ACL 2017

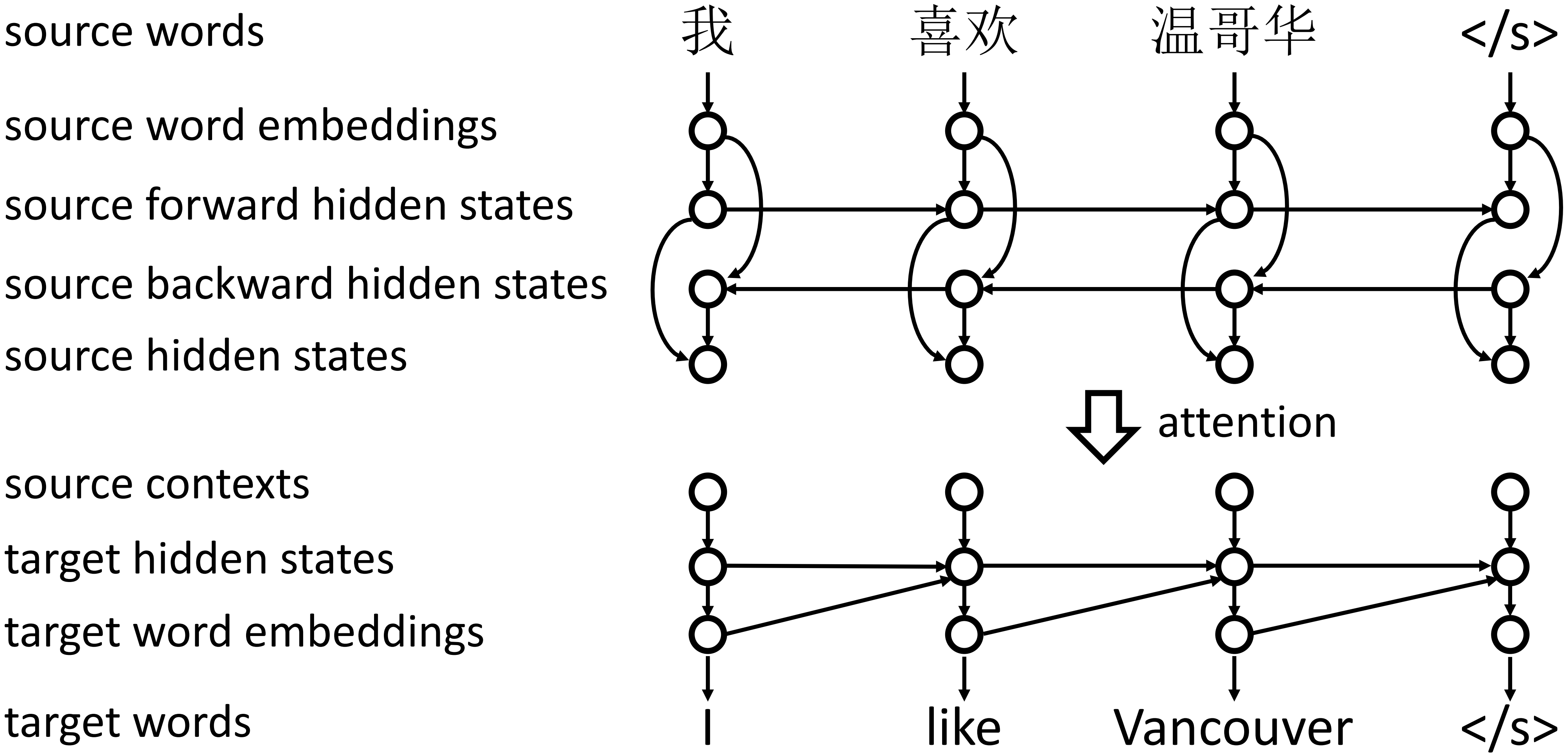
Visualizing and Understanding Neural Machine Translation

Yanzhuo Ding, Yang Liu, Huanbo Luan, Maosong Sun

Tsinghua University

Neural Machine Translation

- Idea: using neural networks to translate languages (Bahdanau et al., 2015)



source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

source contexts

target hidden states

target word embeddings

target words

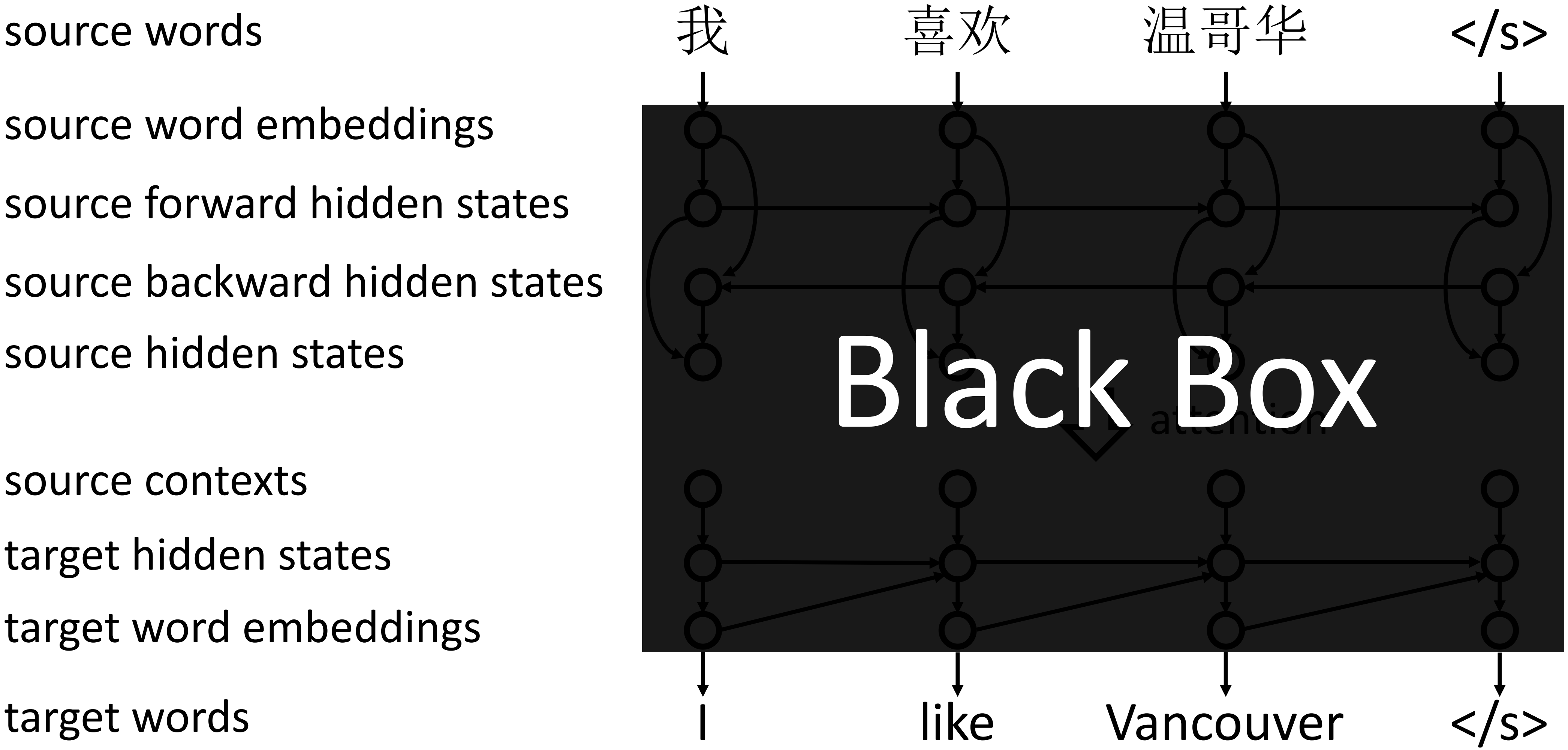
我 喜欢 温哥华 </s>

attention

I like Vancouver </s>

Challenge

- It is hard to visualize and understand the internal workings



source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

source contexts

target hidden states

target word embeddings

target words

我 喜欢 温哥华 </s>

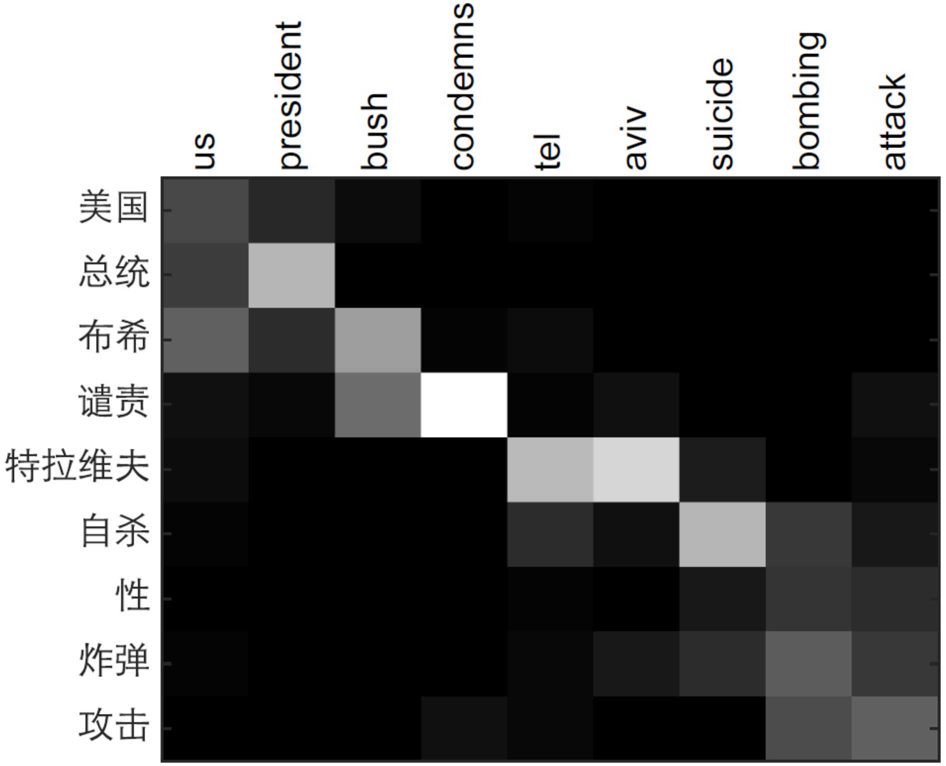
Black Box

I like Vancouver </s>

Related Work

attention mechanism

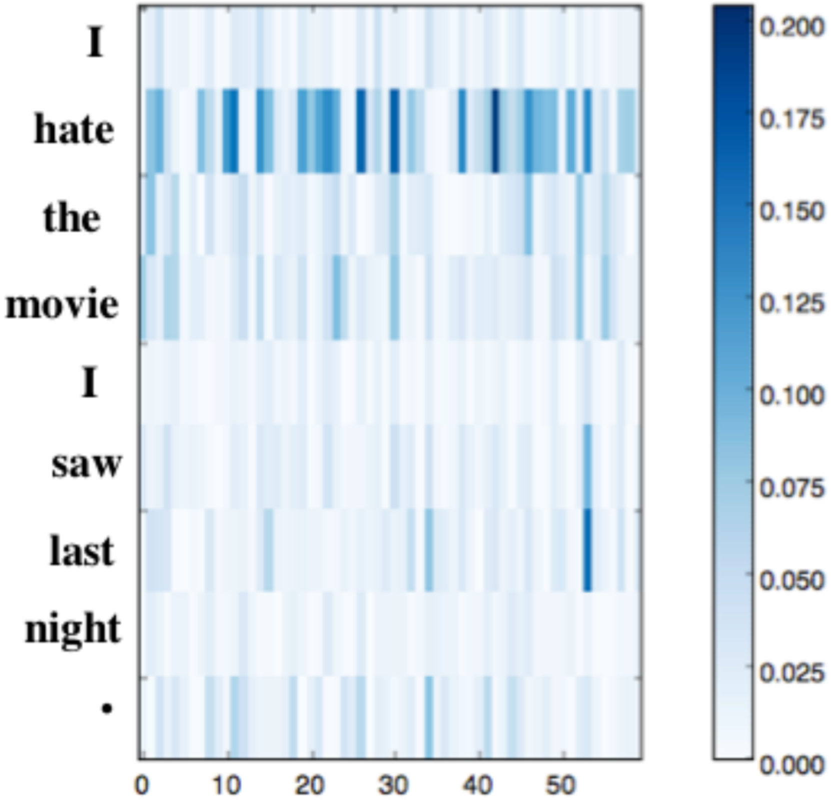
(Bahdanau et al., 2015)



restricted to the connection between input and output

first-derivative saliency

(Li et al., 2016)

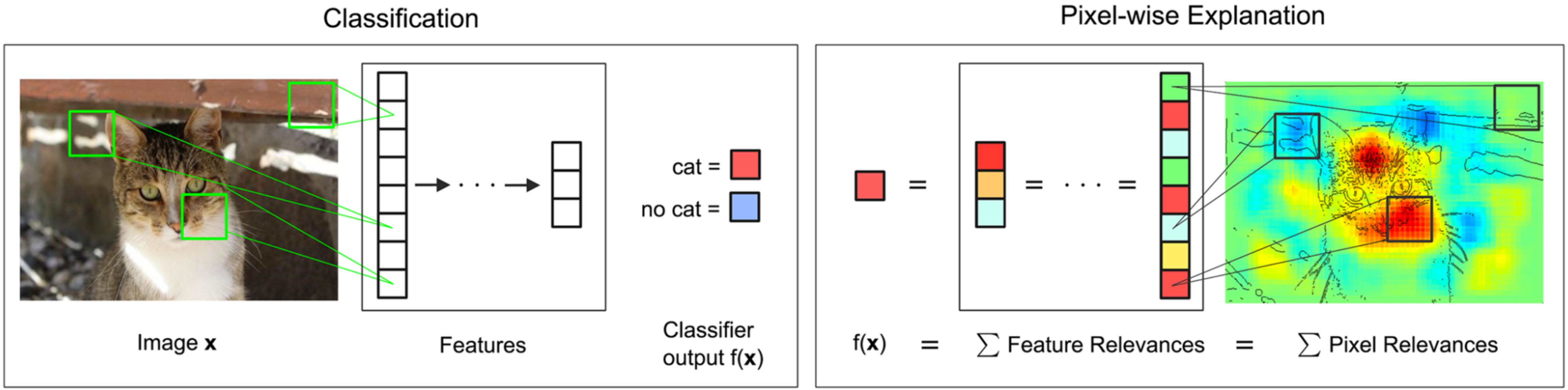


require neural activations to be differentiable

Related Work

layer-wise relevance propagation (LRP)

(Bach et al., 2015)

