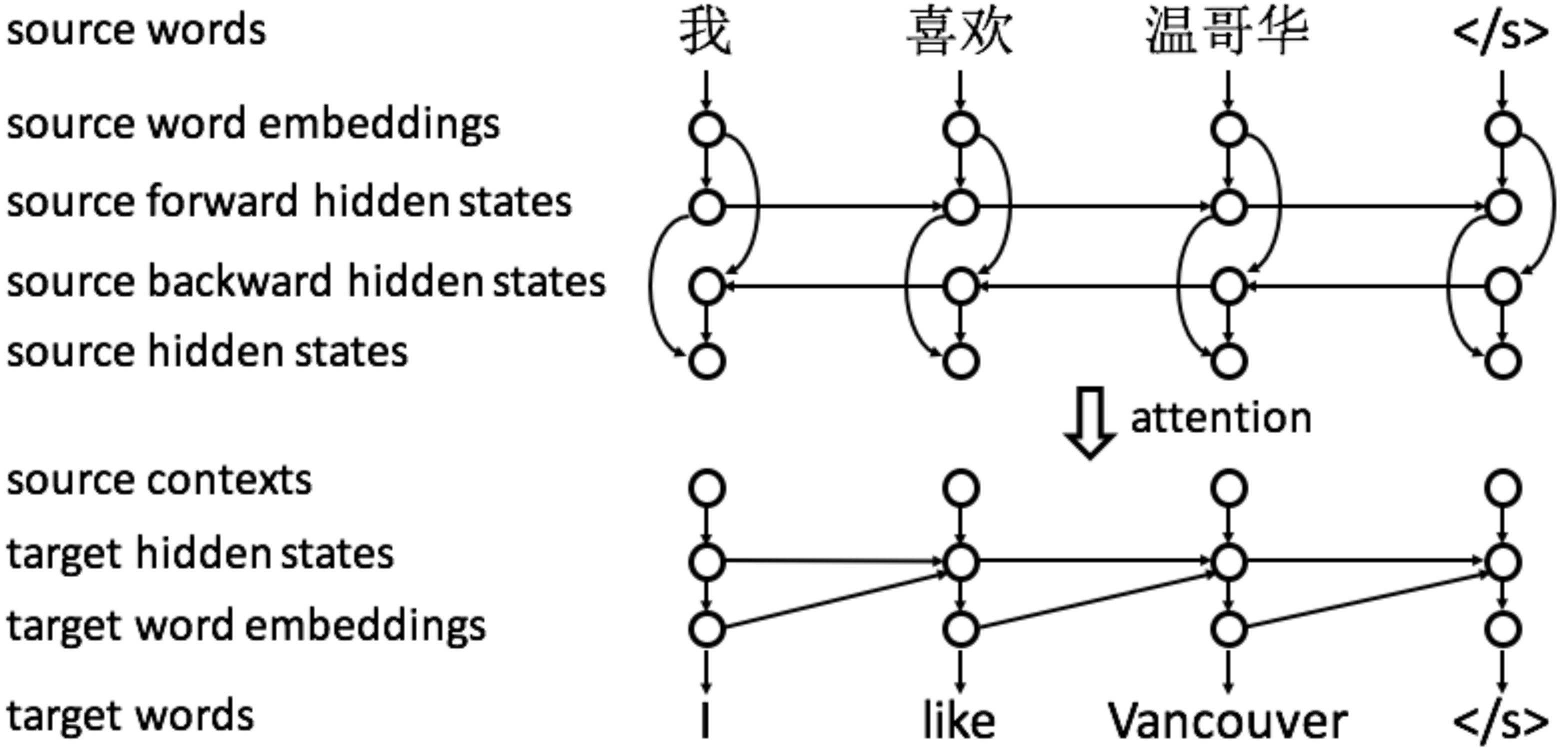
Visualizing and Understanding Neural Machine Translation

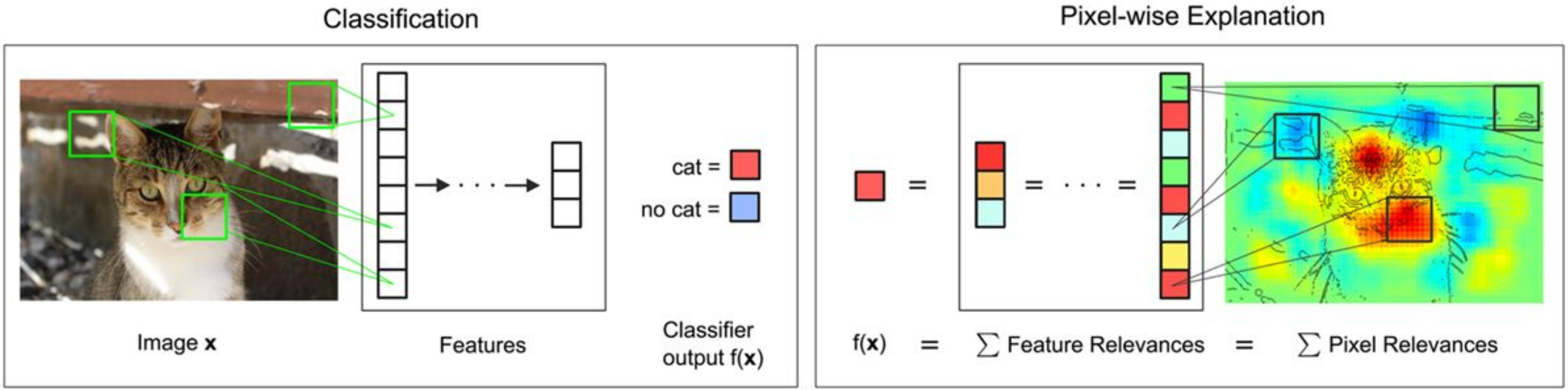
Yanzhuo Ding, Yang Liu, Huanbo Luan, Maosong Sun

Tsinghua University

Presented by: Yuchen He


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>

Layer-wise relevance propagation (LRP)


