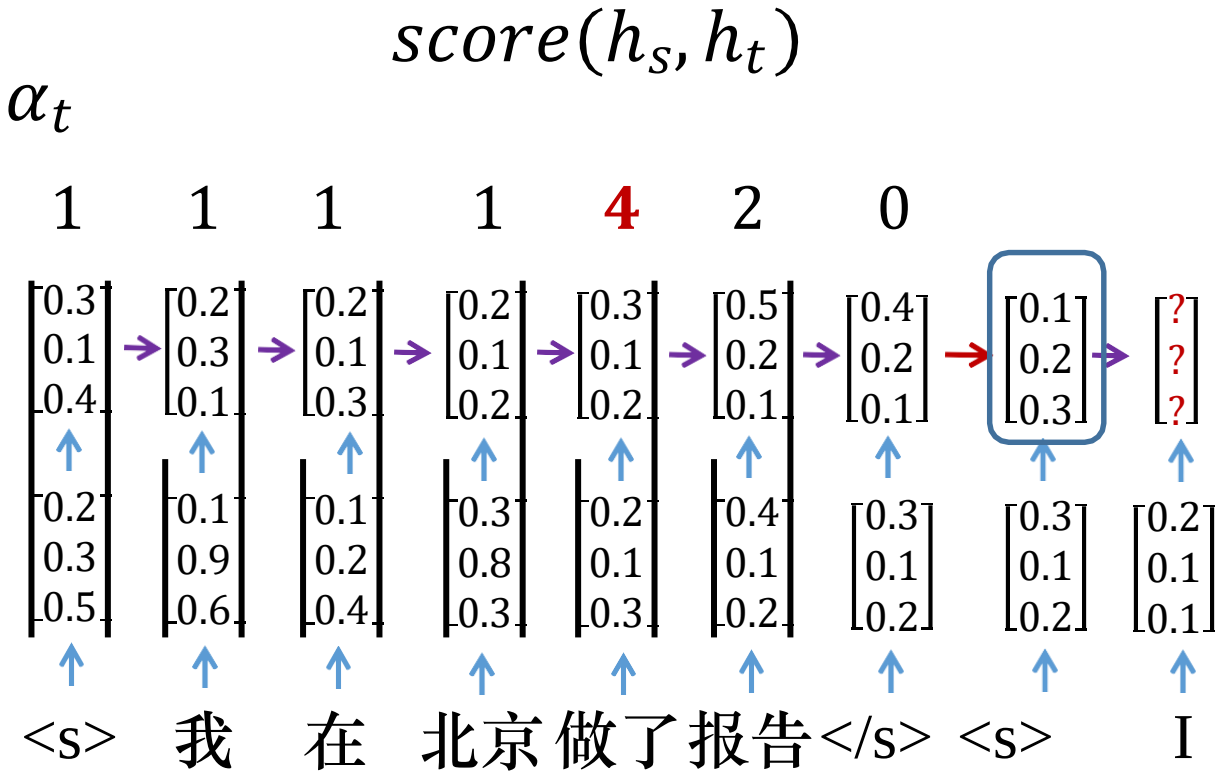


神经机器翻译-注意机制

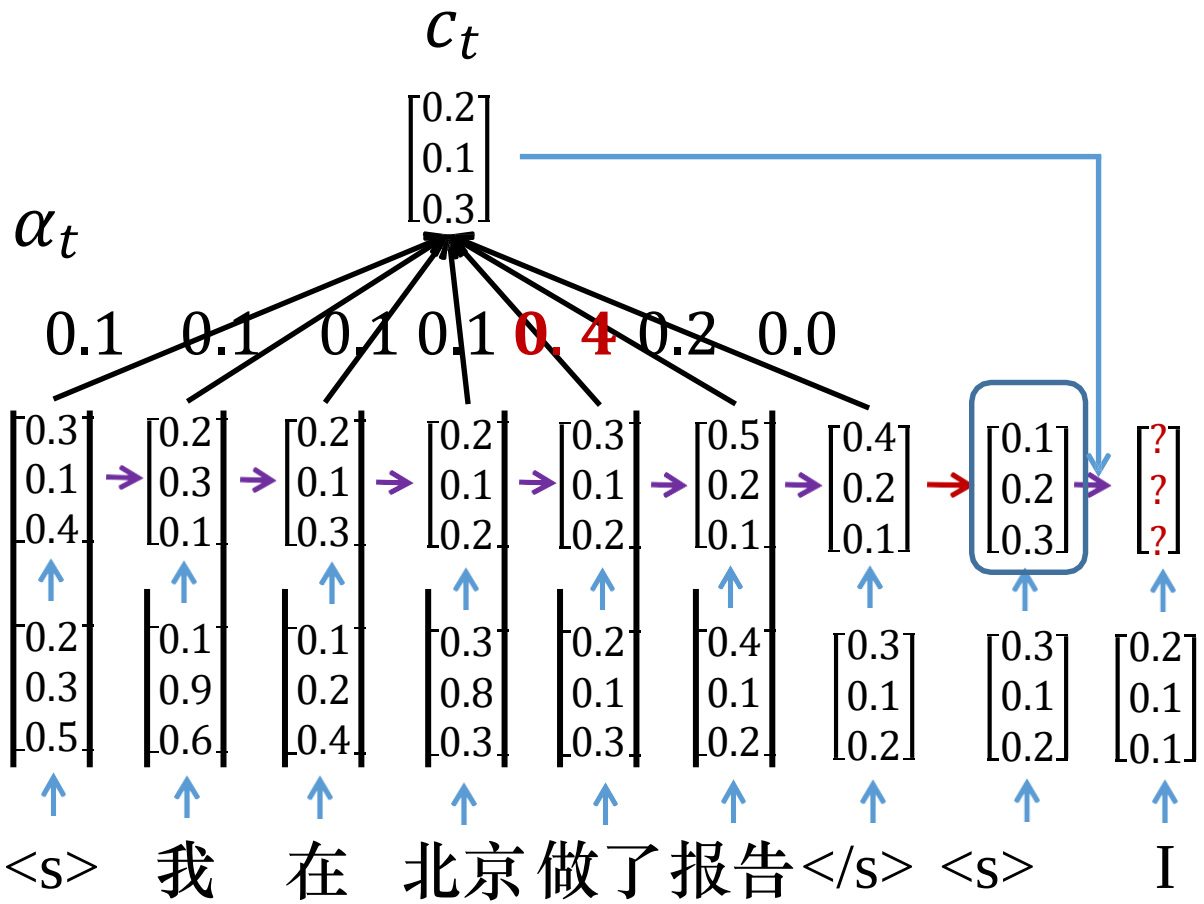
$$score(h_s, h_t) = 0$$



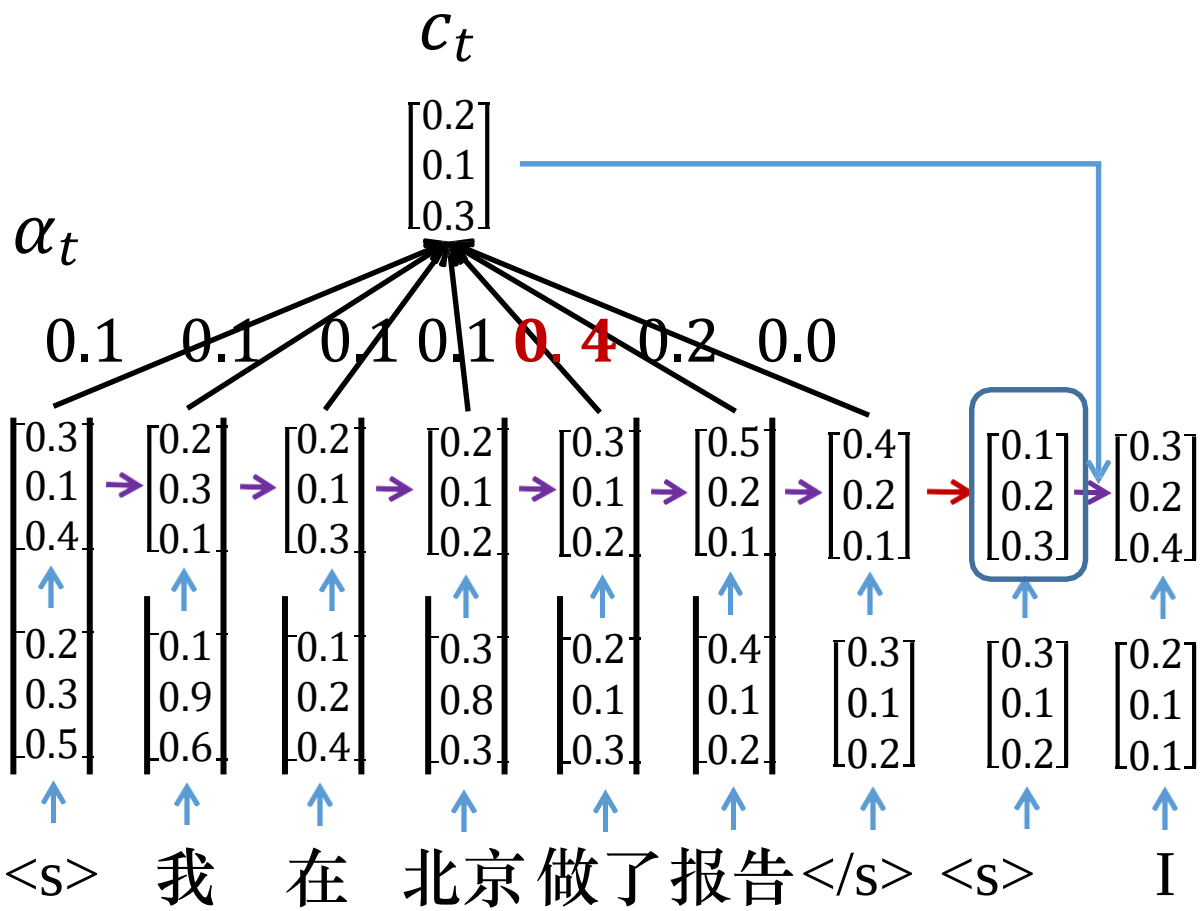
神经机器翻译-注意机制


$score(h_s, h_t)$
α_t
1 1 1 1 4 2 0
0.3 0.1 0.4
0.2 0.3 0.5
0.2 0.3 0.1
0.1 0.9 0.6
0.2 0.1 0.3
0.1 0.2 0.4
0.2 0.1 0.2
0.3 0.8 0.3
0.3 0.1 0.2
0.2 0.1 0.3
0.5 0.2 0.1
0.4 0.1 0.2
0.4 0.2 0.1
0.3 0.1 0.2
0.1 0.2 0.3
0.3 0.1 0.2
? ? ?
0.2 0.1 0.1
<s> 我 在 北京 做了 报告 </s> <s> I