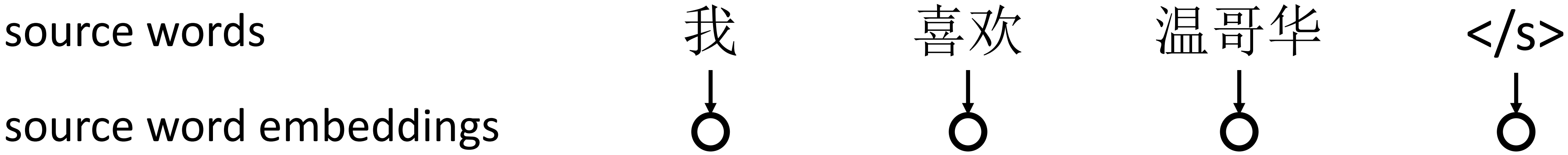

Classification
Pixel-wise Explanation
cat =
no cat =
Image **x**
Features
Classifier output f(**x**)
=
= · · · =
f(**x**) = ∑ Feature Relevances = ∑ Pixel Relevances

calculating the relevance between two arbitrary neurons without requiring differentiability

Our Work

source words 我 喜欢 温哥华 </s>

Our Work



source words

source word embeddings

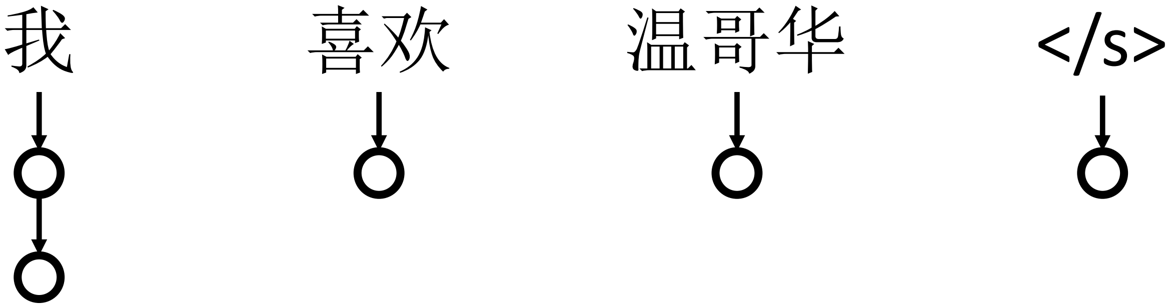
我

喜欢

温哥华

</s>

Our Work



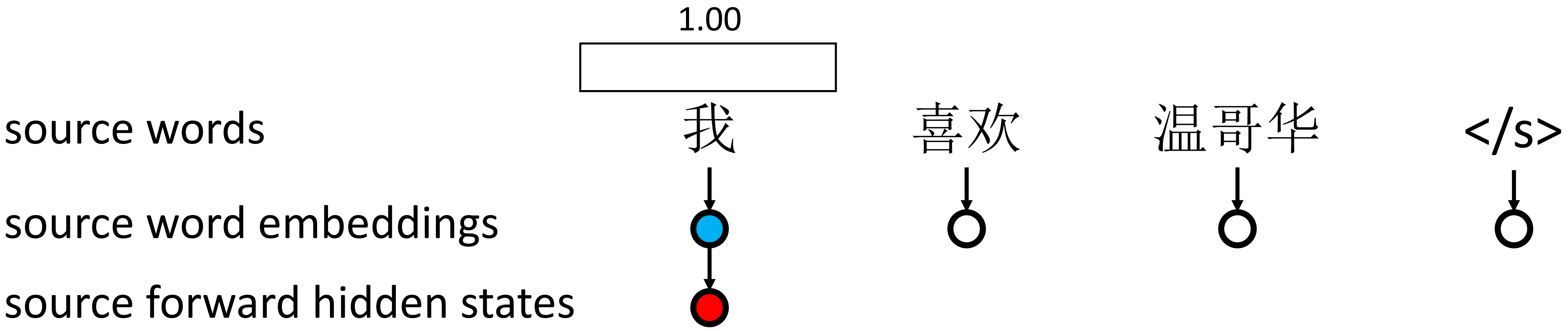
source words

source word embeddings

source forward hidden states

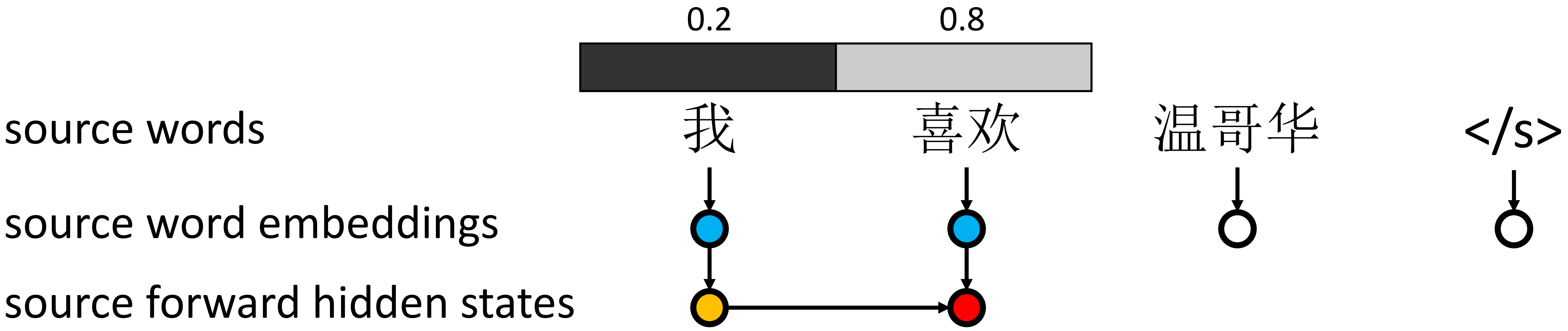
我 喜欢 温哥华 </s>

Our Work



🔴	targeted vector of neurons
🔵	relevant vector of neurons
🟡	intermediate vector of neurons
○	irrelevant vector of neurons
1.0	relevance

Our Work



0.2

0.8

source words

我

喜欢

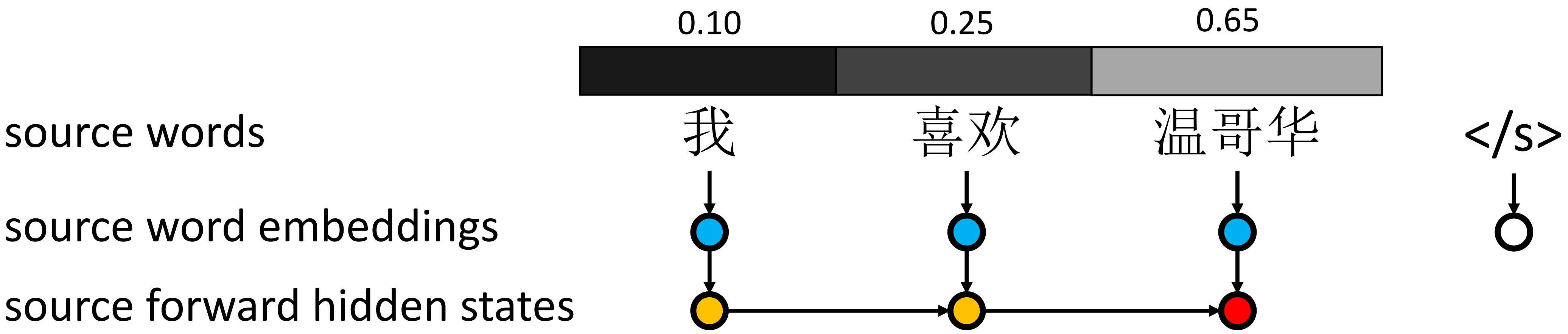
温哥华

</s>

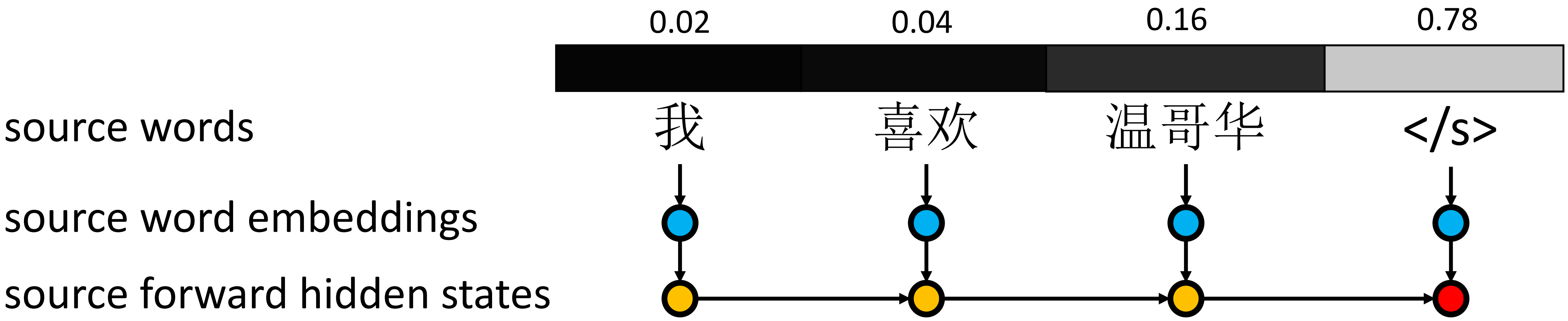
source word embeddings

source forward hidden states

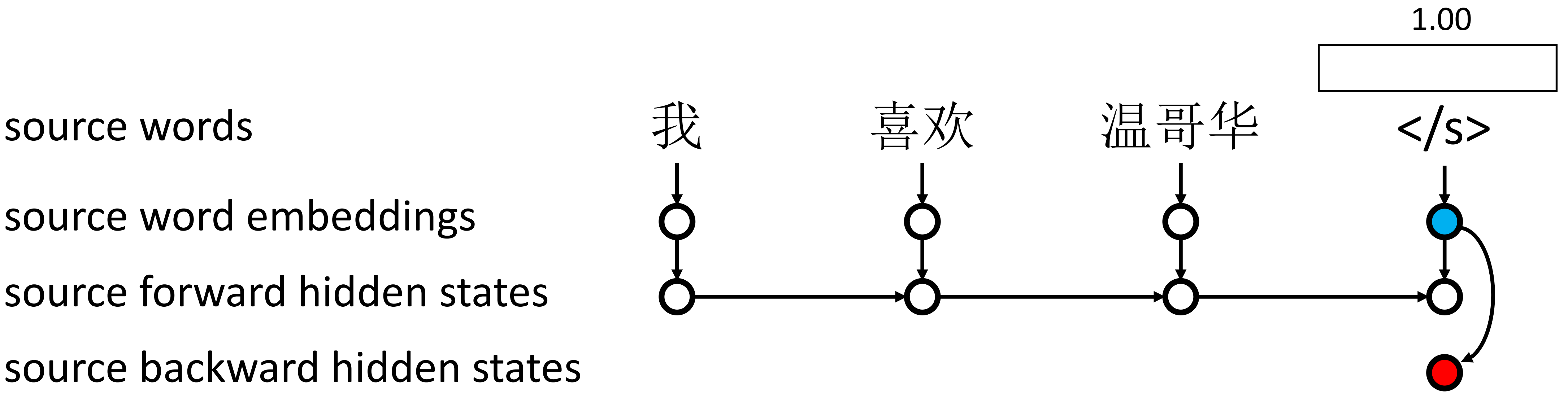
Our Work


0.10
0.25
0.65
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states

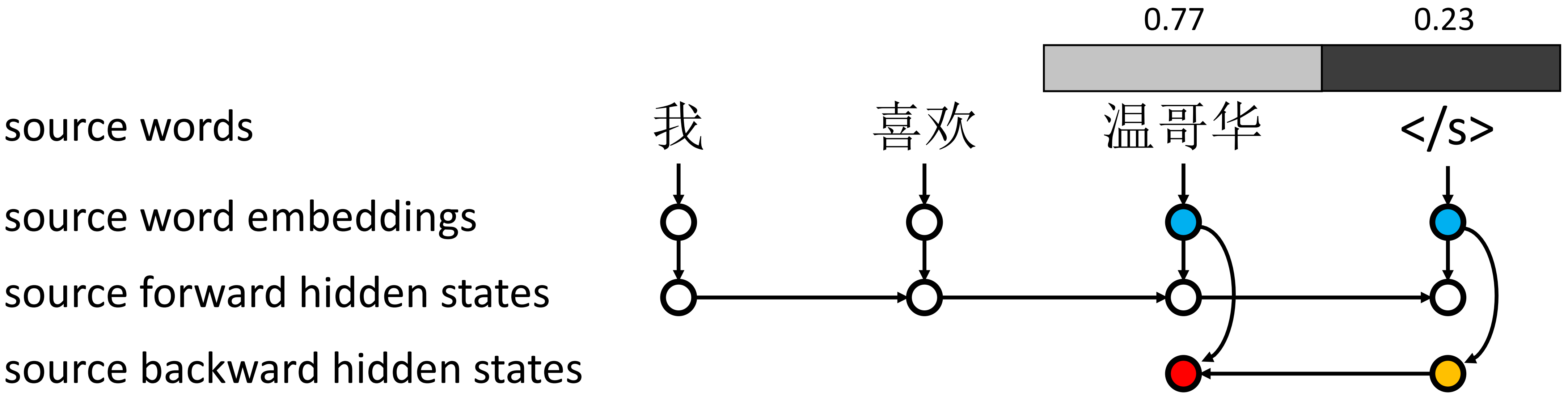
Our Work


0.02
0.04
0.16
0.78
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states

Our Work


1.00
source words
source word embeddings
source forward hidden states
source backward hidden states
我
喜欢
温哥华
</s>