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

Can calculate the relevance between two arbitrary neurons

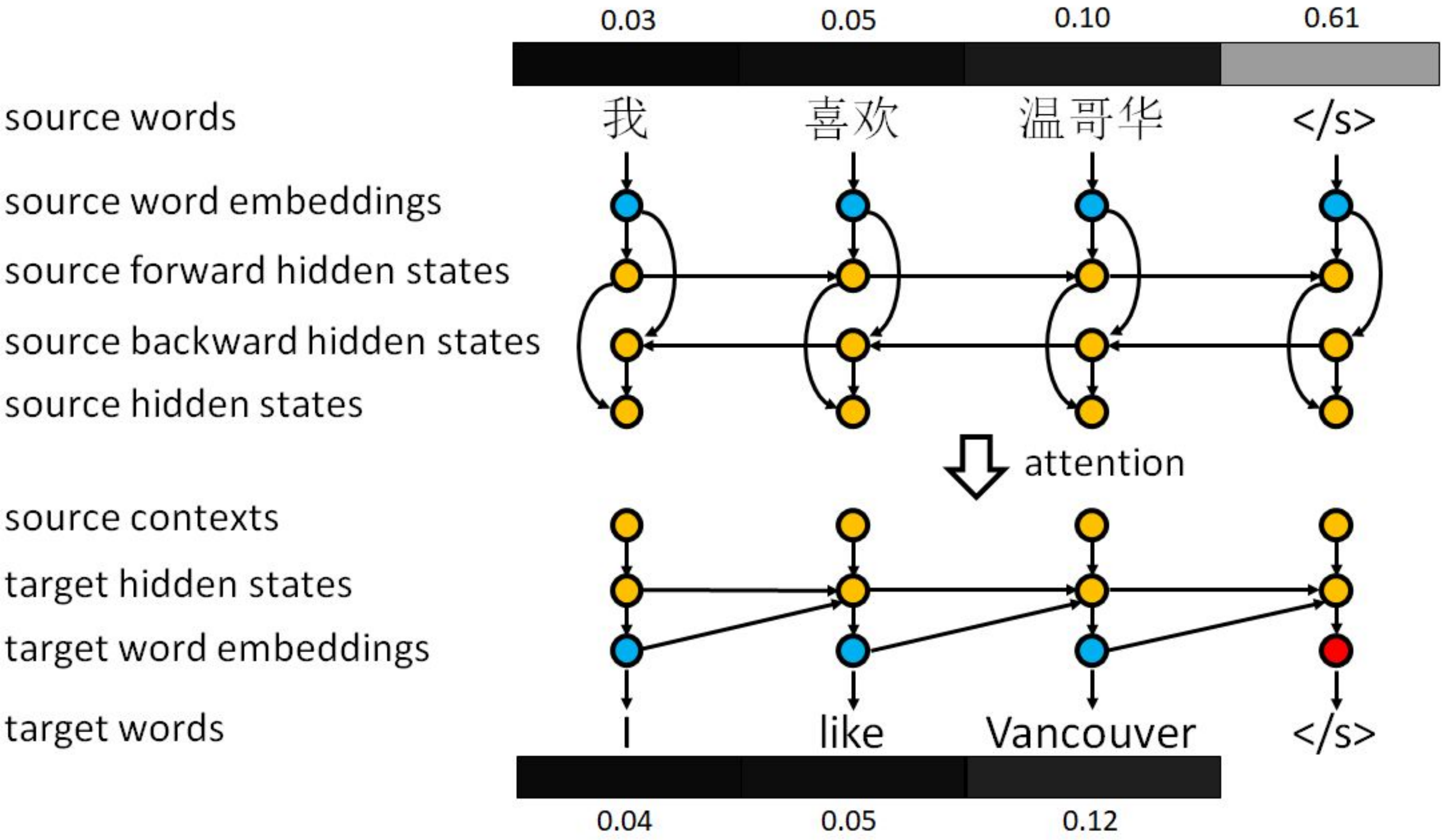
Measures/visualizes how much each pixel is related to the final classification

Goal

- To quantify and visualize the relevance between a neural network layer and contextual word vectors(source & target word embeddings)

Offers more insights in interpreting how target words are generated

Relevance vector

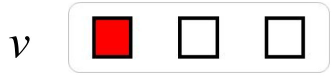


Calculating Neuron-Level Relevance

Base case: (relevance of v to itself)

$$r_{v \leftarrow v} = v$$

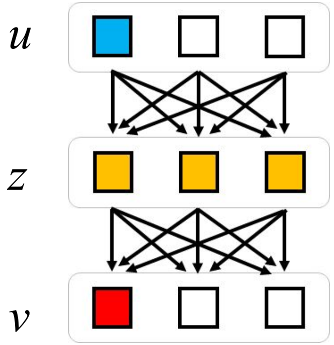
for any neuron v



Recursive case: (relevance of u to v)

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

for any neurons u, v

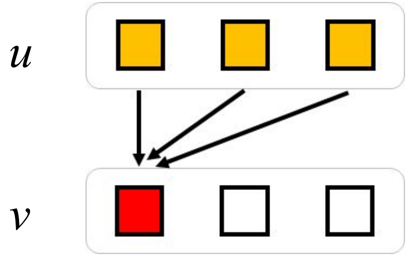


$\mathrm{OUT}(u)$ comprises all u's ***directly connected descendant*** neurons in the network.

Calculating Weight Ratios

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

for any neurons *u, v*



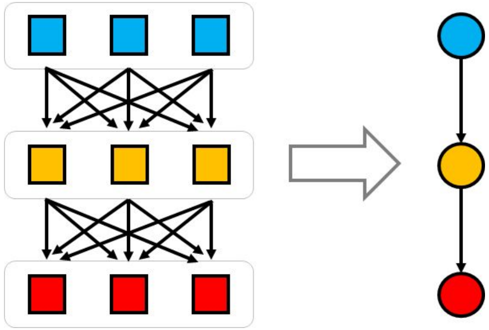
$\mathbf{W}_{u,v} u$ is the weight of *u* to *v* in the existing neural network

$\mathrm{IN}(u)$ comprises all *u*'s ***directly connected ancestor*** neurons in the network.