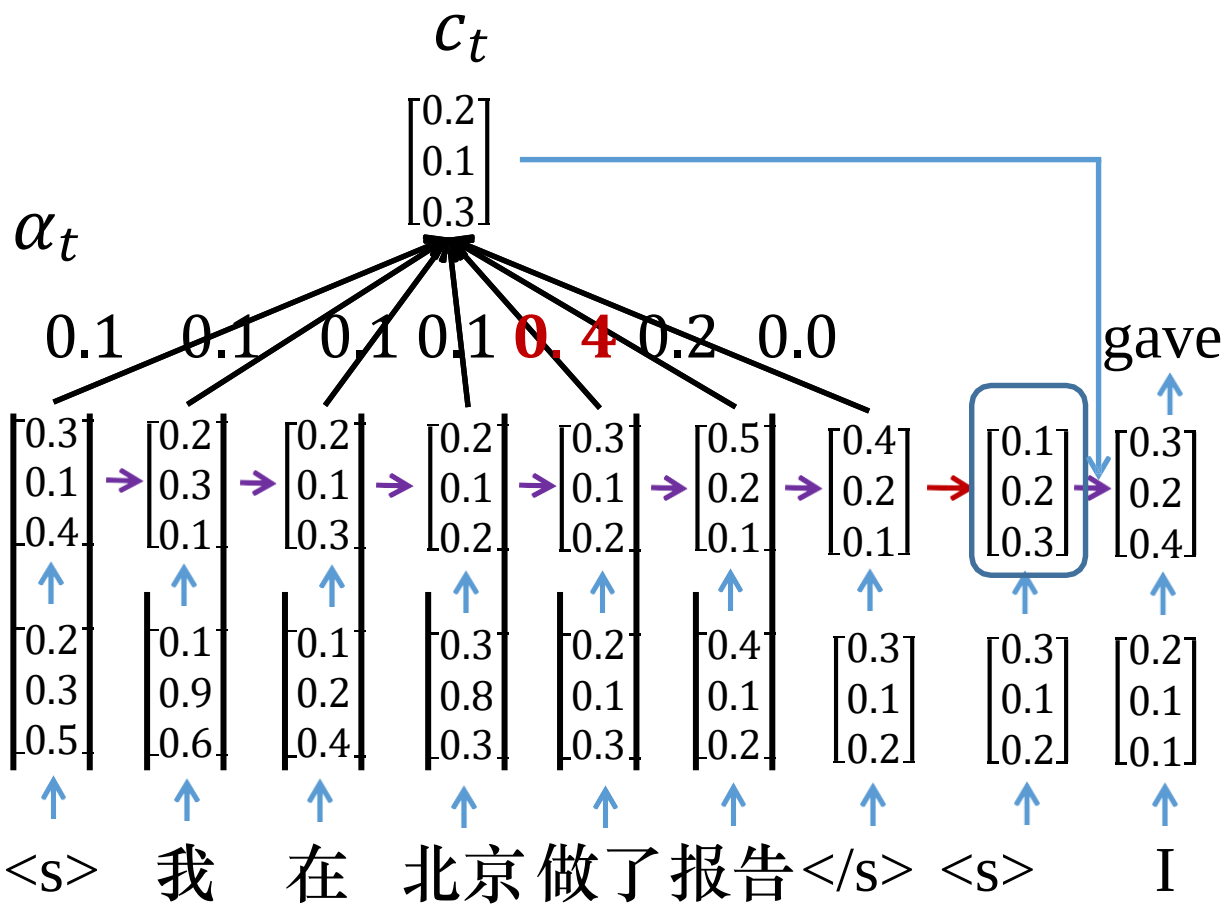
神经机器翻译-注意机制



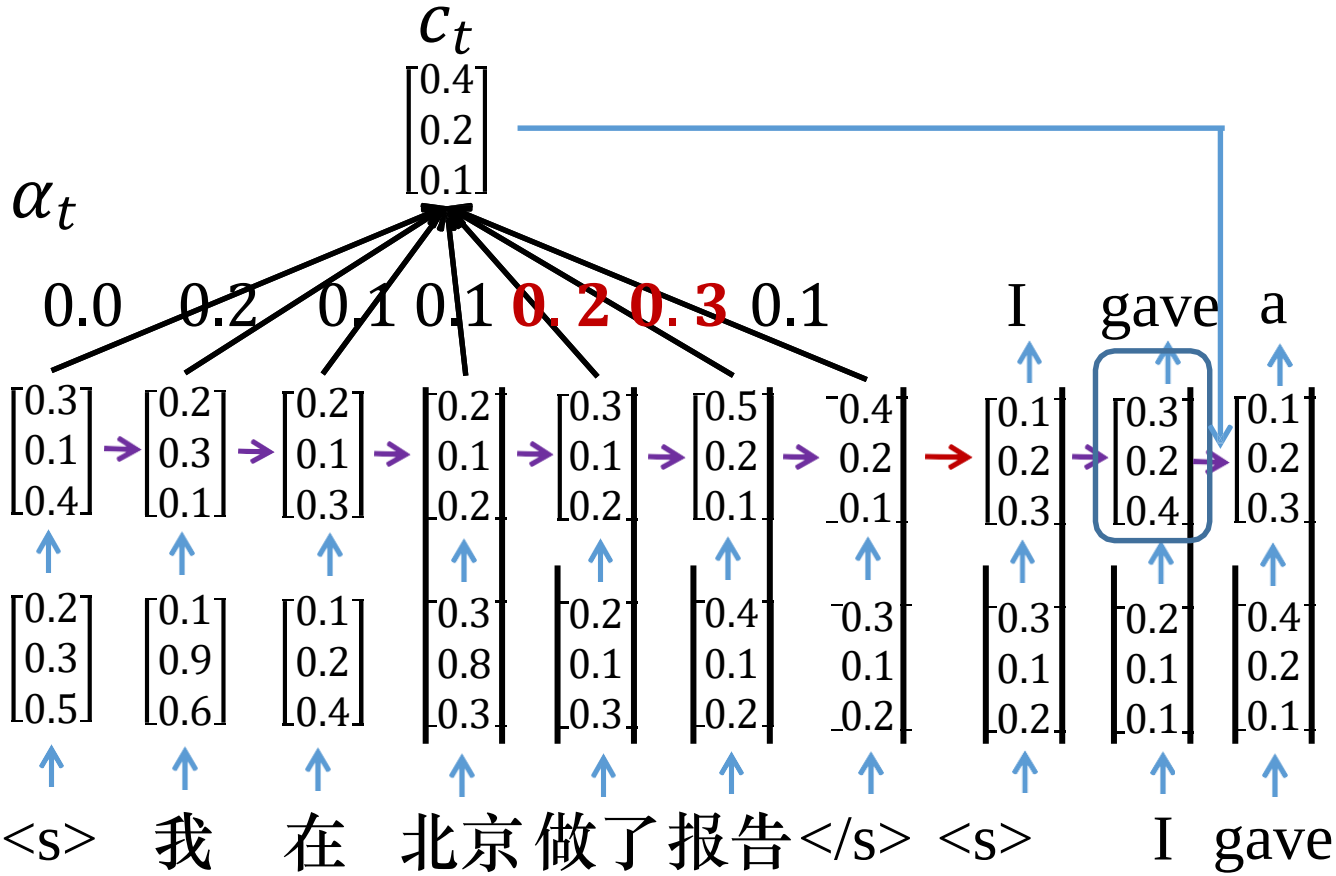
神经机器翻译-注意机制



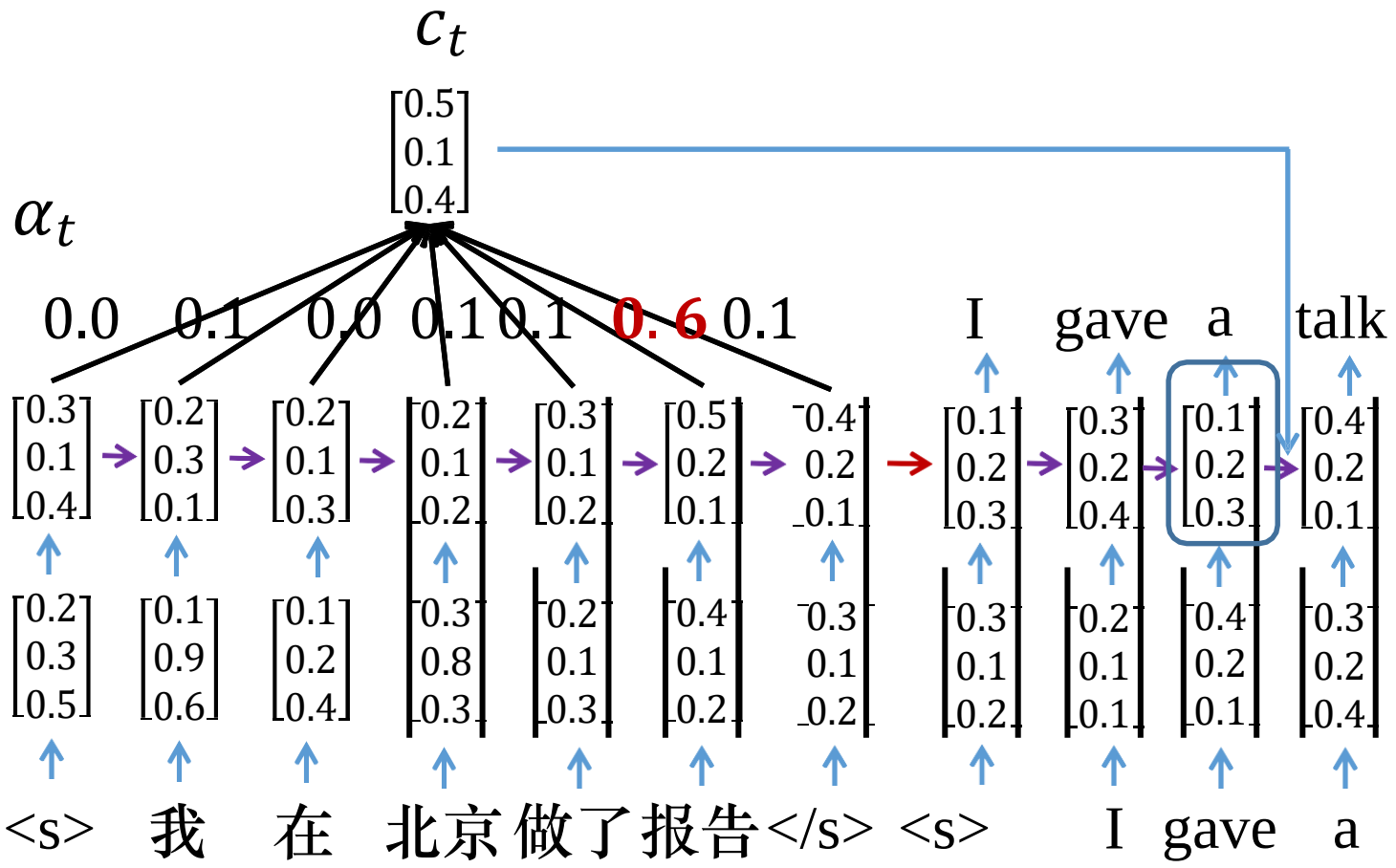
神经机器翻译-注意机制



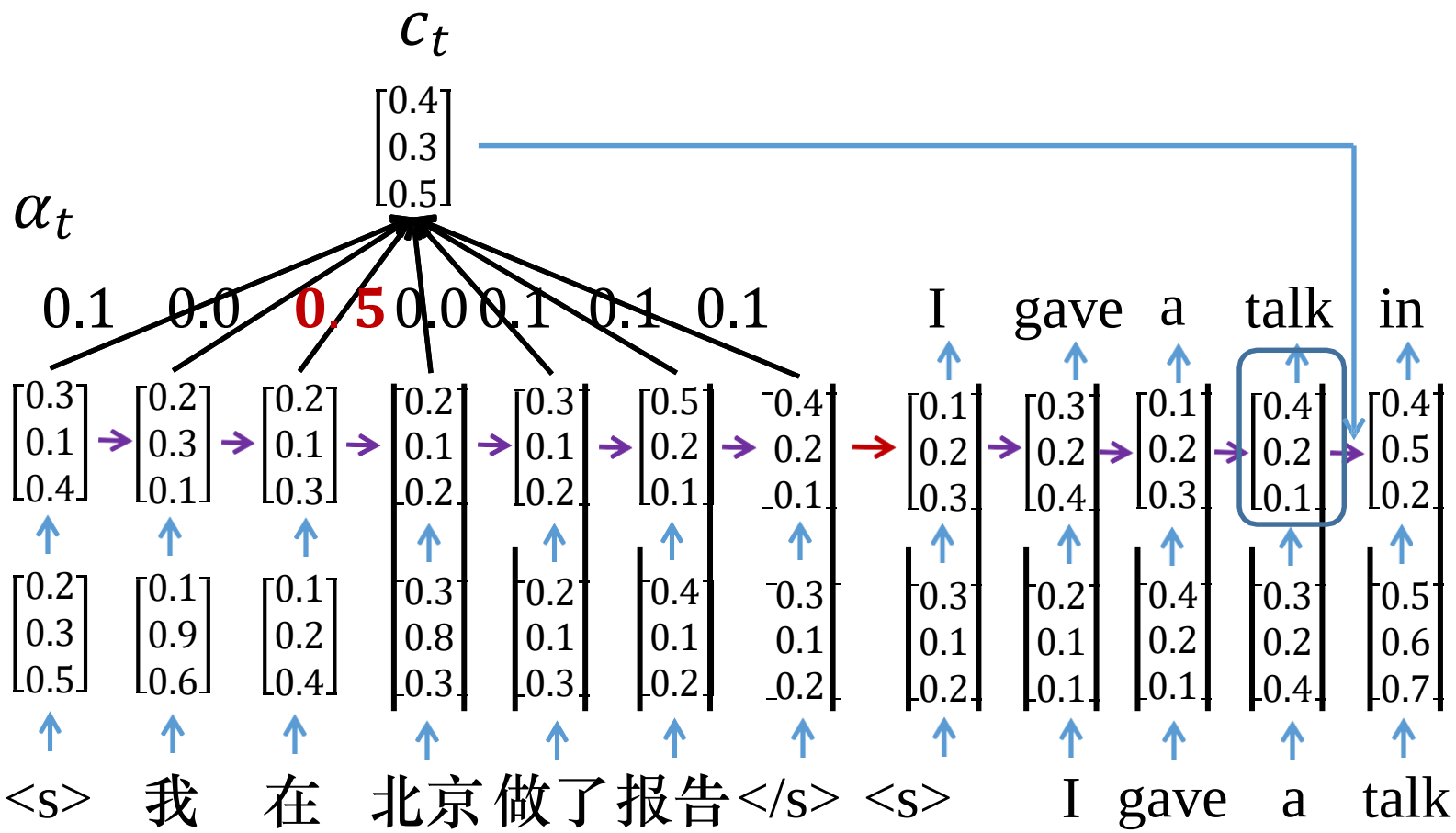
神经机器翻译-注意机制