Our Work



0.77 0.23

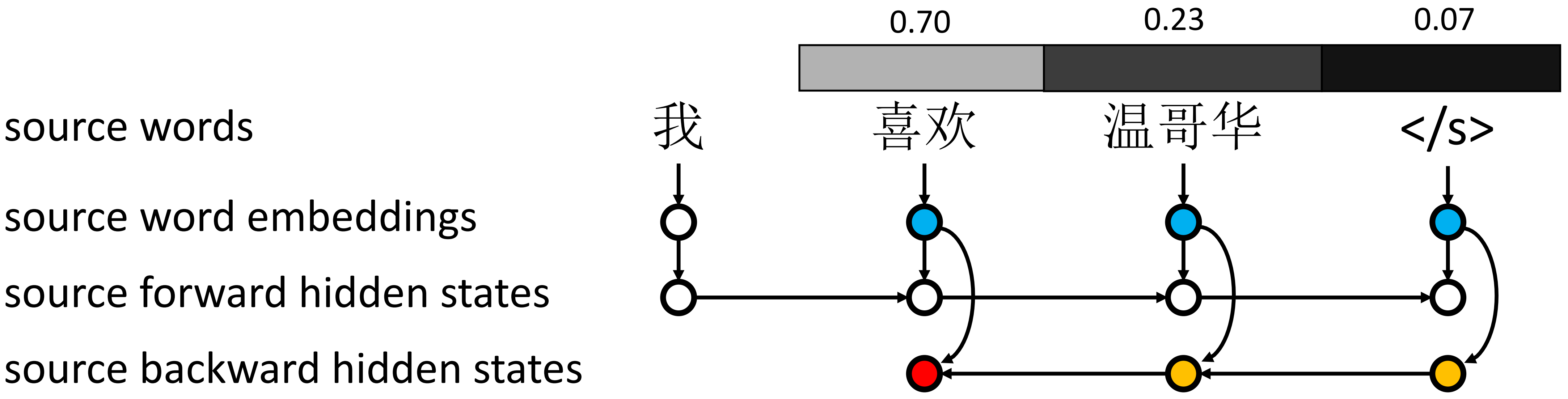
source words 我 喜欢 温哥华 </s>

source word embeddings

source forward hidden states

source backward hidden states

Our Work



0.70

0.23

0.07

source words

我

喜欢

温哥华

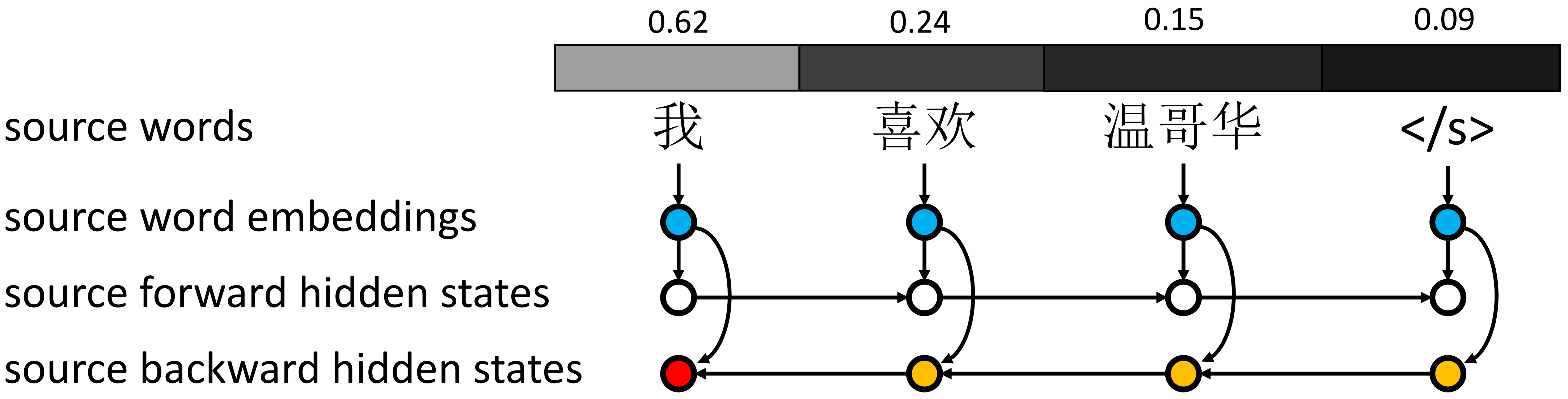
</s>

source word embeddings

source forward hidden states

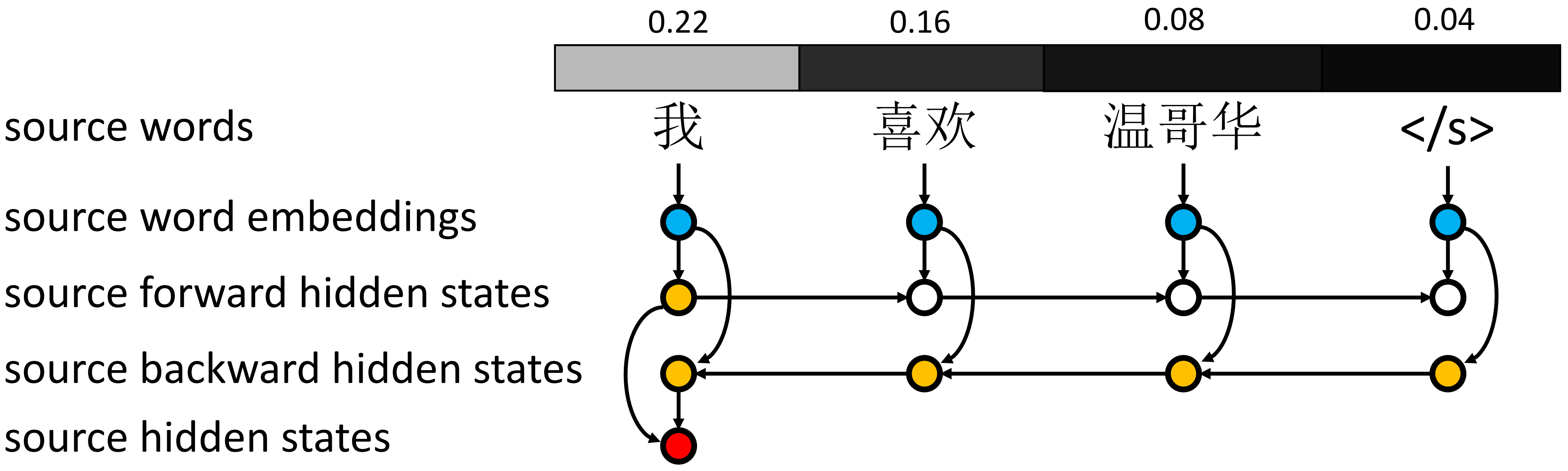
source backward hidden states

Our Work



0.62
0.24
0.15
0.09
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states
source backward hidden states