Putting things together

Sum up $r_{u_n \leftarrow v_m}$ and get vector-level relevance $R_{\mathbf{u} \leftarrow \mathbf{v}}$

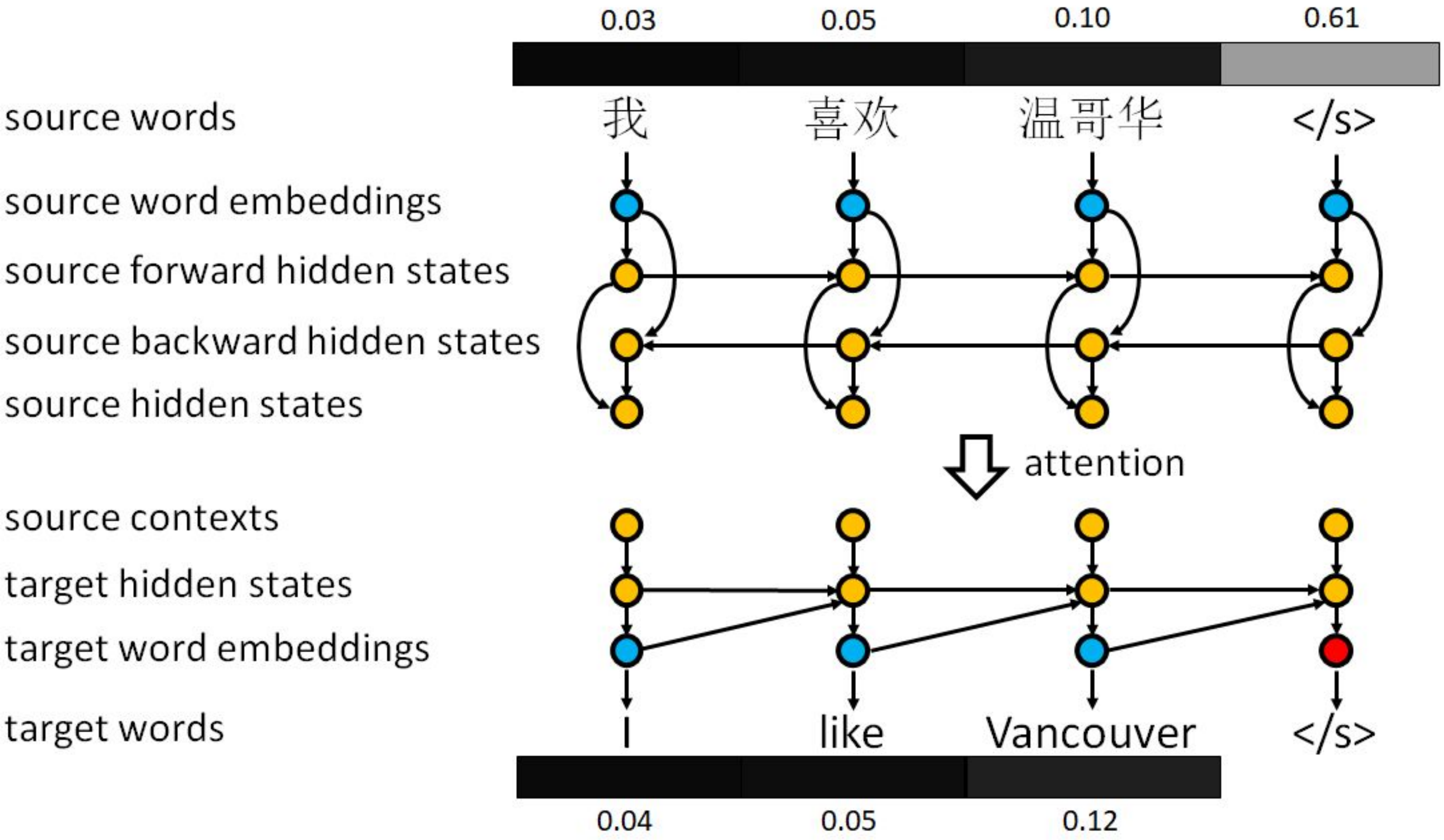
$$R_{\mathbf{u} \leftarrow \mathbf{v}} = \sum_{m=1}^{M} \sum_{n=1}^{N} r_{u_n \leftarrow v_m}$$



Generate and normalize relevance vector $R_{\mathbf{v}}$ as a sequence of $R_{\mathbf{u} \leftarrow \mathbf{v}}$ for all related contextual word vectors

$$R_{\mathbf{v}} = \{R_{\mathbf{u}_1 \leftarrow \mathbf{v}}, \ldots, R_{\mathbf{u}_{|C(\mathbf{v})|} \leftarrow \mathbf{v}}\}$$