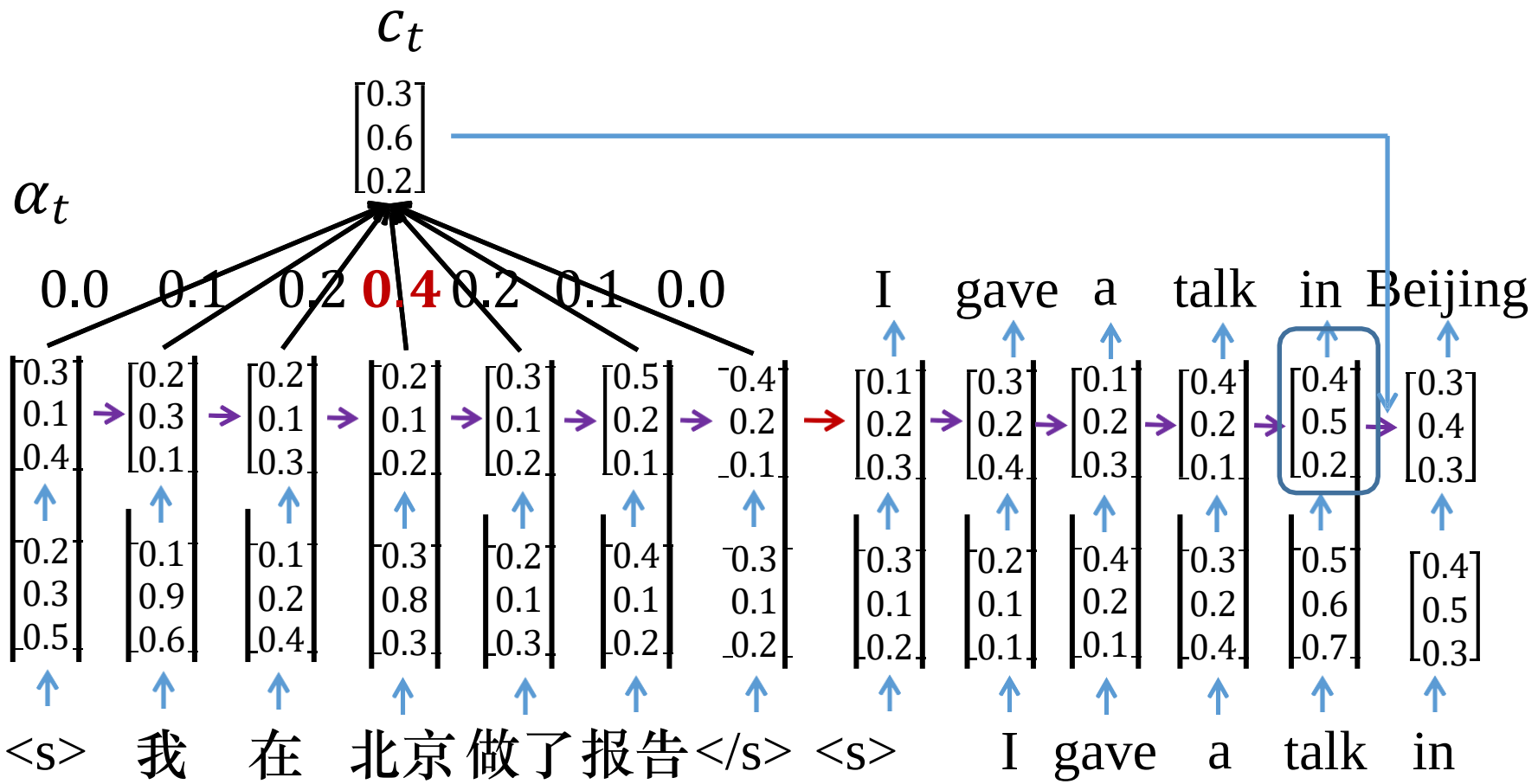
神经机器翻译-注意机制



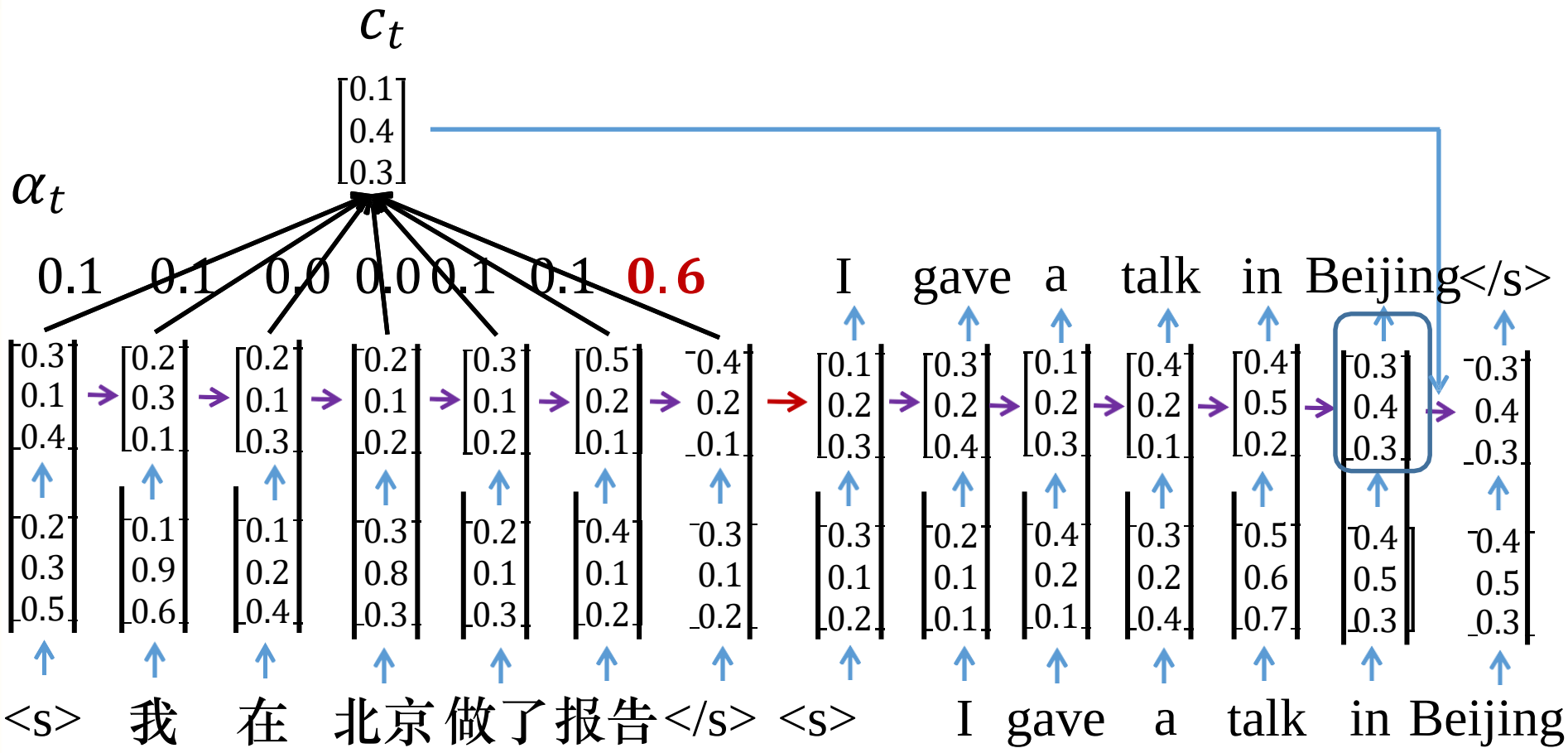
神经机器翻译-注意机制


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

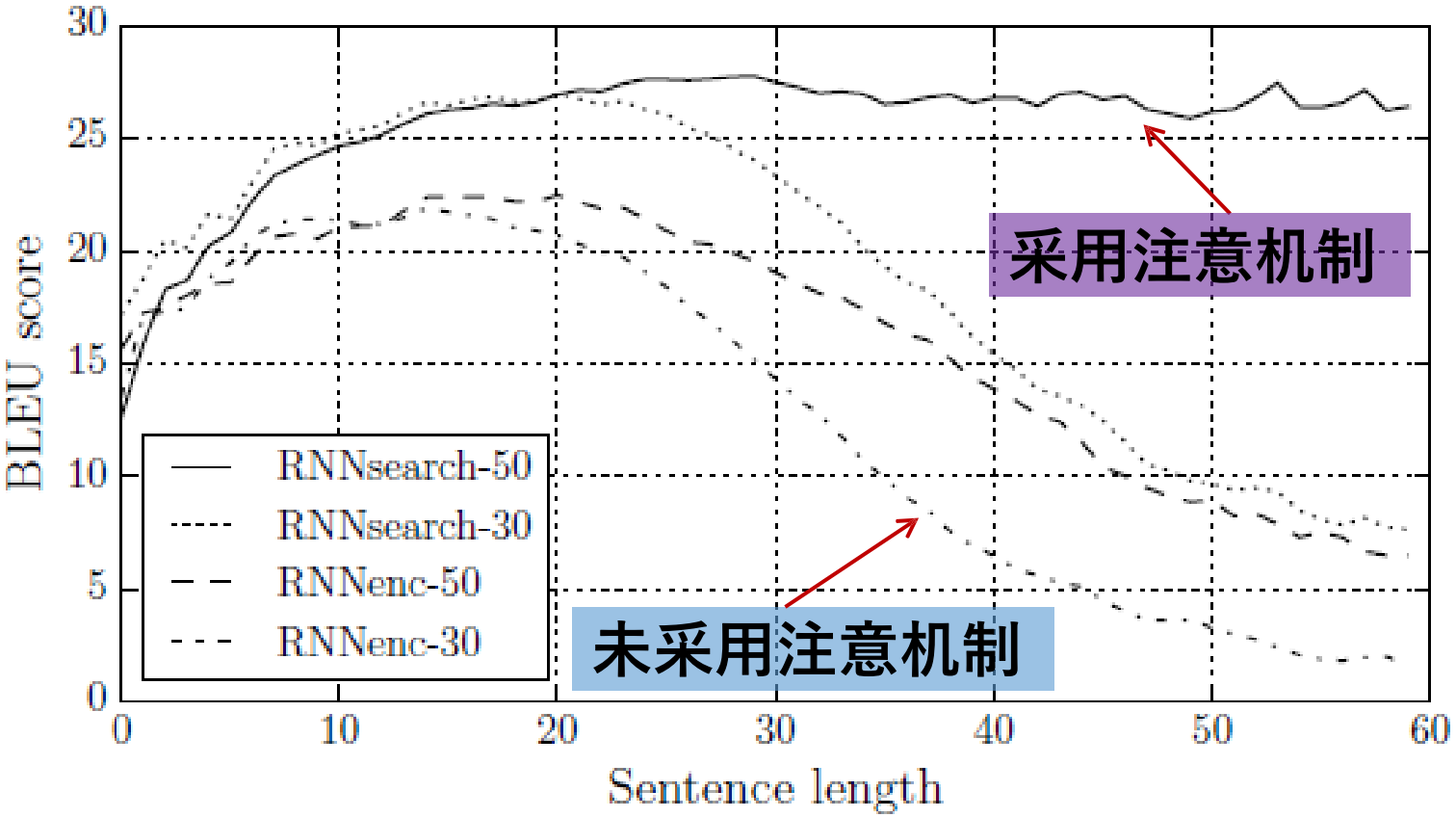
神经机器翻译-注意机制