Our Work



0.22 0.16 0.08 0.04

source words 我 喜欢 温哥华 </s>

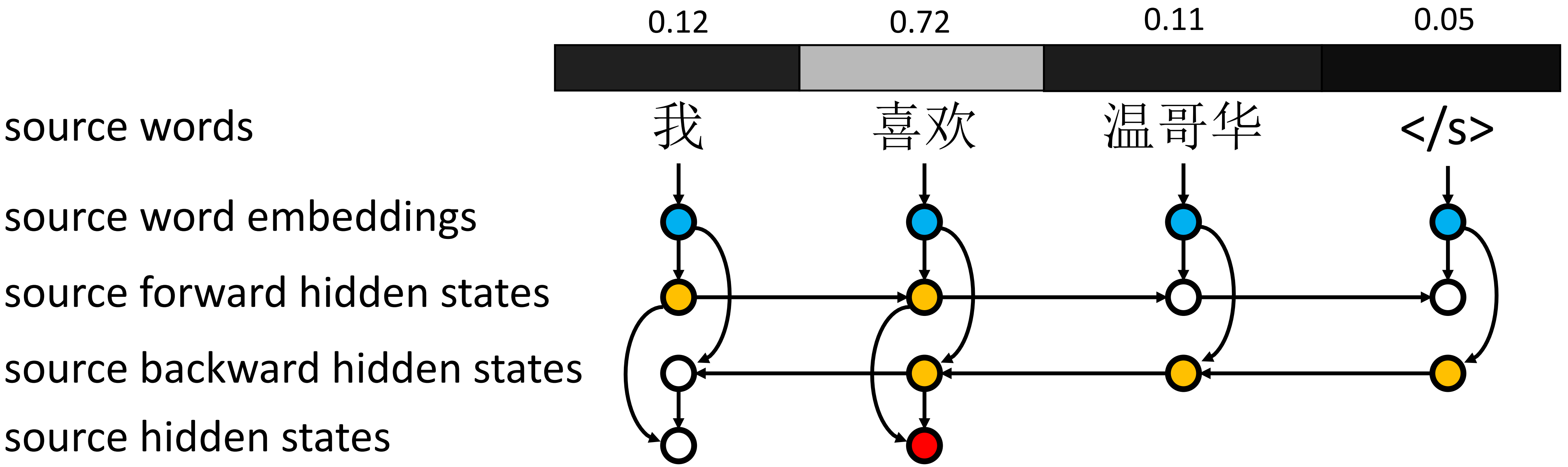
source word embeddings

source forward hidden states

source backward hidden states

source hidden states

Our Work



0.12 0.72 0.11 0.05

source words

我 喜欢 温哥华 </s>

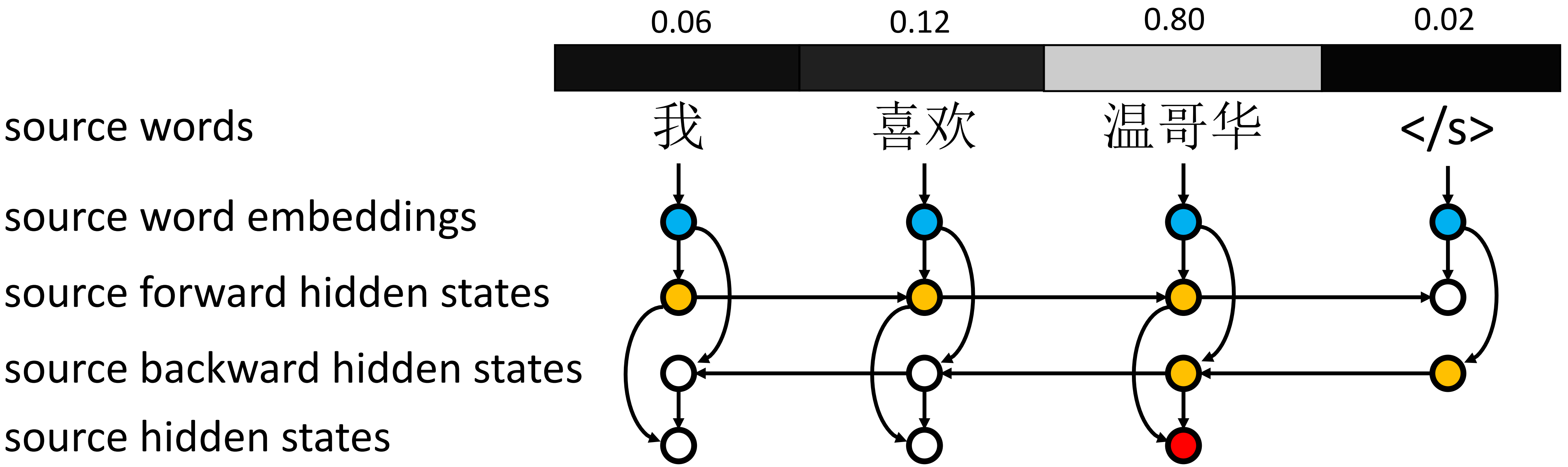
source word embeddings

source forward hidden states

source backward hidden states

source hidden states

Our Work



0.06 0.12 0.80 0.02

source words

我 喜欢 温哥华 </s>

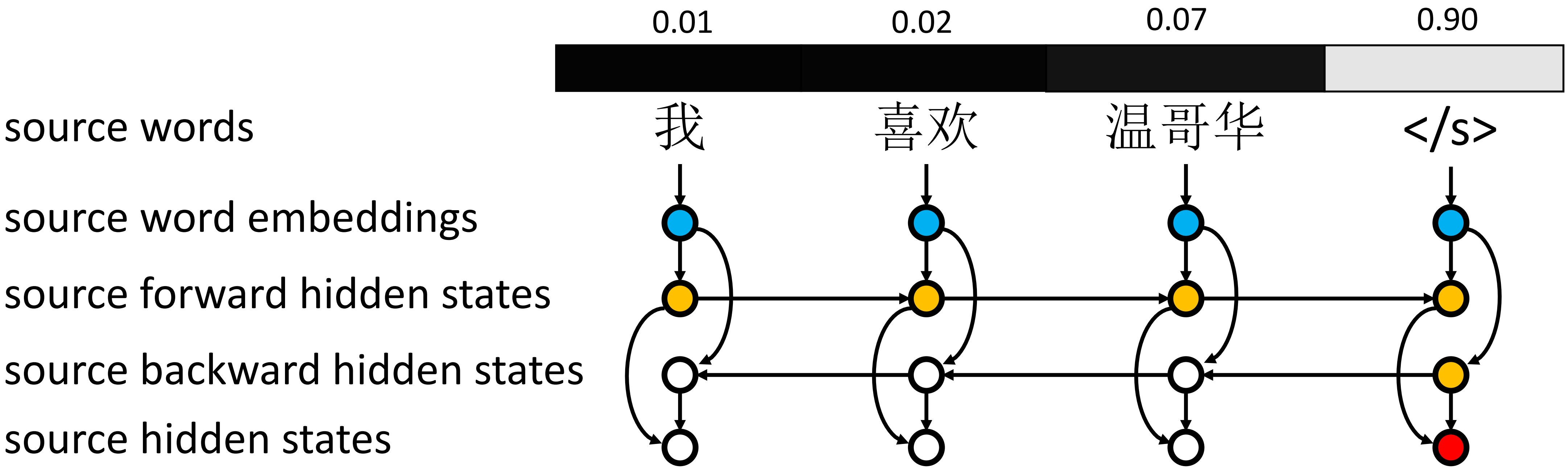
source word embeddings

source forward hidden states

source backward hidden states

source hidden states

Our Work



0.01 0.02 0.07 0.90

source words 我 喜欢 温哥华 </s>

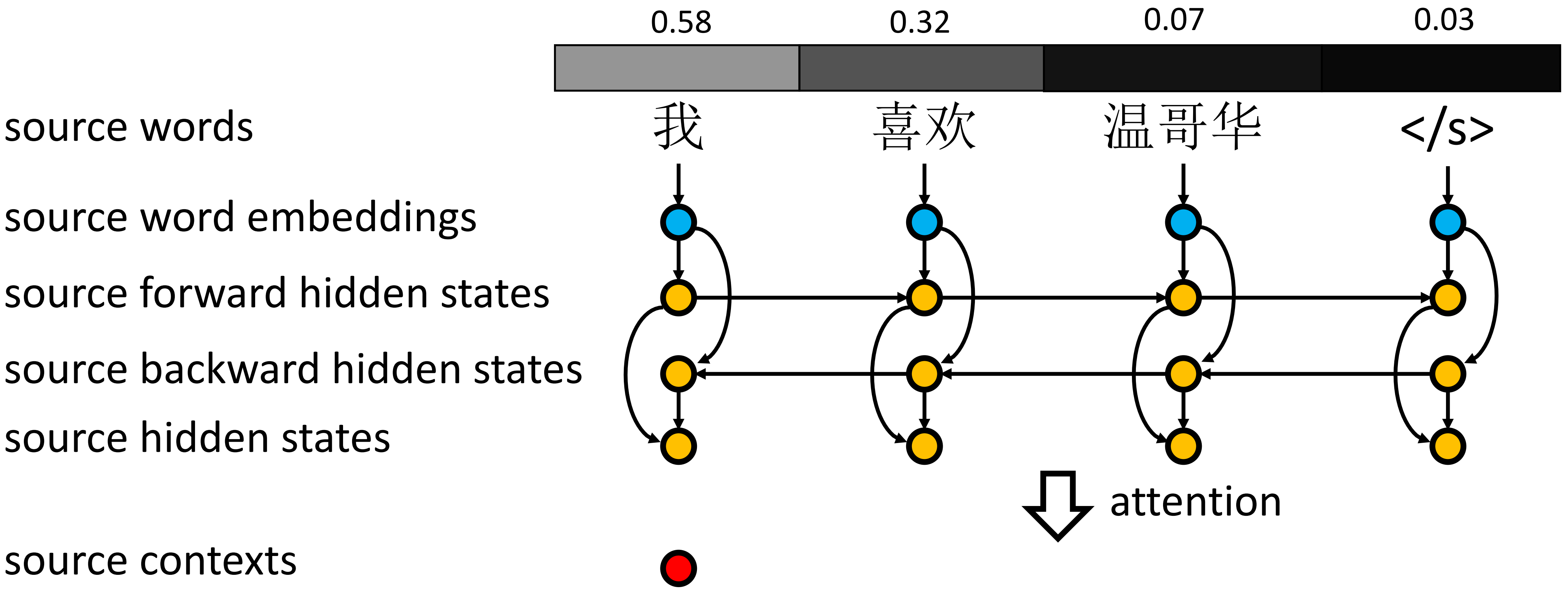
source word embeddings

source forward hidden states

source backward hidden states

source hidden states

Our Work



0.58 0.32 0.07 0.03

source words 我 喜欢 温哥华 </s>

source word embeddings

source forward hidden states

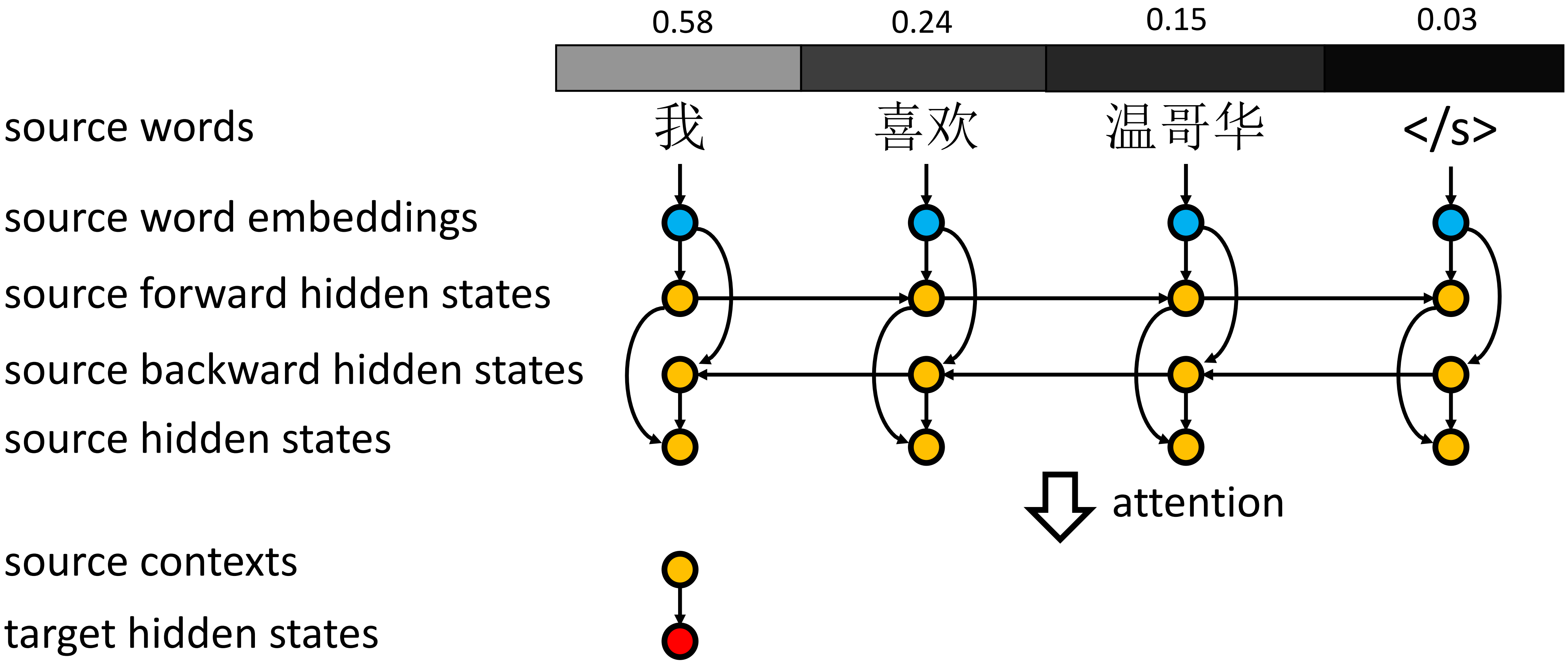
source backward hidden states

source hidden states

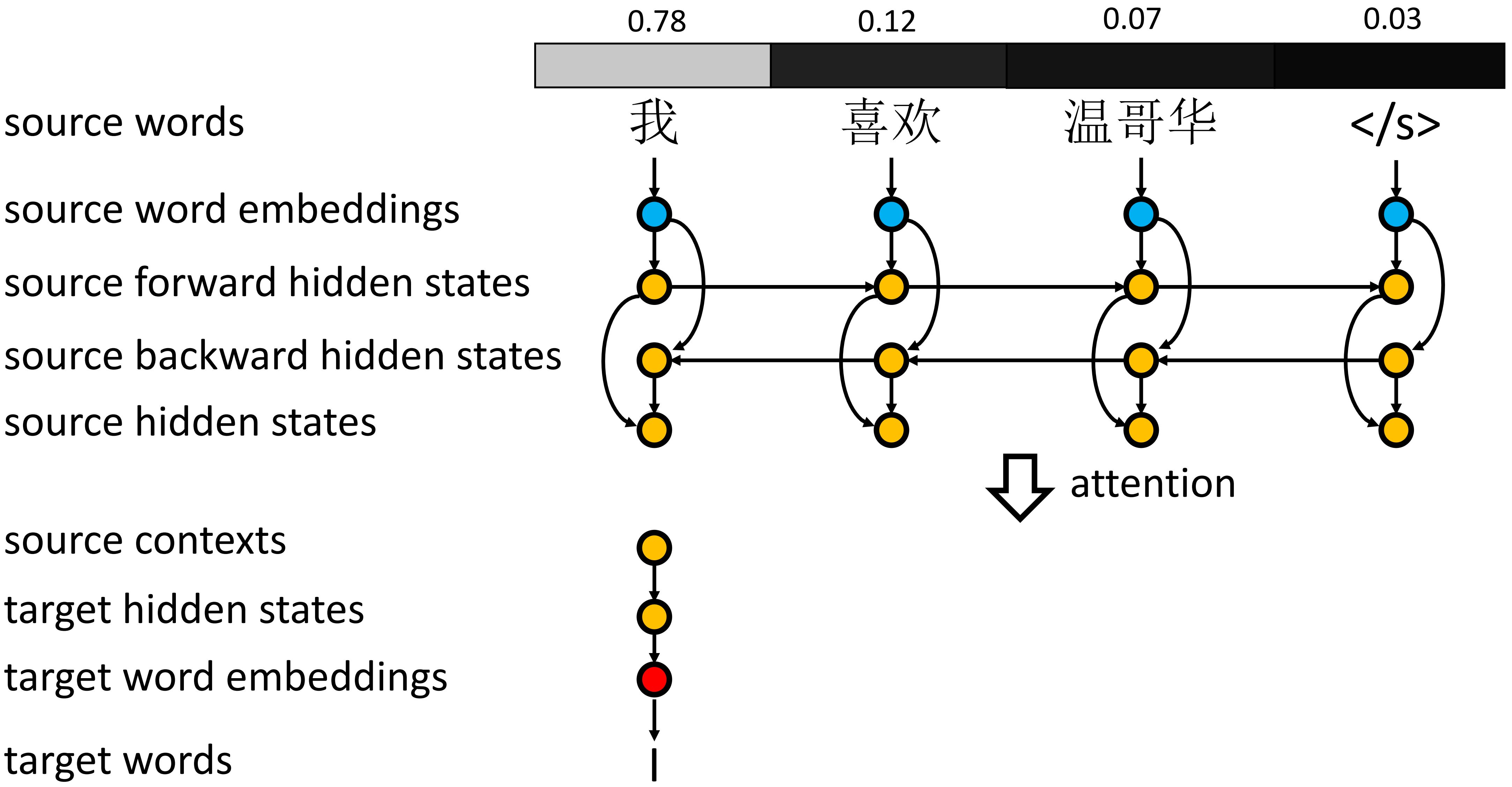
attention

source contexts

Our Work


0.58
0.24
0.15
0.03
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
attention
source contexts
target hidden states