Relevance vector

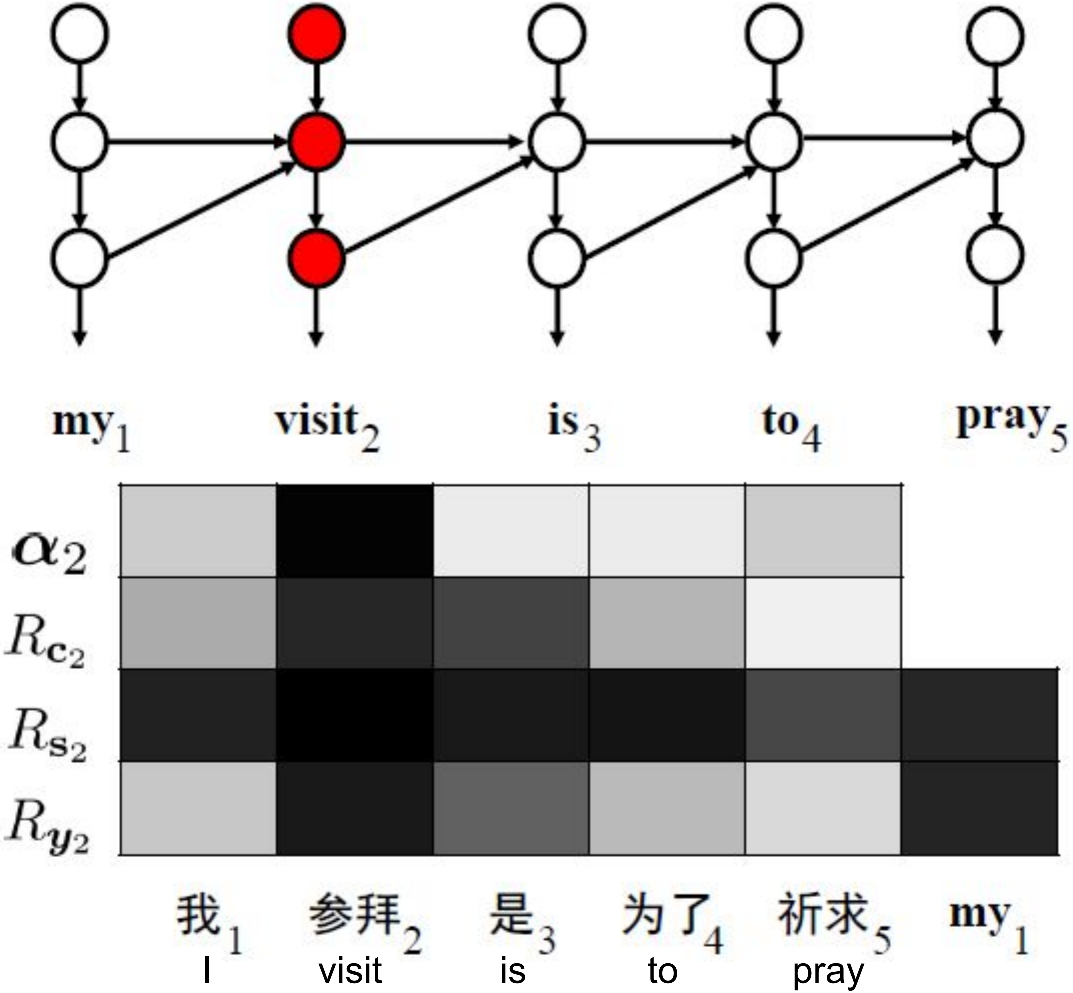


Application

Help debug attention-based NMT systems

- Word omission
- Word repetition
- Unrelated words
- Negation reversion

“Relevance matrix”



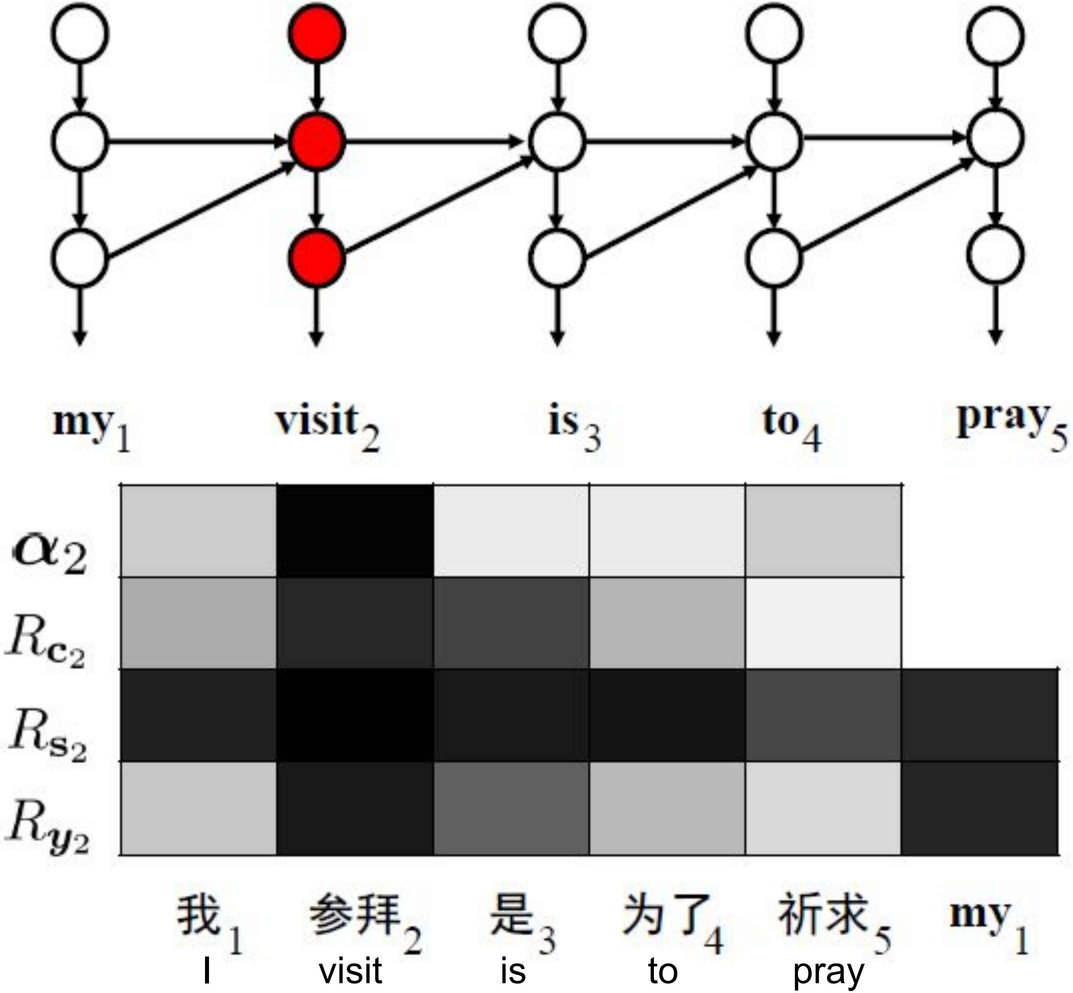
attention weights

source context vector

target hidden state

target word embedding

“Relevance matrix”