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

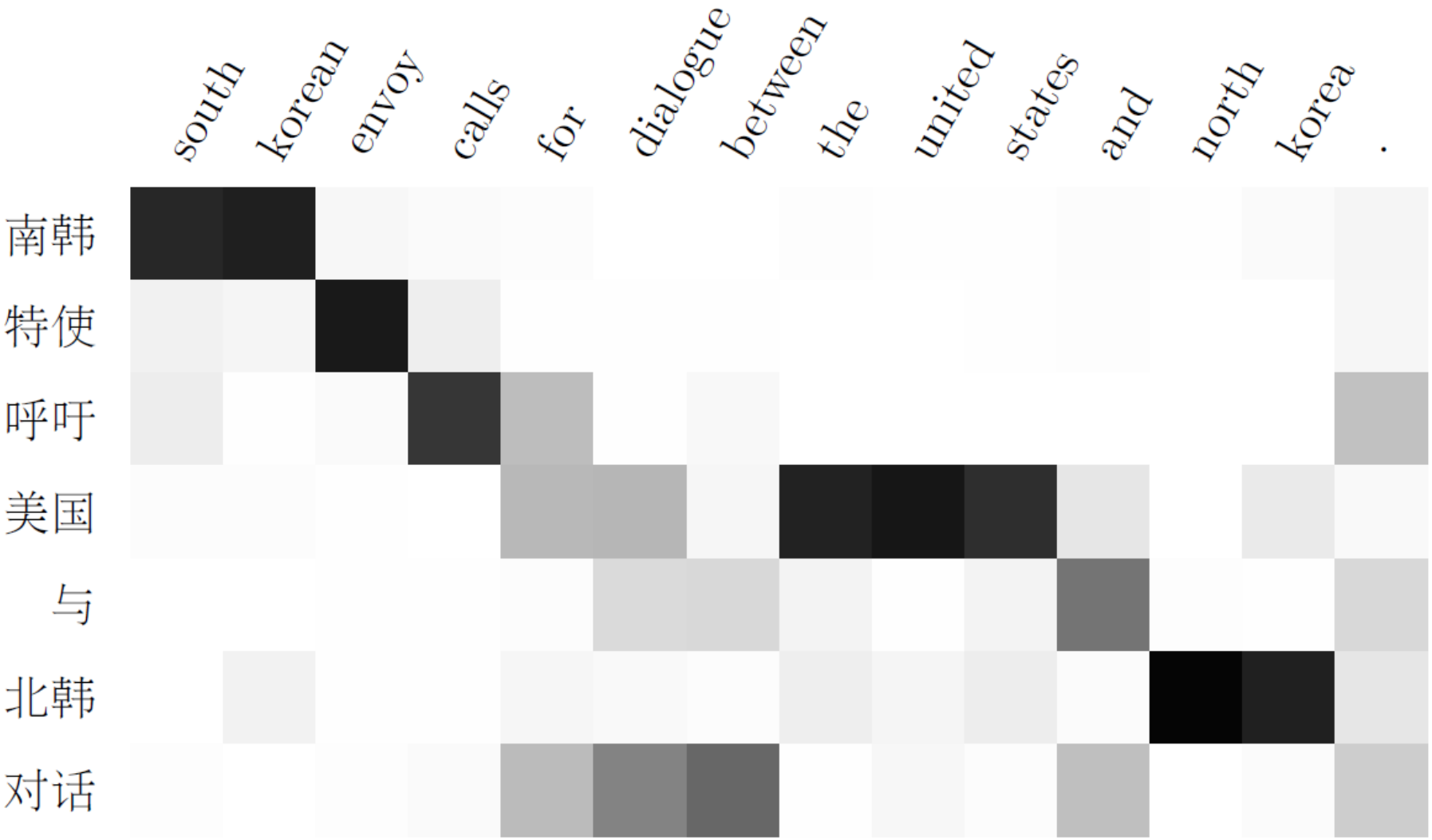

c_t
α_t
0.1 0.1 0.0 0.0 0.1 0.1 0.6
<s> 我 在 北京 做了 报告 </s>
<s> I gave a talk in Beijing </s>
I gave a talk in Beijing

神经机器翻译-计算单元



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

14.3: Transformer

Transformer

Can we get rid of RNN completely ?

RNN are too sequential. Parallelization is not possible since these intermediate contexts are generated sequentially