Our Work



0.78 0.12 0.07 0.03

source words 我 喜欢 温哥华 </s>

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

attention

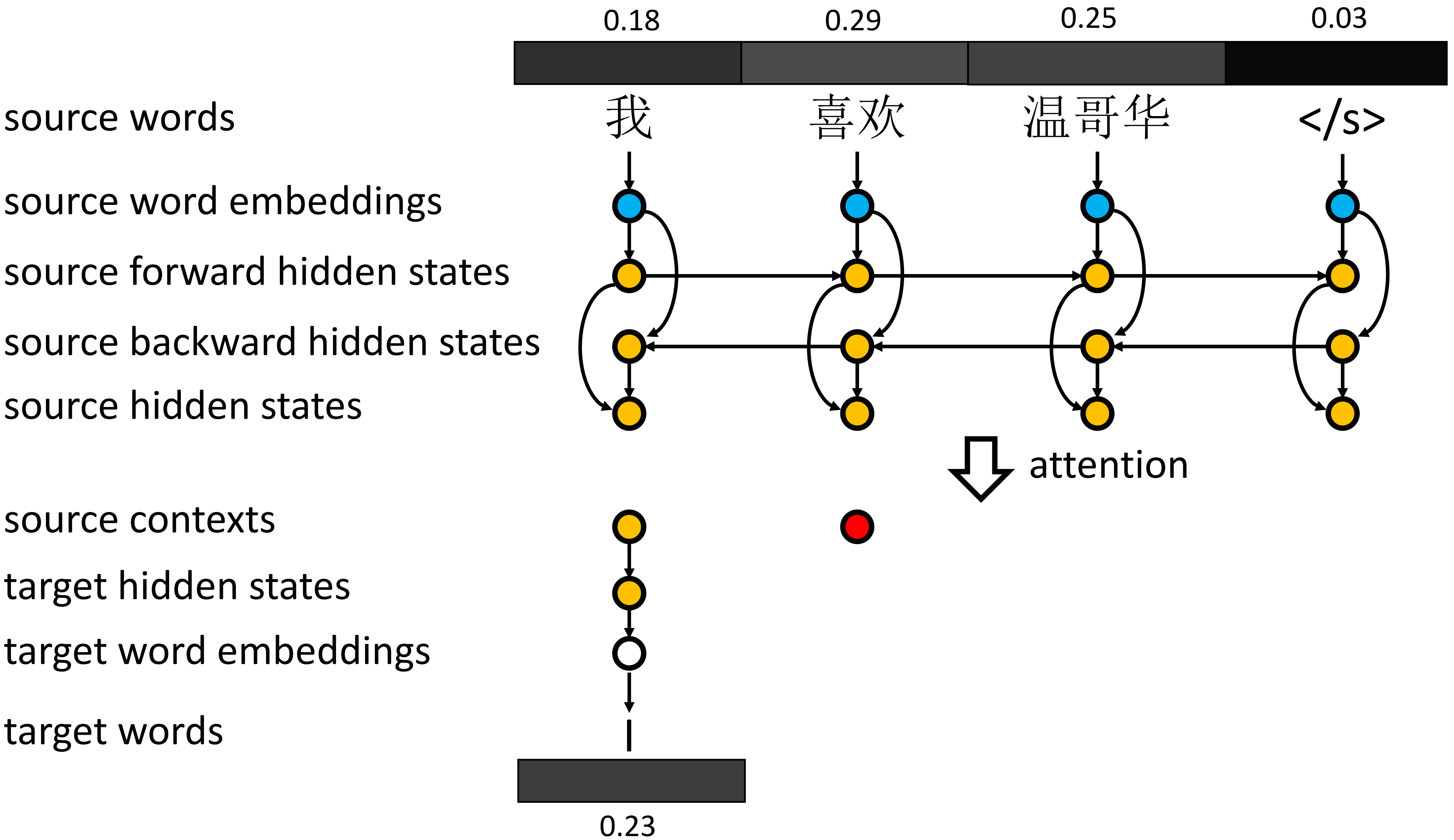
source contexts

target hidden states

target word embeddings

target words I

Our Work



0.18 0.29 0.25 0.03

source words 我 喜欢 温哥华 </s>

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

attention

source contexts

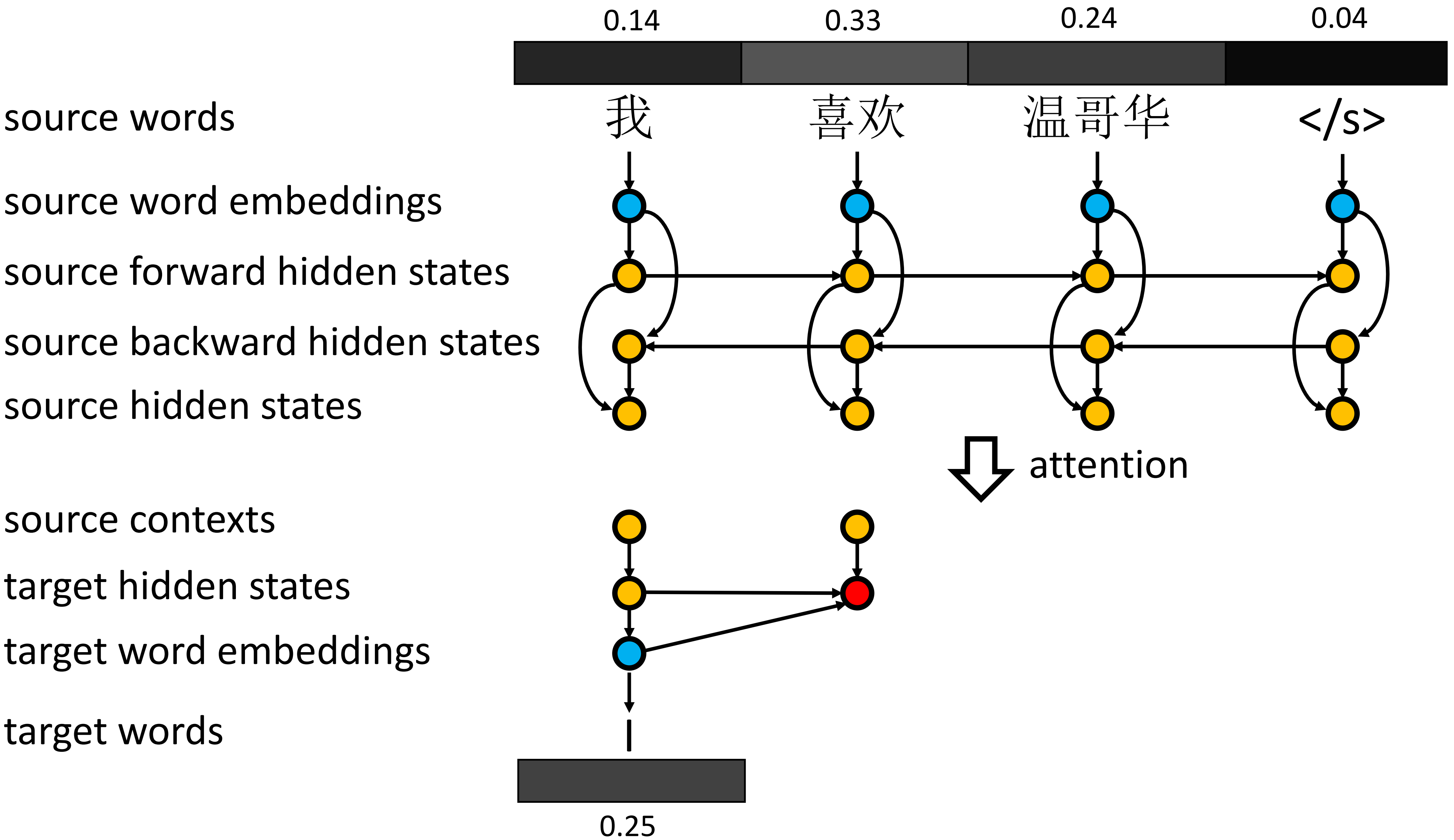
target hidden states

target word embeddings

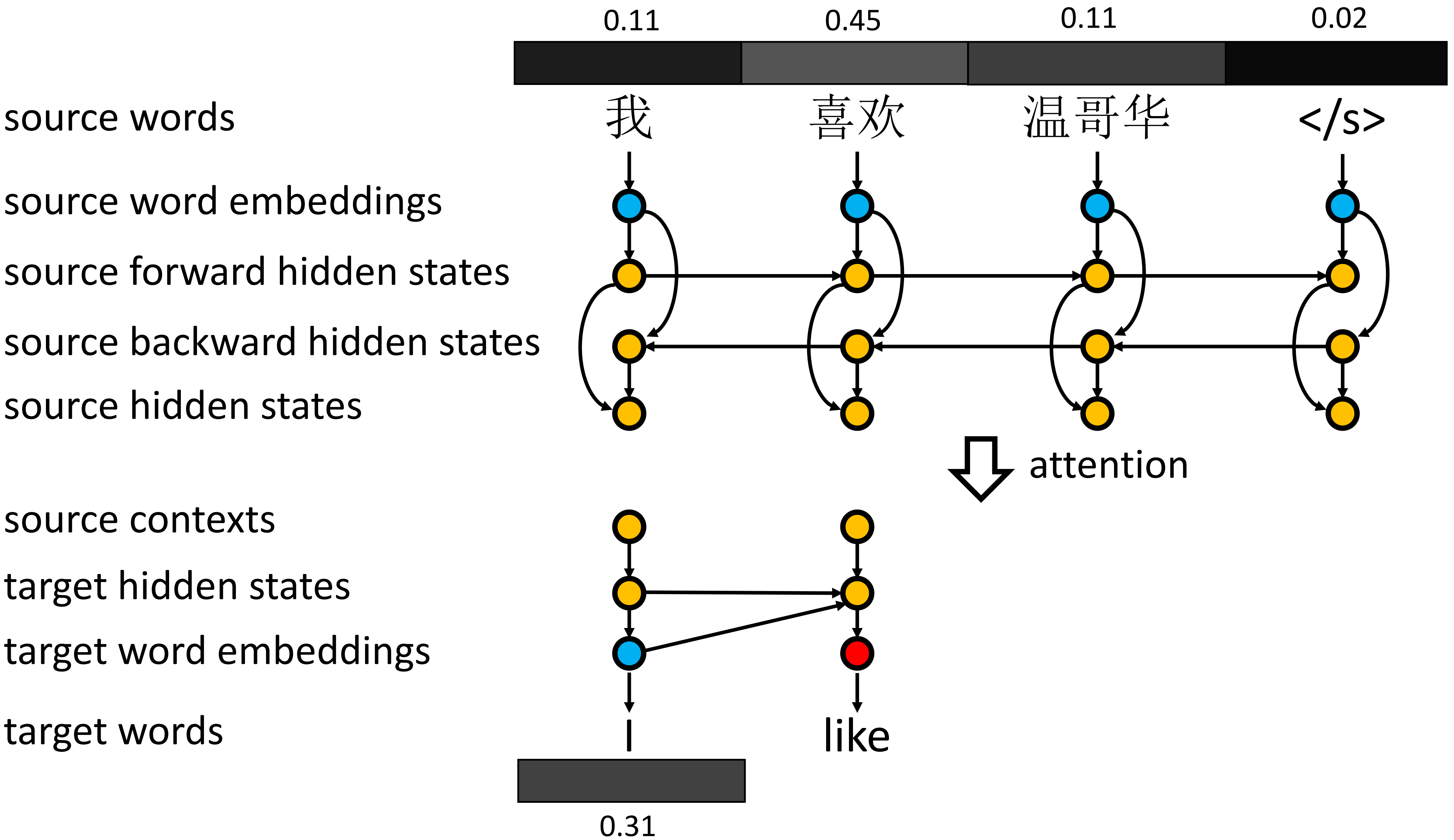
target words I

0.23

Our Work


0.14
0.33
0.24
0.04
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
attention
source contexts
target hidden states
target word embeddings
target words
I
0.25