attention weights

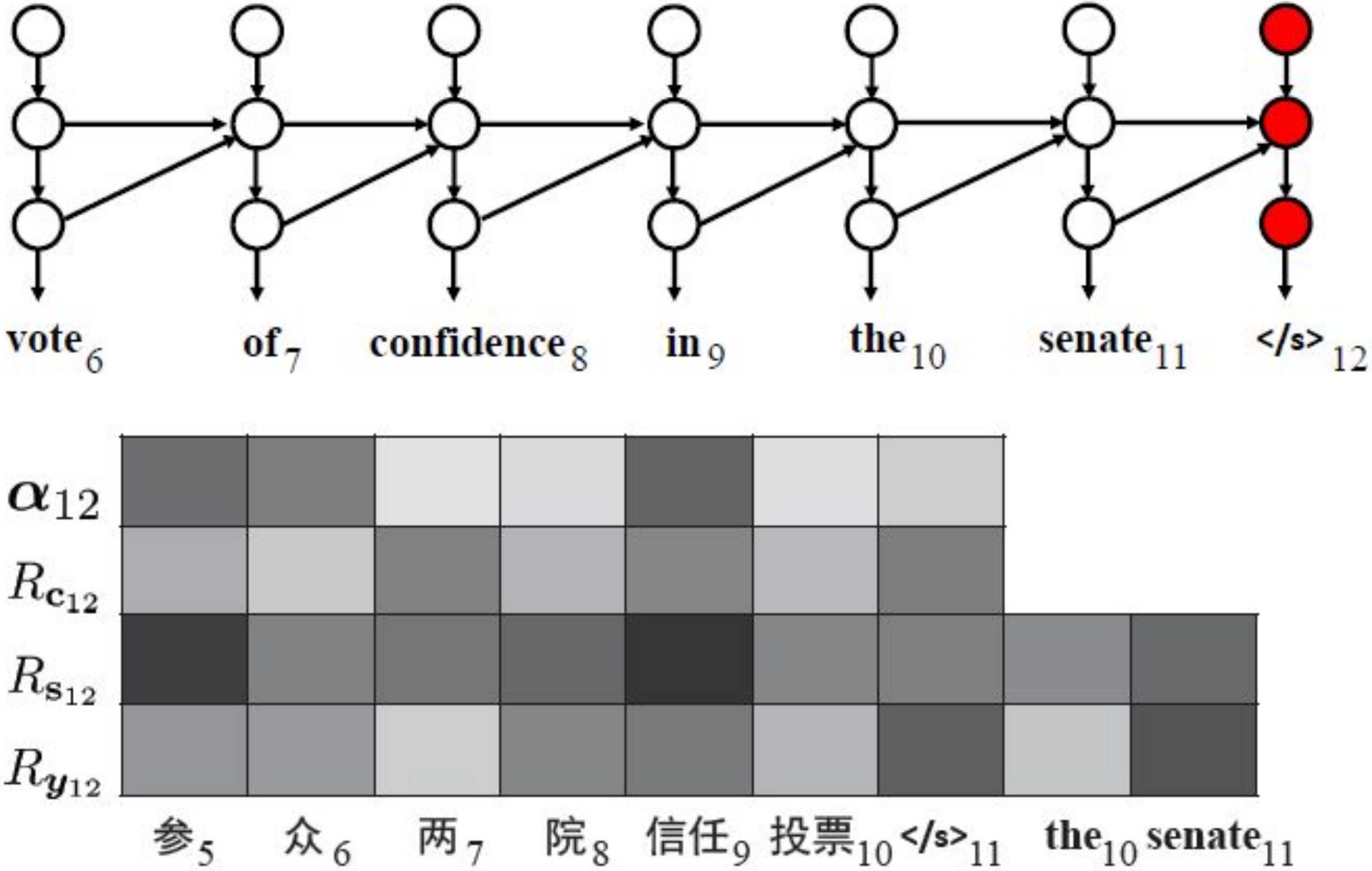
source context vector

target hidden state

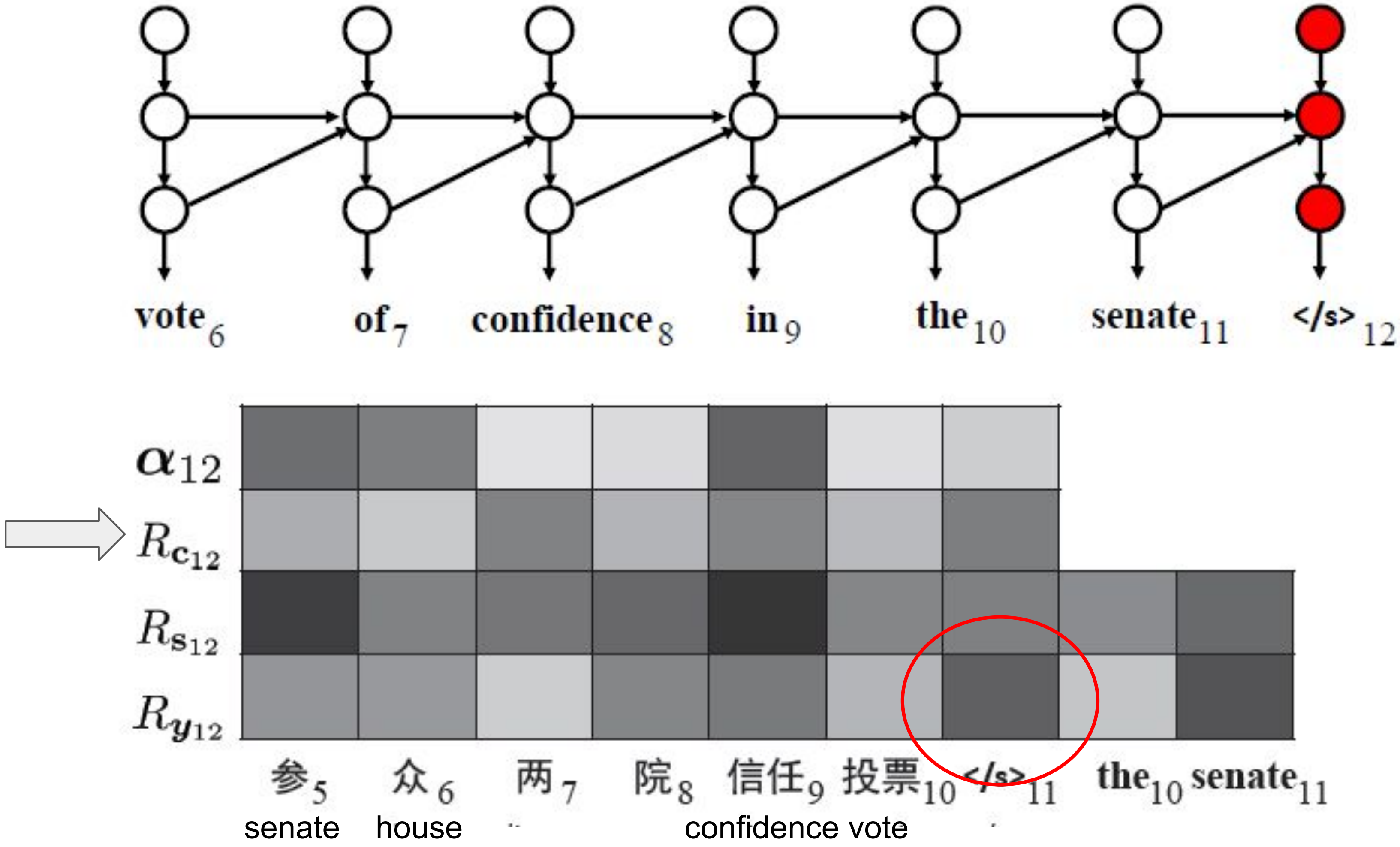
target word embedding

Word omission

Input	巴基斯坦总统穆沙拉夫赢得参众两院信任投票
Reference	Pakistani president Musharraf wins votes of confidence in senate and house
Output	Pakistani president win over democratic vote of confidence in senate (missing words)