Transformer

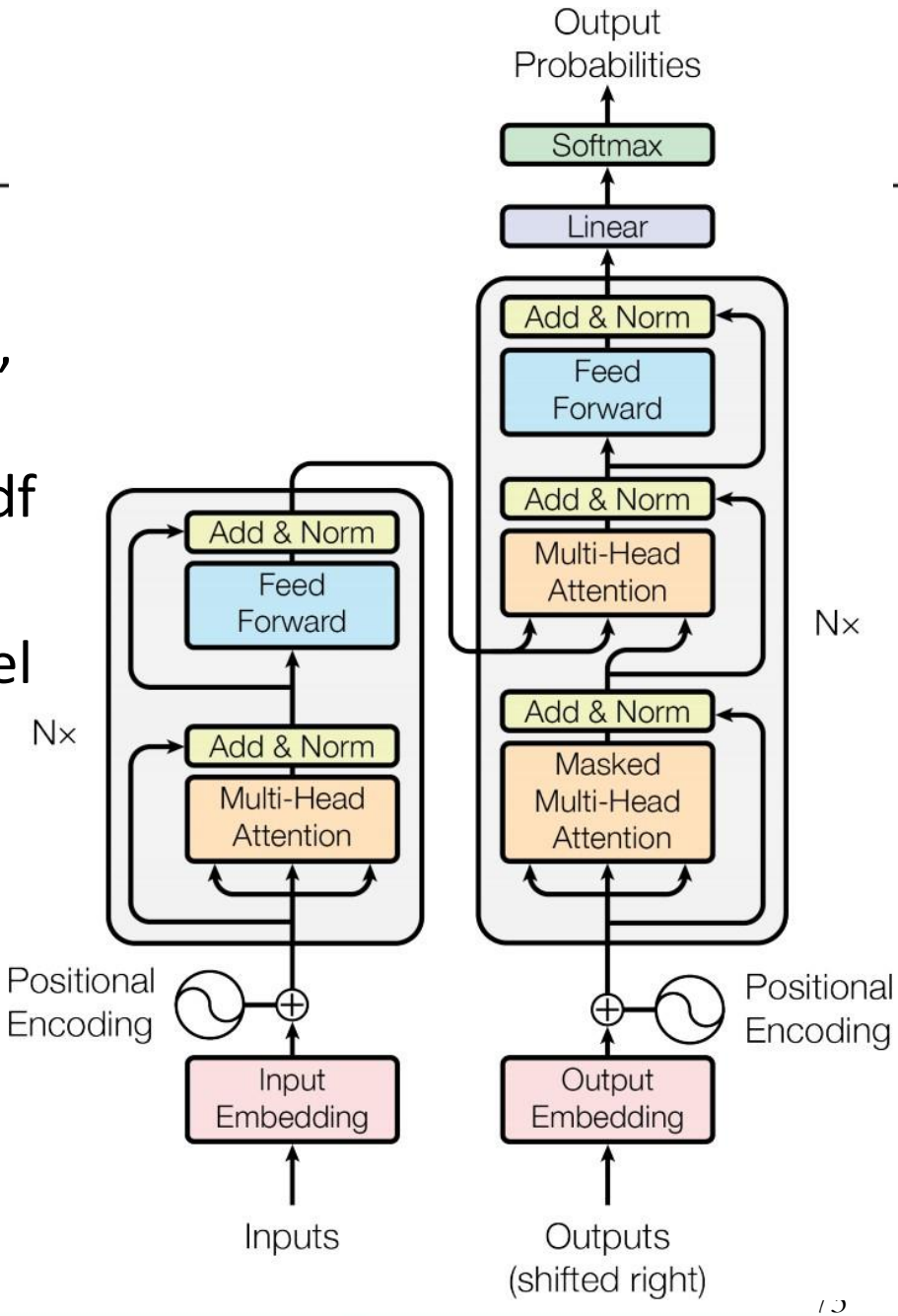
There are three kinds of attention

1. within encoder
2. encoder decoder
3. within decoder

Transformer

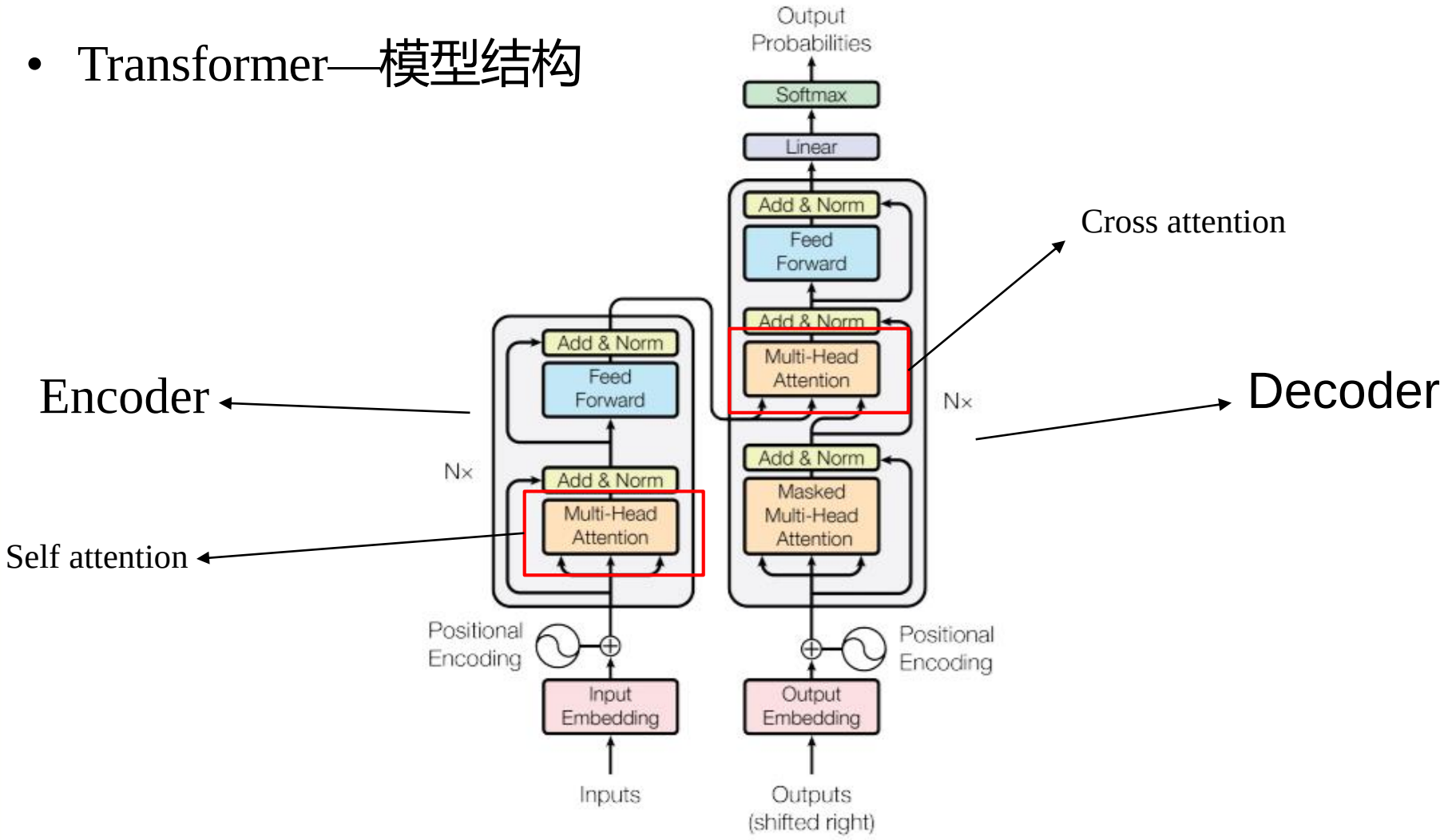
Attention is all you need. 2017.
Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, Polosukhin
https://arxiv.org/pdf/1706.03762.pdf

- Non-recurrent sequence-to-sequence encoder-decoder model
- Task: machine translation with parallel corpus
- Predict each translated word
- Final cost/error function is standard cross-entropy error on top of a softmax classifier

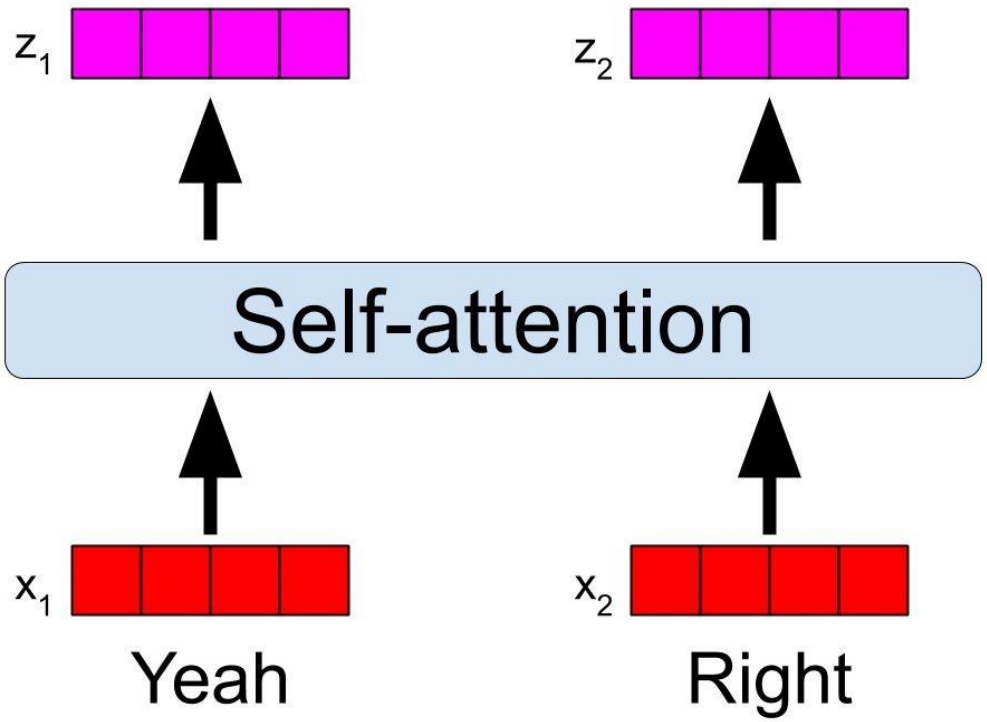


Transformer

- Transformer—模型结构

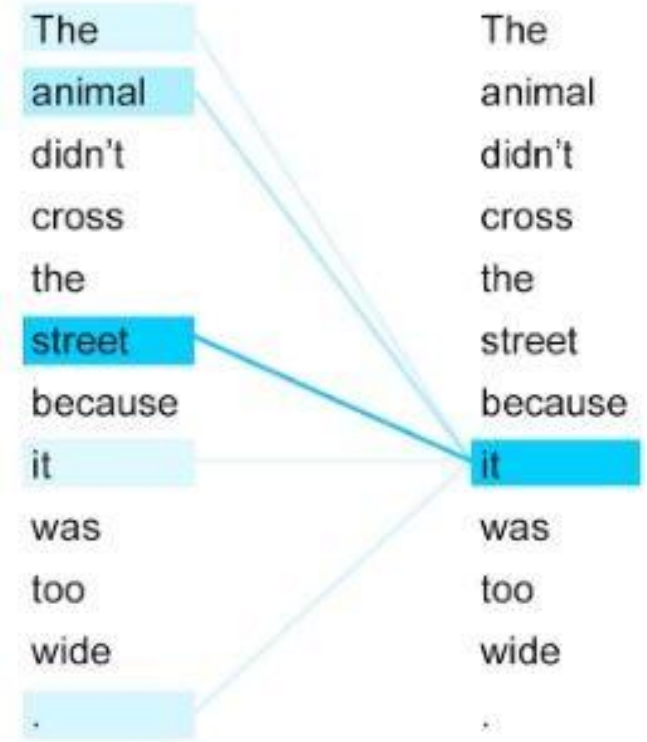


Transformer (self-attention)



A new representation Z

Transformer (self-attention)



Calculate for each word , its dependencies on others

'it' = 0.7 * street + .2 * animal + …..

Transformer (self-attention)

- Self-attention mechanism directly models relationships between all words in a sentence, regardless of their respective position
- Self attention allows connections between words within a sentence.

Transformer

Use self attention instead of RNNs, CNN

- Computation can be parallelized
- Ability to learn long term dependencies

14.4: 机器翻译应用

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
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.
2016年9月上线，有详细系统描述
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

GNMT：谷歌神经翻译系统

Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi
yonghui,schuster,zhifengc,qvl,mnorouzi@google.com

Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, Jeffrey Dean

GNMT：谷歌神经翻译系统

“In our experiments we found that for NMT systems to achieve good accuracy, both the encoder and decoder RNNs have to be deep enough to capture subtle irregularities in the source and target languages.”