Our Work



source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

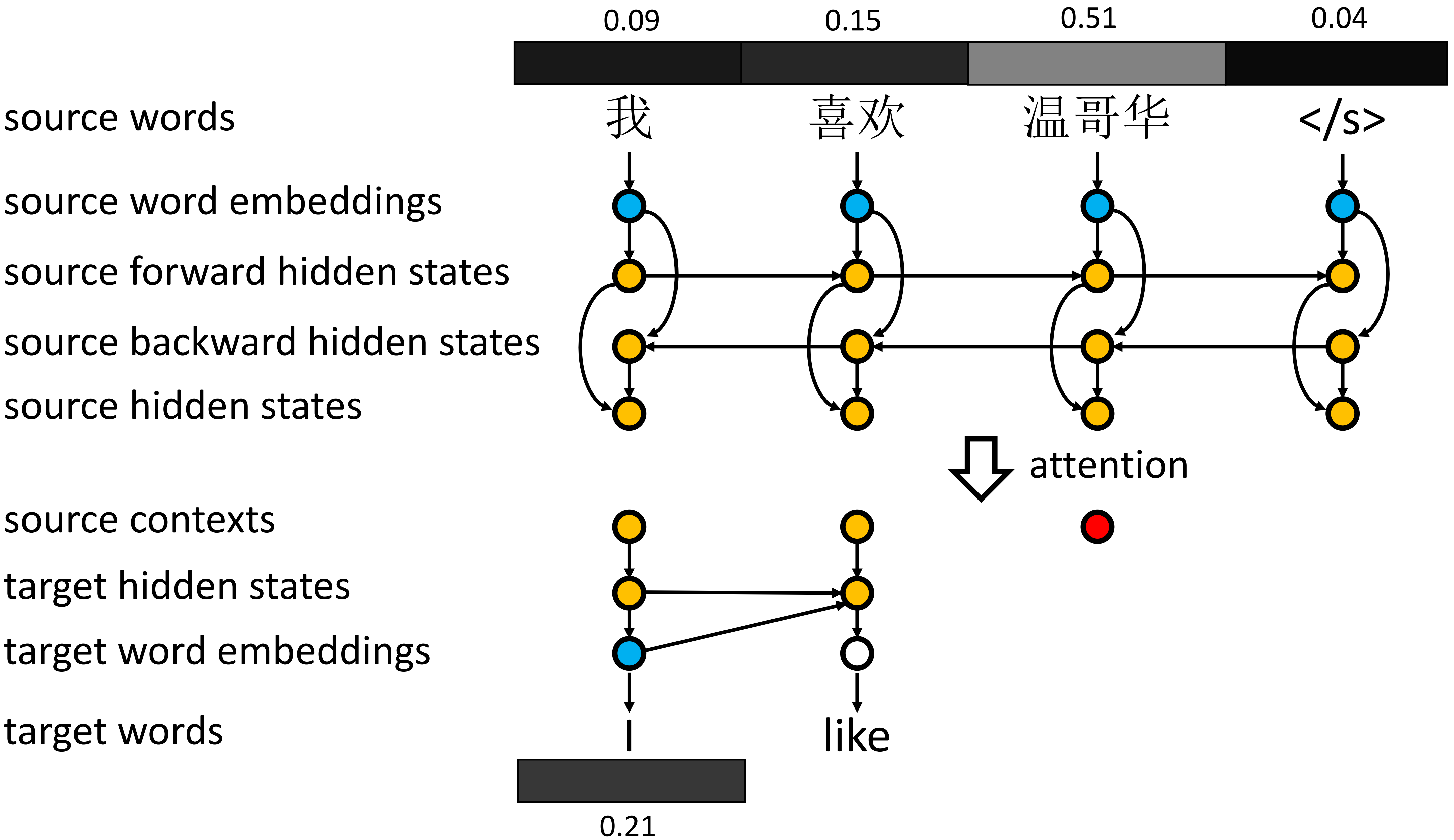
source contexts

target hidden states

target word embeddings

target words

Our Work



source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

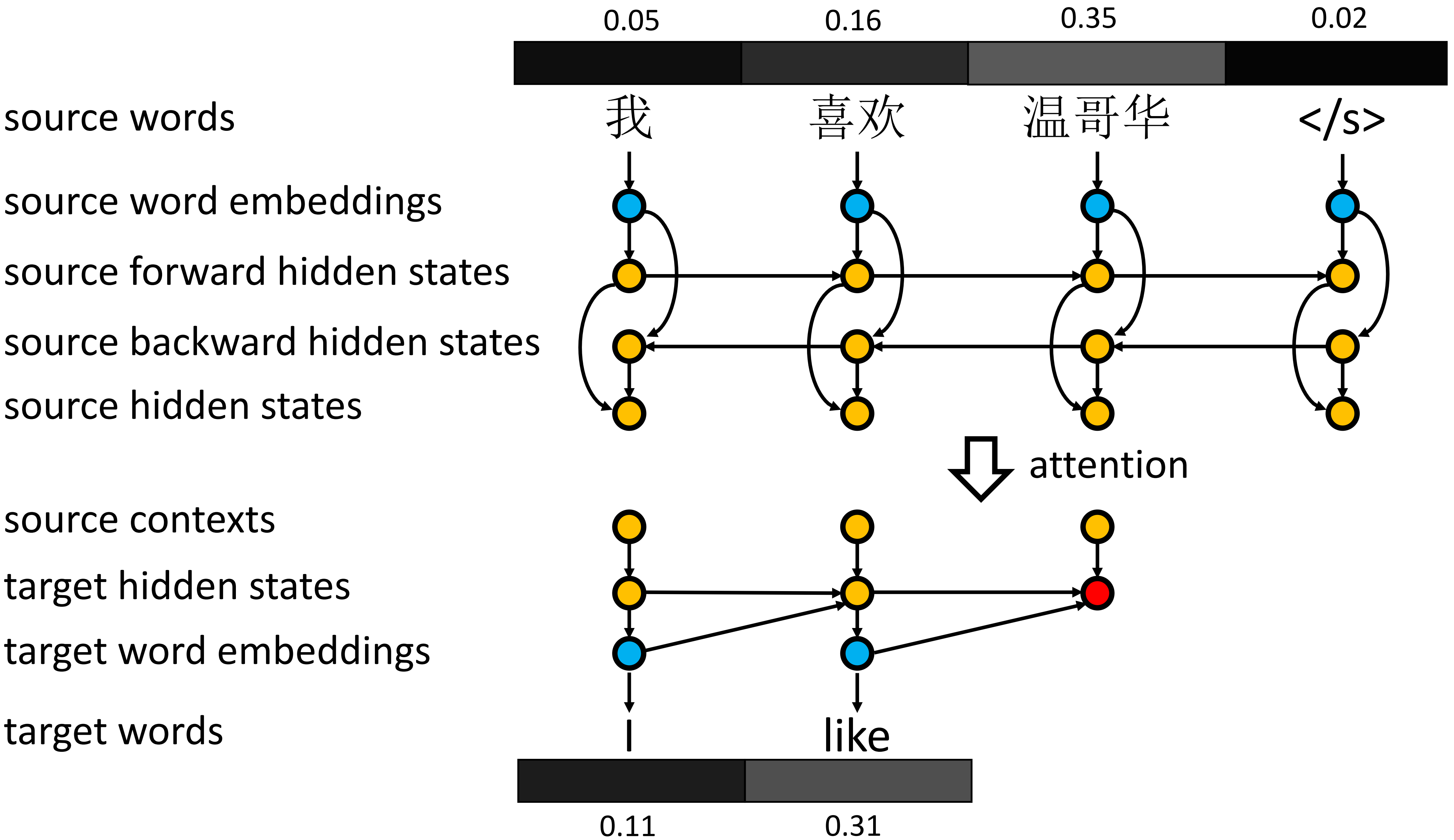
source contexts

target hidden states

target word embeddings

target words

Our Work



0.05 0.16 0.35 0.02

source words 我 喜欢 温哥华 </s>

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

attention

source contexts

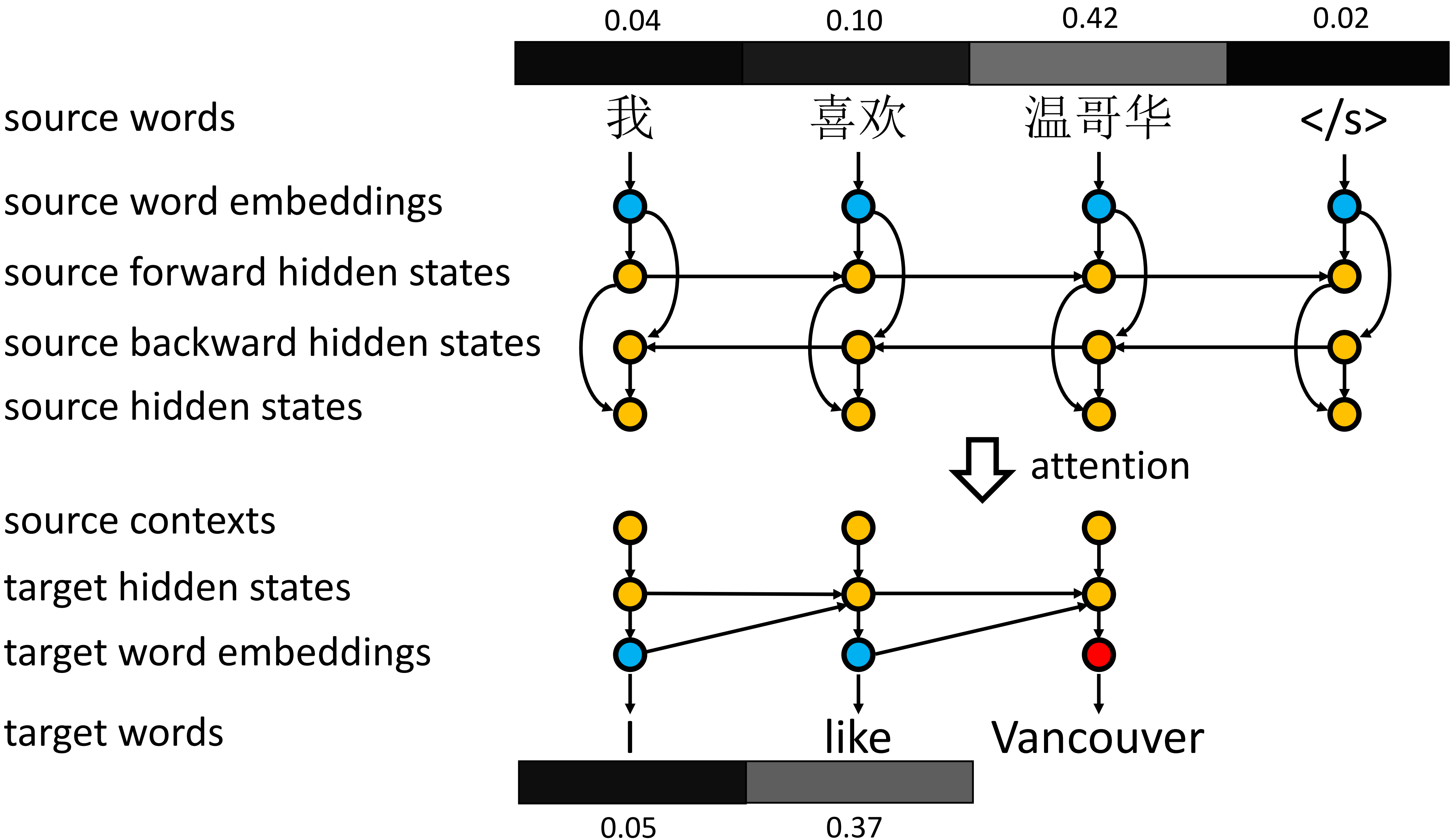
target hidden states

target word embeddings

target words I like

0.11 0.31

Our Work



0.04 0.10 0.42 0.02

source words 我 喜欢 温哥华 </s>

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

attention

source contexts

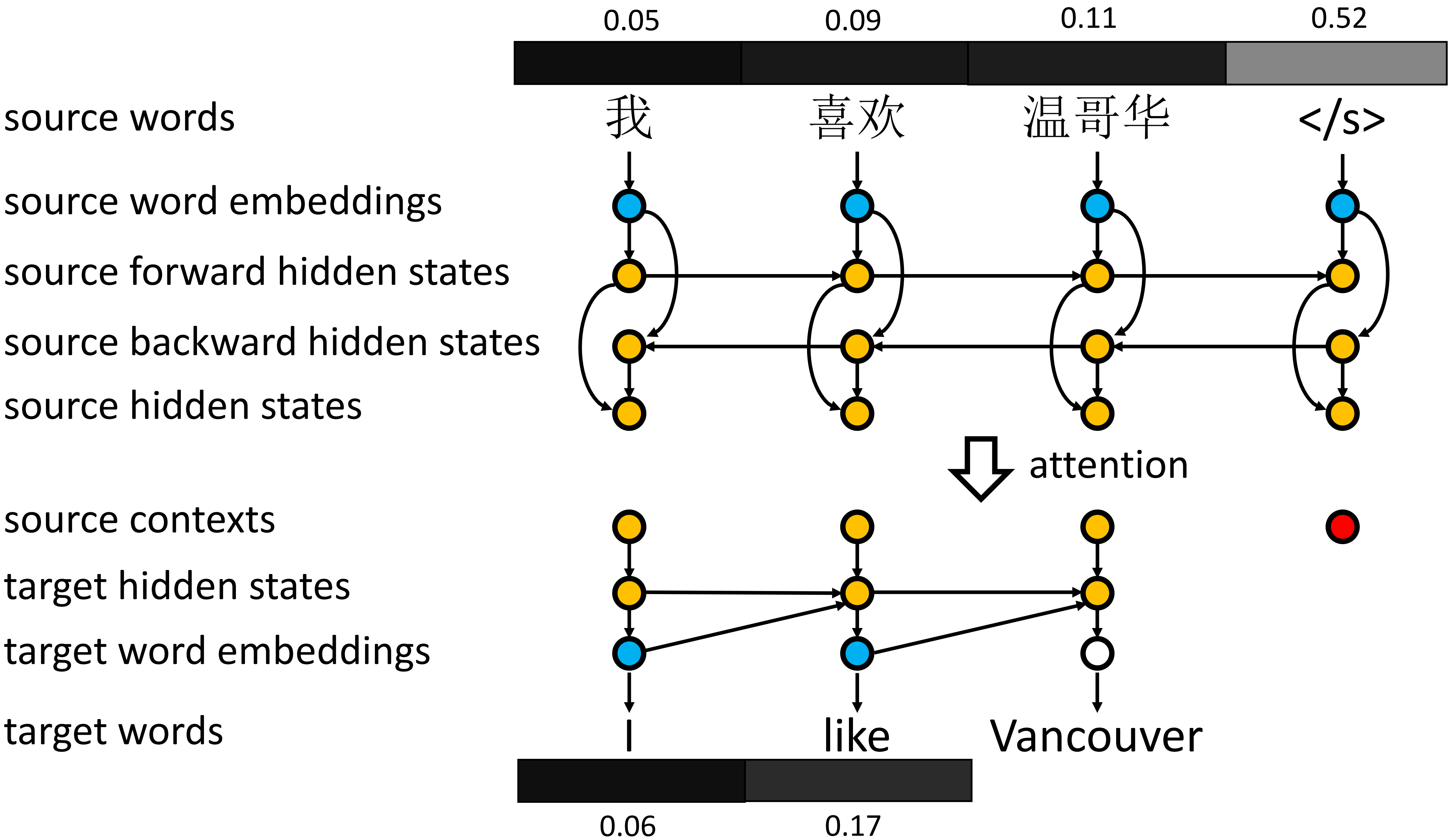
target hidden states

target word embeddings

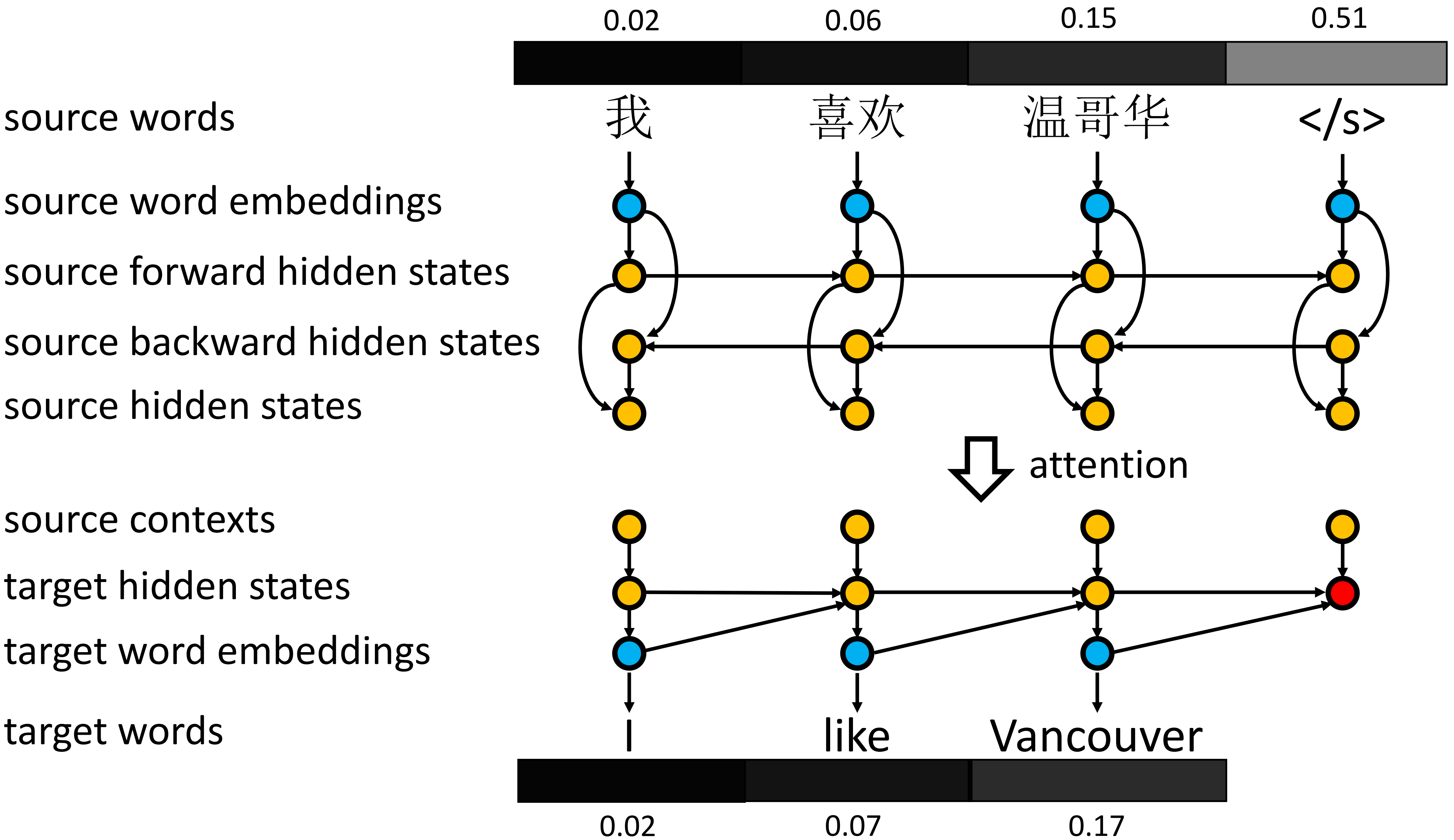
target words I like Vancouver

0.05 0.37

Our Work

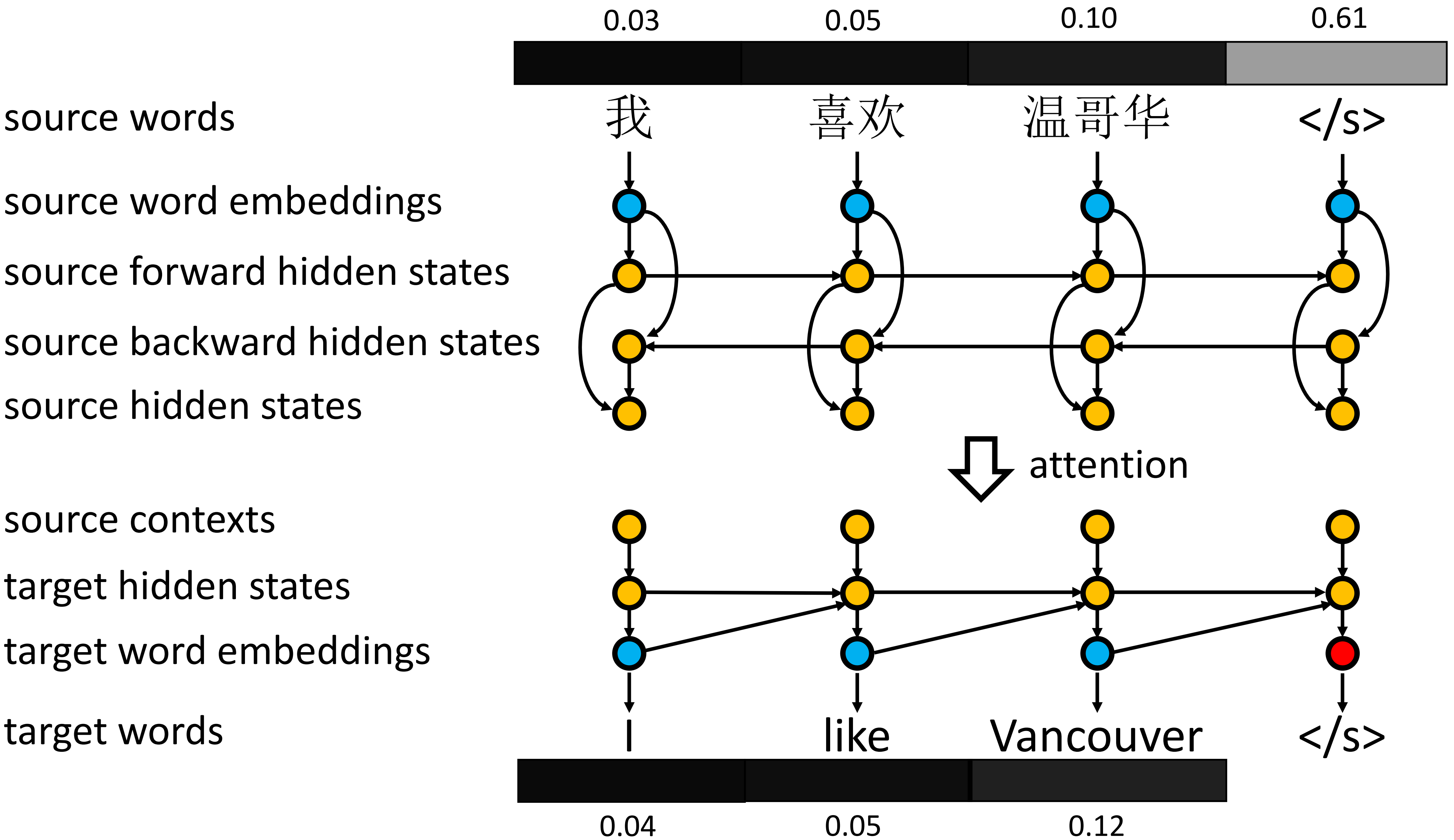

0.05
0.09
0.11
0.52
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
attention
source contexts
target hidden states
target word embeddings
target words
I
like
Vancouver
0.06
0.17