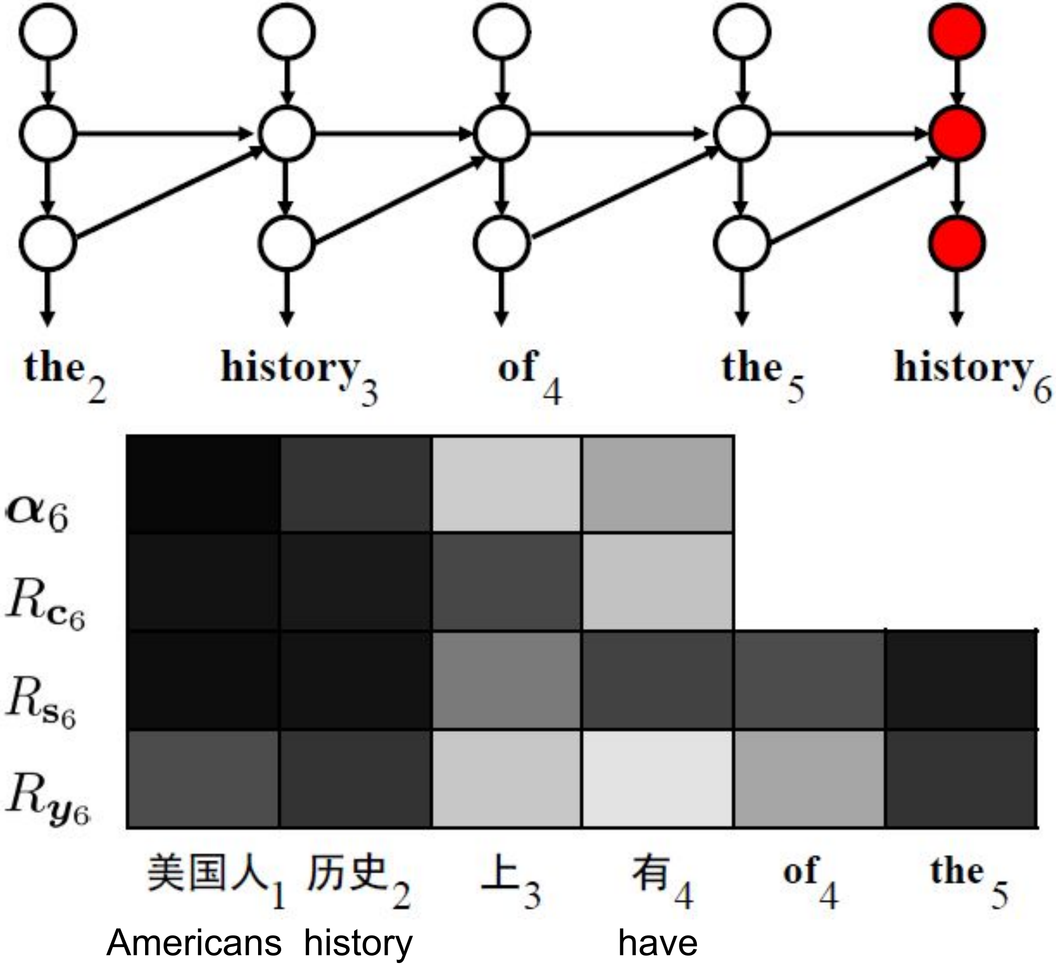


senate house confidence vote

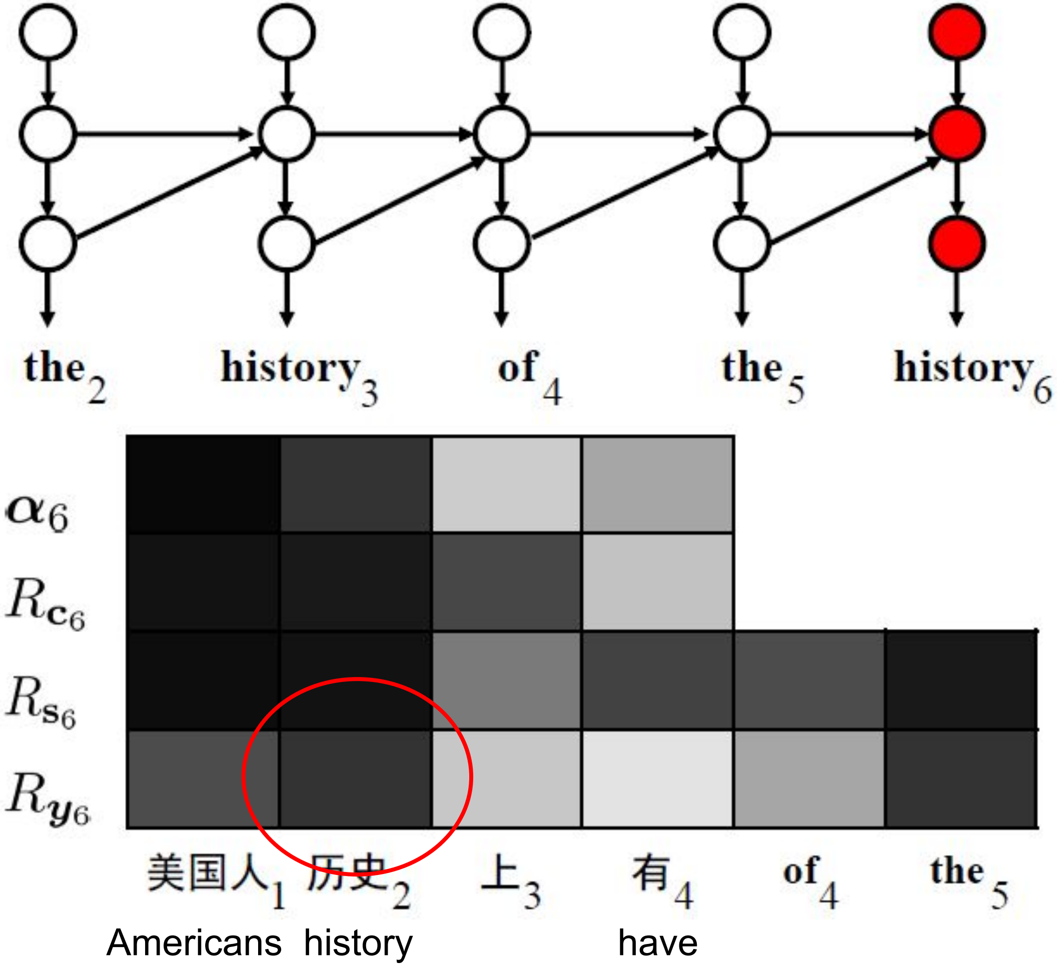

vote$_6$
of$_7$
confidence$_8$
in$_9$
the$_{10}$
senate$_{11}$
</s>$_{12}$
α_{12}
$R_{c_{12}}$
$R_{s_{12}}$
$R_{y_{12}}$
参$_5$
众$_6$
两$_7$
院$_8$
信任$_9$
投票$_{10}$
</s>$_{11}$
the$_{10}$
senate$_{11}$
senate
house
confidence vote

Word repetition

Input	美国人历史上有讲诚信的传统，有犯错认错的传统
Reference	In history, Americans have the tradition of honesty and would not hesitate to admit their mistakes
Output	In the history of the history of the history of the Americans, there is a tradition of faith in the history of mistakes