GNMT：attention

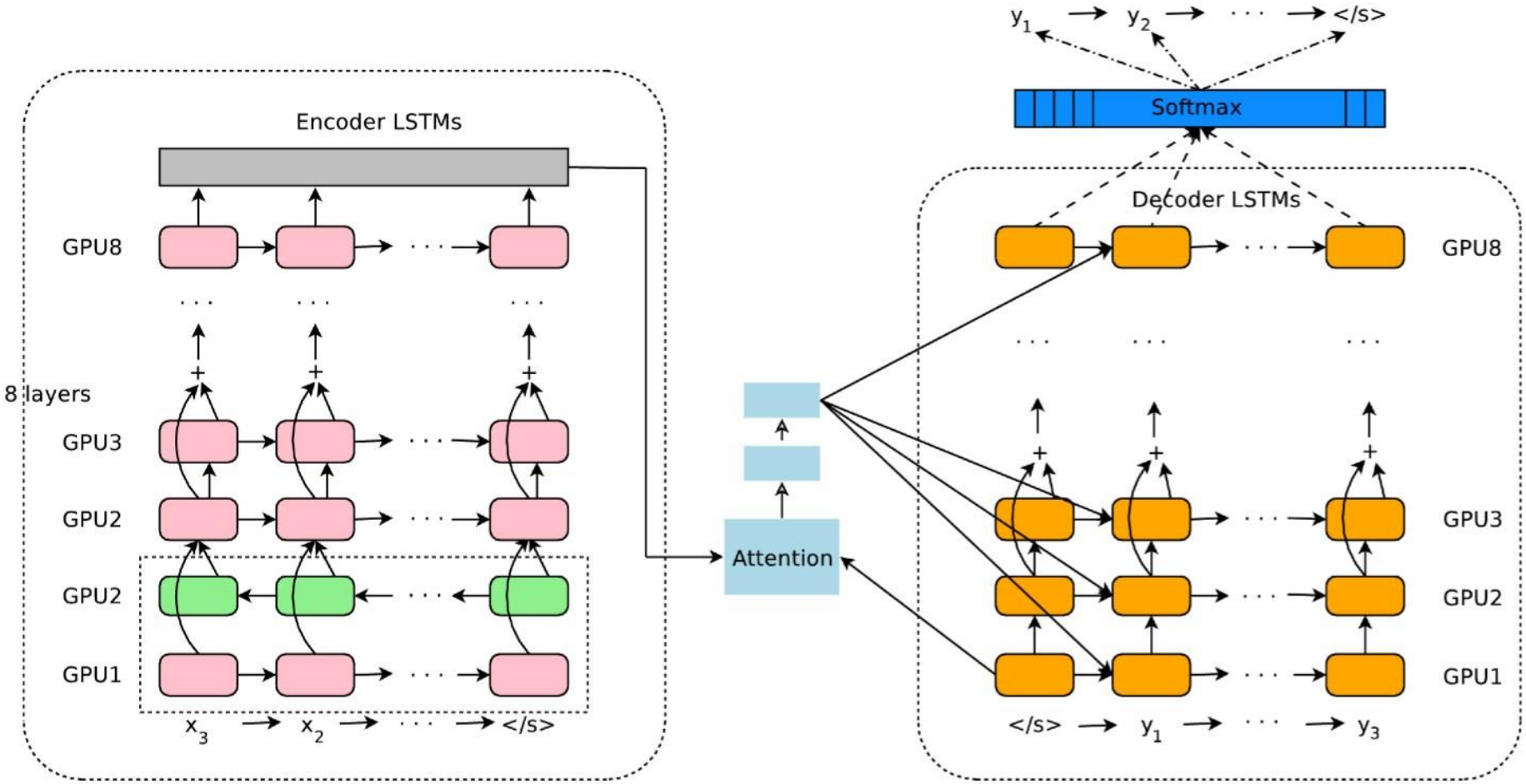
- More specifically, let $\mathbf{y}_{i-1}$ be the decoder-RNN output from the past decoding time step (in our implementation, we use the output from the bottom decoder layer).
- Attention context $\mathbf{a}_i$ for the current time step is computed according to the following formulas:

$$s_t = AttentionFunction(\mathbf{y}_{i-1}, \mathbf{x}_t) \quad \forall t, \quad 1 \leq t \leq M$$

$$p_t = \exp(s_t) / \sum_{t=1}^{M} \exp(s_t) \qquad \forall t, \quad 1 \leq t \leq M$$

$$\mathbf{a}_i = \sum_{t=1}^{M} p_t . \mathbf{x}_t$$

GNMT：谷歌神经翻译系统



GNMT：谷歌神经翻译系统

GNMT：Residual Connections

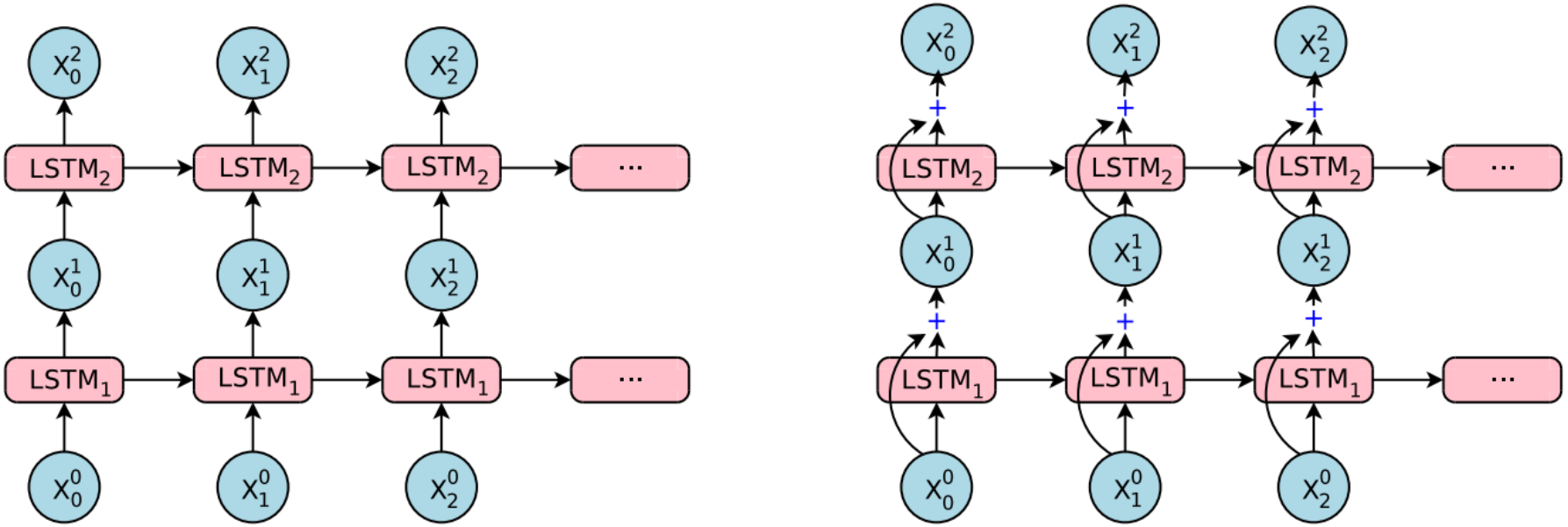


Figure 2: The difference between normal stacked LSTM and our stacked LSTM with residual connections. On the left: simple stacked LSTM layers [41]. On the right: our implementation of stacked LSTM layers with residual connections. With residual connections, input to the bottom LSTM layer ($\mathbf{x_i^0}$'s to LSTM$_1$) is element-wise added to the output from the bottom layer ($\mathbf{x_i^1}$'s). This sum is then fed to the top LSTM layer (LSTM$_2$) as the new input.

GNMT：Model Parallelism

- Model parallelism is used to improve the speed of the gradient computation on each replica.
- The encoder and decoder networks are partitioned along the depth dimension and are placed on multiple GPUs, effectively running each layer on a different GPU.
- Since all but the first encoder layer are uni-directional, layer *i* + 1 can start its computation before layer i is fully finished.

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

致谢

- 中科院自动化所《自然语言处理》，宗成庆&张家俊；

Thank you!

权小军 中山大学数据科学与计算机学院