Our Work



0.02
0.06
0.15
0.51
source words
我
喜欢
温哥华
</s>
source word embeddings
source forward hidden states
source backward hidden states
source hidden states
attention
source contexts
target hidden states
target word embeddings
target words
I
like
Vancouver
0.02
0.07
0.17

Our Work



source words

source word embeddings

source forward hidden states

source backward hidden states

source hidden states

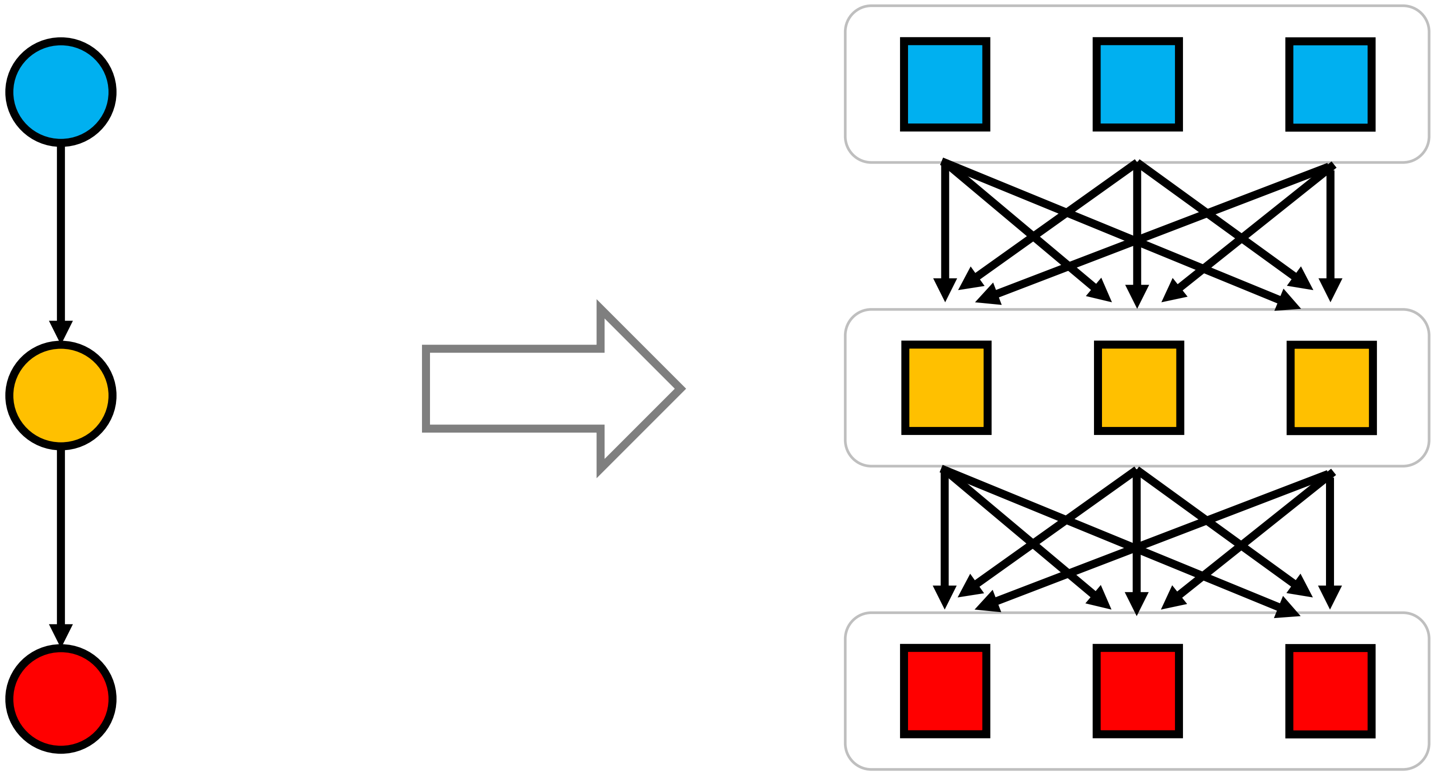
source contexts

target hidden states

target word embeddings

target words

Vector- and Neuron-Level Relevance

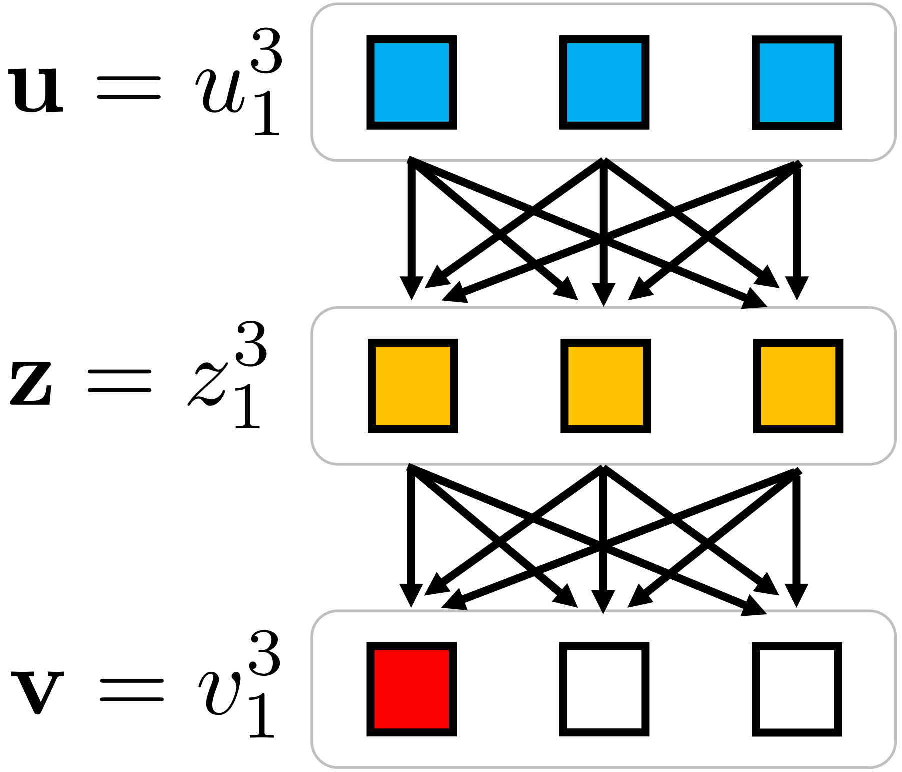


vector-level relevance

neural-level relevance

Neuron-Level Relevance

- Idea: decompose the activation of the targeted neuron among relevant neurons



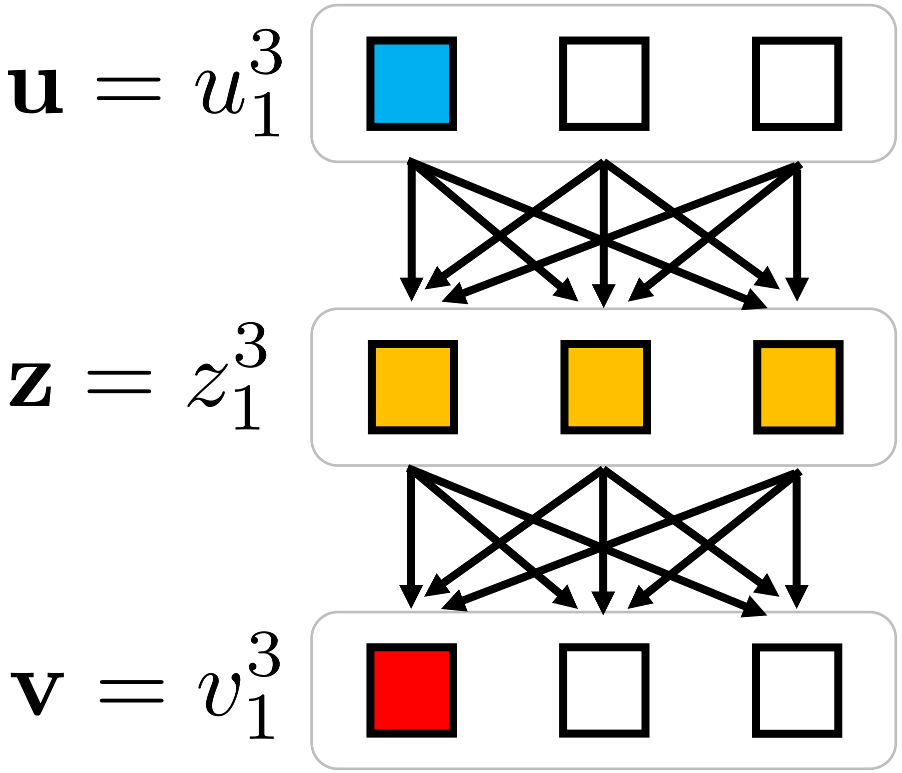
$$v_m = \sum_{\mathbf{u} \in \mathcal{C}(v_m)} \sum_{n=1}^{N} r_{u_n \leftarrow v_m}$$

For example

$$v_1 = \sum_{n=1}^{3} r_{u_n \leftarrow v_1}$$

Calculating Neuron-Level Relevance

- Recursive calculation in a backward propagation



$$\boxed{r_{u \leftarrow v} = \sum_{z \in \mathrm{OUT}(u)} w_{u \rightarrow z} r_{z \leftarrow v}}$$

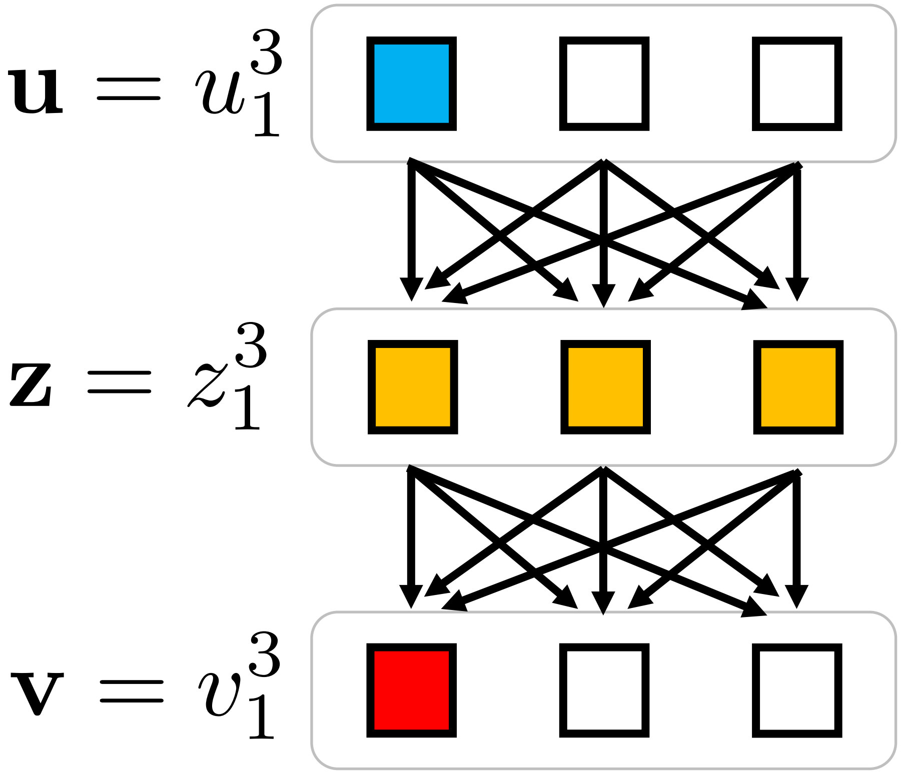
For example

$$r_{u_1 \leftarrow v_1} = \sum_{k=1}^{3} w_{u_1 \rightarrow z_k} r_{z_k \leftarrow v_1}$$

$$r_{z_k \leftarrow w_1} = w_{z_k \rightarrow v_1} v_1$$

Calculating Weight Ratios

- Recursive calculation in a forward propagation

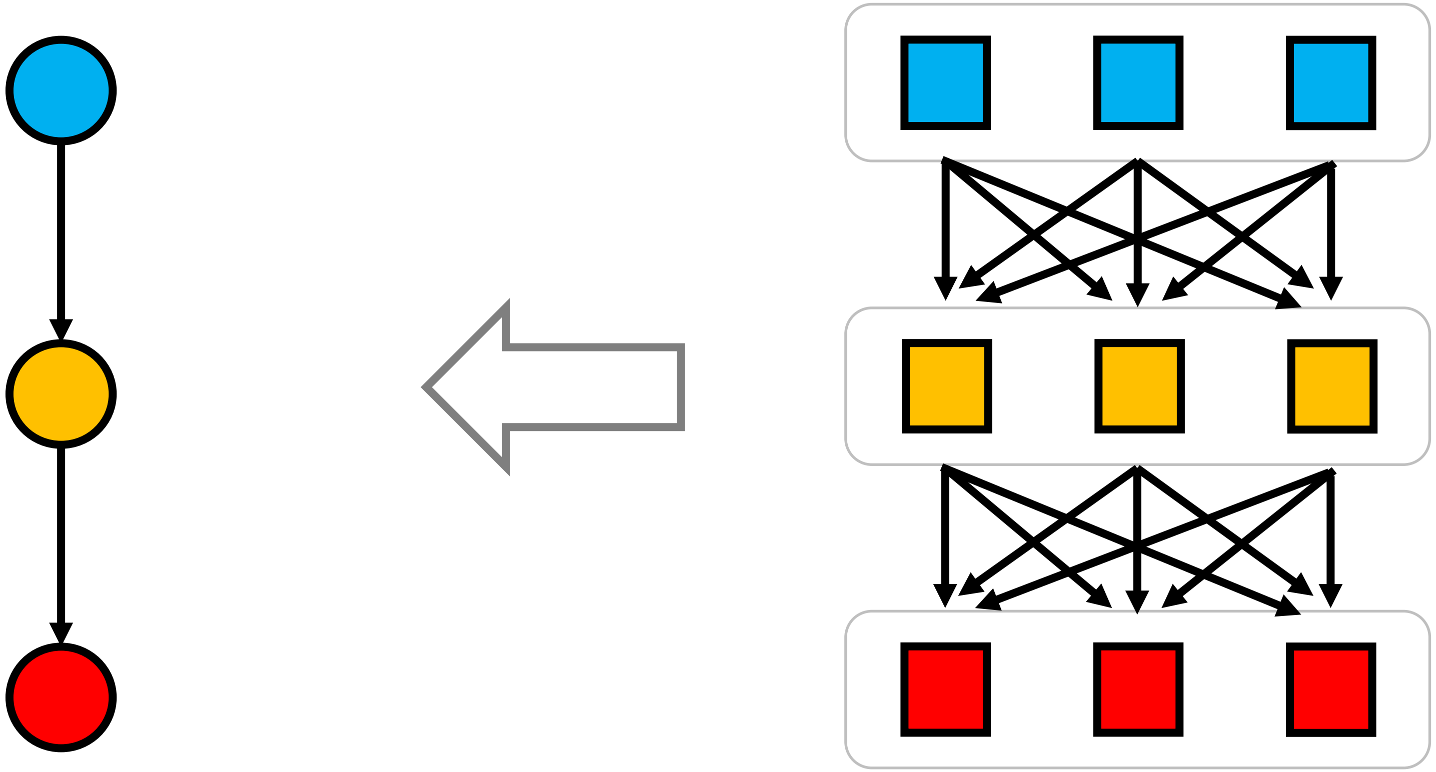


$$\boxed{w_{u \to v} = \frac{\mathbf{W}_{u,v} u}{\sum_{u' \in \mathrm{IN}(v)} \mathbf{W}_{u',v} u'}}$$

For example

$$w_{u_1 \to z_1} = \frac{\mathbf{W}^{(1)}_{1,1} u_1}{\mathbf{W}^{(1)}_{1,1} u_1 + \mathbf{W}^{(1)}_{2,1} u_2 + \mathbf{W}^{(1)}_{3,1} u_3}$$

Calculating Vector-Level Relevance



$$R_{\mathbf{u}\leftarrow\mathbf{v}} = \sum_{m=1}^{M}\sum_{n=1}^{N} r_{u_n \leftarrow v_m}$$

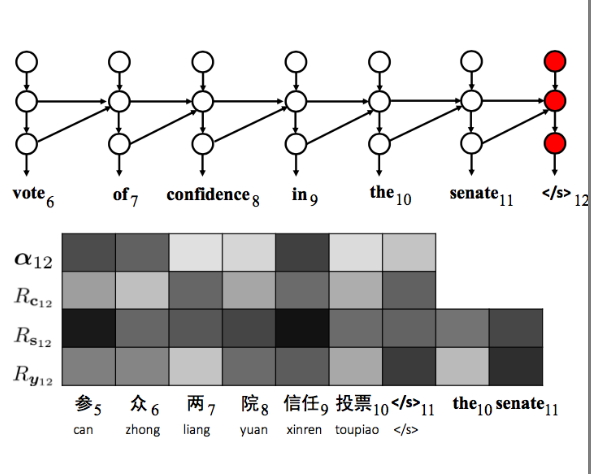
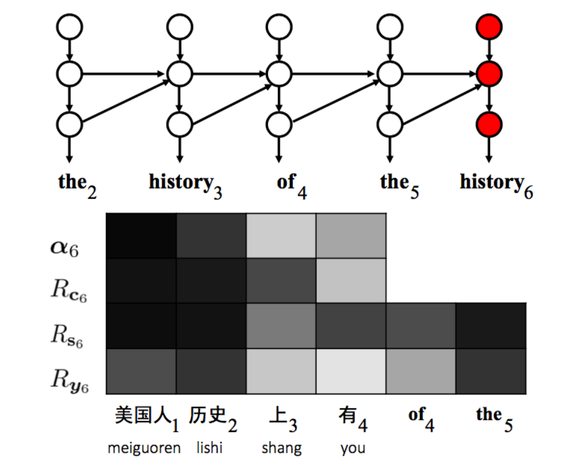
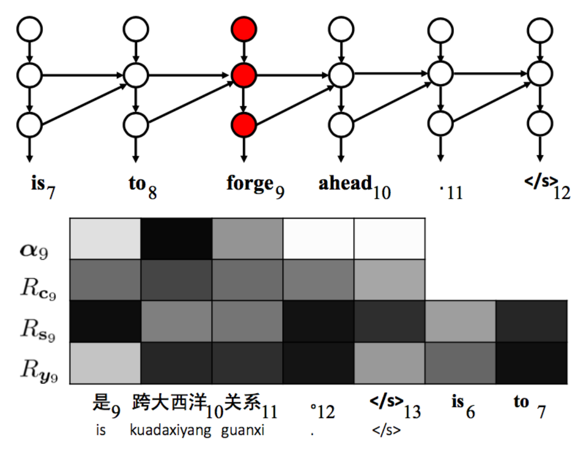
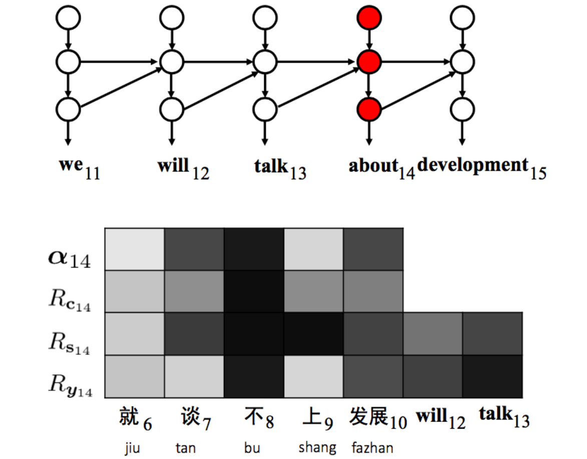
Algorithm

- Specify targeted vector of neurons
- Calculate weight ratios in a forward propagation
- Calculate relevance in a backward propagation

layer-wise propagation for neural machine translation

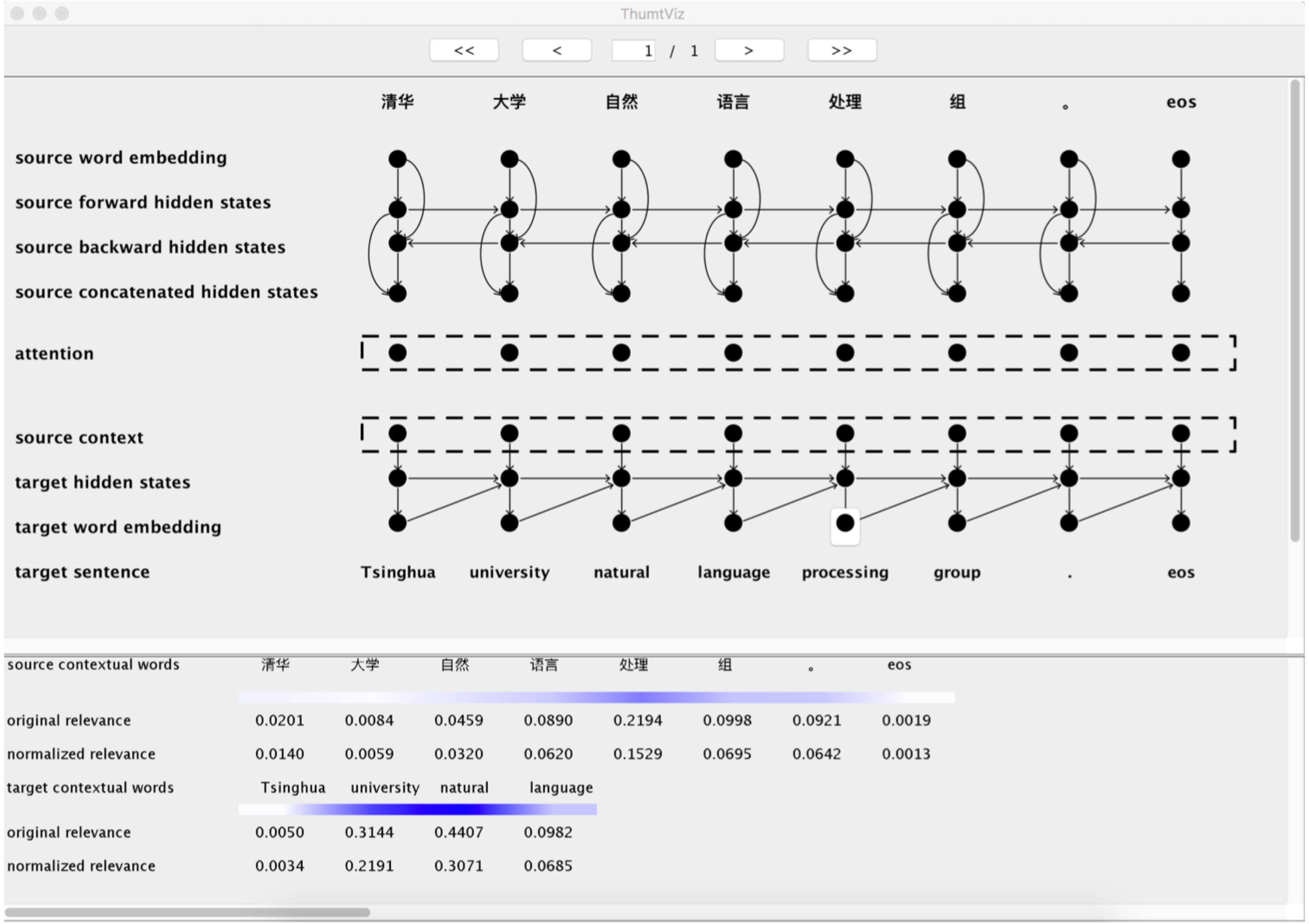
Application

- Help to debug attention-based NMT systems

word omission	word repetition	unrelated words	negation reversion
 vote$_6$ of$_7$ confidence$_8$ in$_9$ the$_{10}$ senate$_{11}$ </s>$_{12}$ α_{12} $R_{c_{12}}$ $R_{s_{12}}$ $R_{y_{12}}$ 参$_5$ can 众$_6$ zhong 两$_7$ liang 院$_8$ yuan 信任$_9$ xinren 投票$_{10}$ toupiao </s>$_{11}$ </s> the$_{10}$ senate$_{11}$	 the$_2$ history$_3$ of$_4$ the$_5$ history$_6$ α_6 R_{c_6} R_{s_6} R_{y_6} 美国人$_1$ meiguoren 历史$_2$ lishi 上$_3$ shang 有$_4$ you of$_4$ the$_5$	 is$_7$ to$_8$ forge$_9$ ahead$_{10}$.$_{11}$ </s>$_{12}$ α_9 R_{c_9} R_{s_9} R_{y_9} 是$_9$ is 跨大西洋$_{10}$ kuadaxiyang 关系$_{11}$ guanxi 。$_{12}$. </s>$_{13}$ </s> is$_6$ to$_7$	 we$_{11}$ will$_{12}$ talk$_{13}$ about$_{14}$ development$_{15}$ α_{14} $R_{c_{14}}$ $R_{s_{14}}$ $R_{y_{14}}$ 就$_6$ jiu 谈$_7$ tan 不$_8$ bu 上$_9$ shang 发展$_{10}$ fazhan will$_{12}$ talk$_{13}$

analyzing major translation error types by visualizing relevance step by step

Open-Source Toolkit

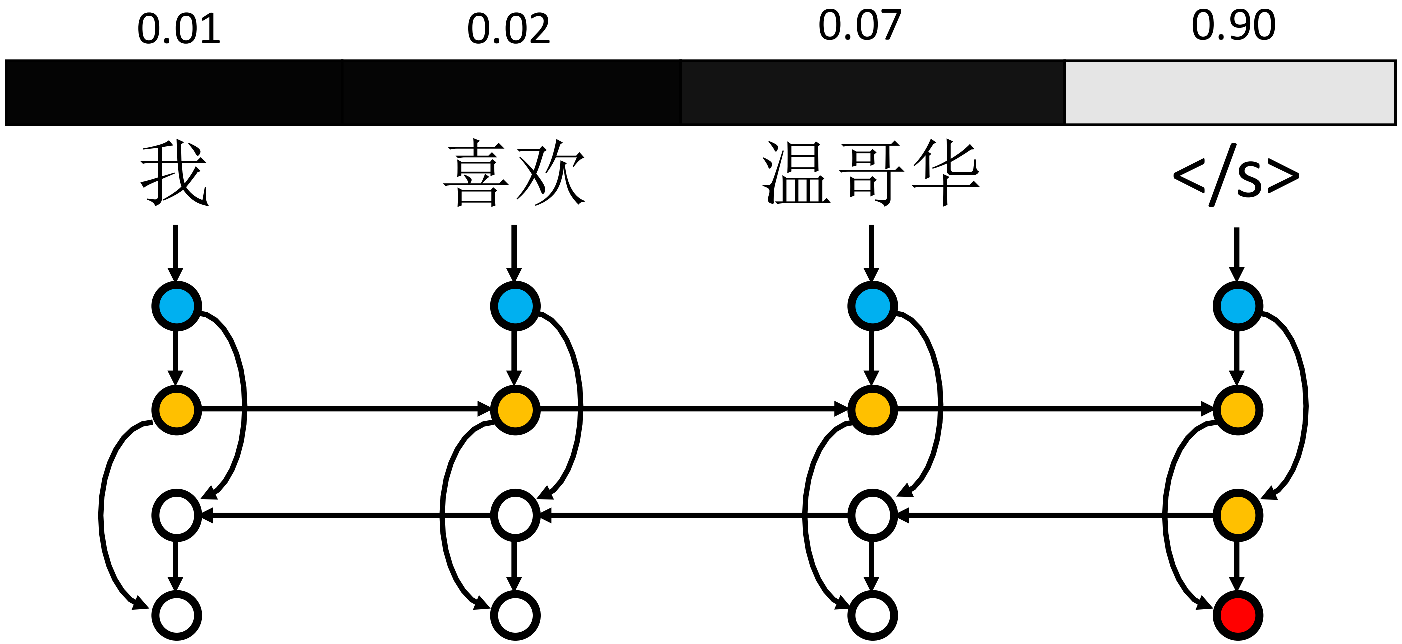


source contextual words	清华	大学	自然	语言	处理	组	。	eos
original relevance	0.0201	0.0084	0.0459	0.0890	0.2194	0.0998	0.0921	0.0019
normalized relevance	0.0140	0.0059	0.0320	0.0620	0.1529	0.0695	0.0642	0.0013

target contextual words	Tsinghua	university	natural	language
original relevance	0.0050	0.3144	0.4407	0.0982
normalized relevance	0.0034	0.2191	0.3071	0.0685

http://thumt.thunlp.org

Conclusion



- It is challenging to interpret how neural networks work
- We leverage layer-wise relevance propagation to visualize NMT
- Our approach can be applied to networks in other NLP tasks

Thanks

http://thumt.thunlp.org