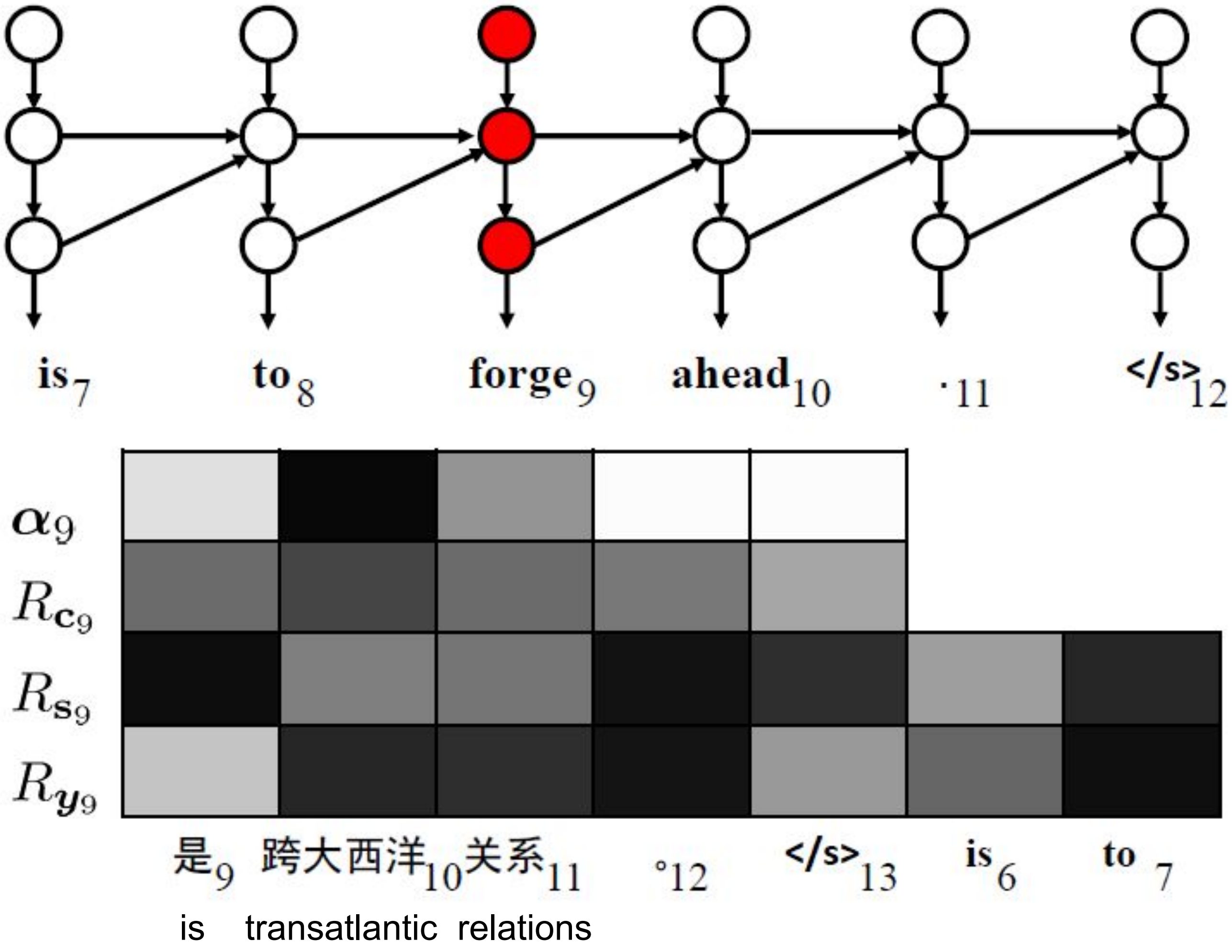


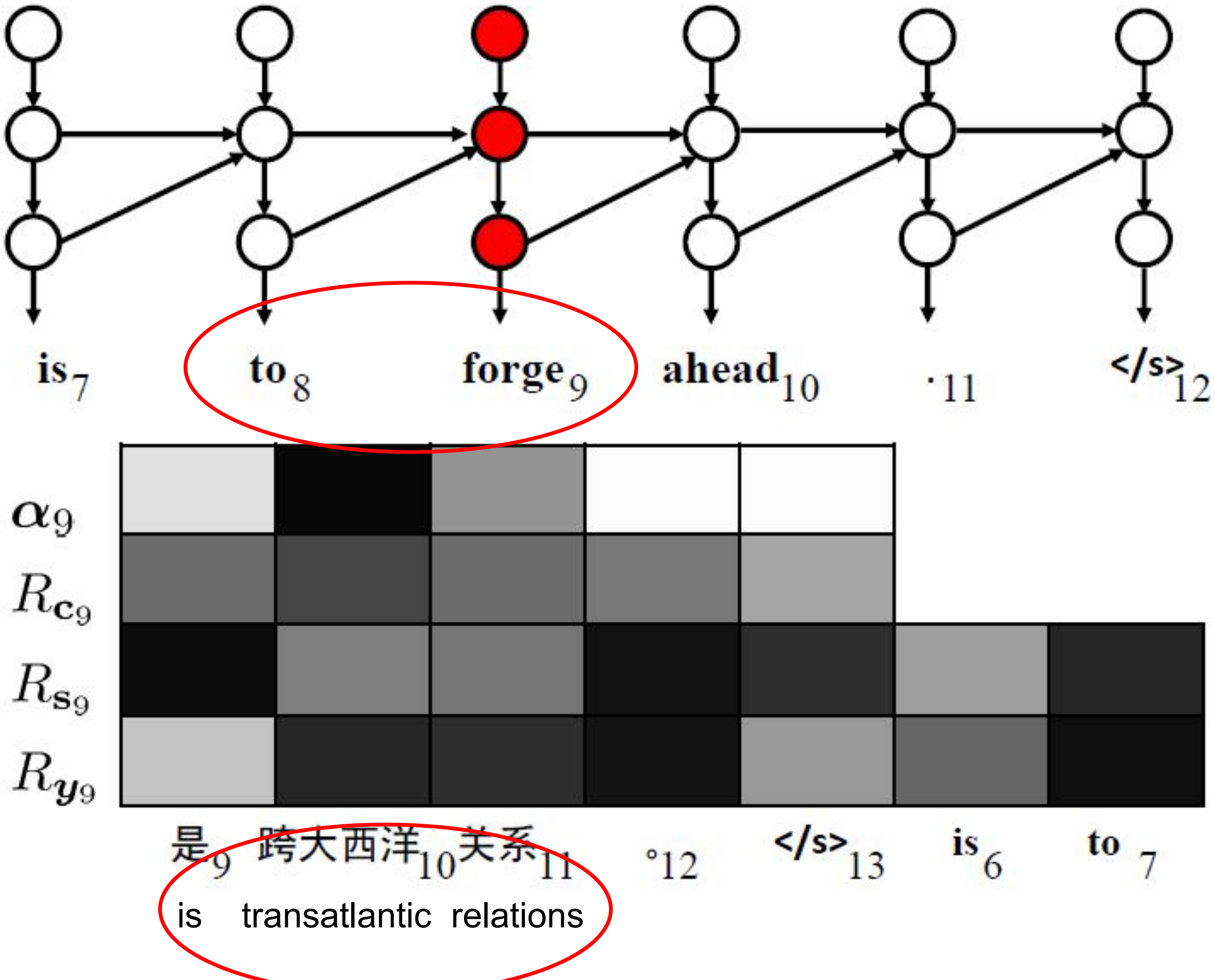
the_2
history_3
of_4
the_5
history_6
$\boldsymbol{\alpha}_6$
$R_{\mathbf{c}_6}$
$R_{\mathbf{s}_6}$
$R_{\boldsymbol{y}_6}$
美国人$_1$
历史$_2$
上$_3$
有$_4$
of_4
the_5
Americans
history
have


the_2
history_3
of_4
the_5
history_6
$\boldsymbol{\alpha}_6$
$R_{\mathbf{c}_6}$
$R_{\mathbf{s}_6}$
$R_{\boldsymbol{y}_6}$
美国人$_1$
历史$_2$
上$_3$
有$_4$
of_4
the_5
Americans
history
have

Unrelated words

Input	此次会议的一个重要议题是跨大西洋关系
Reference	One of the top agendas of the meeting is to discuss the transatlantic relations
Output	A key topic of the meeting is to forge ahead

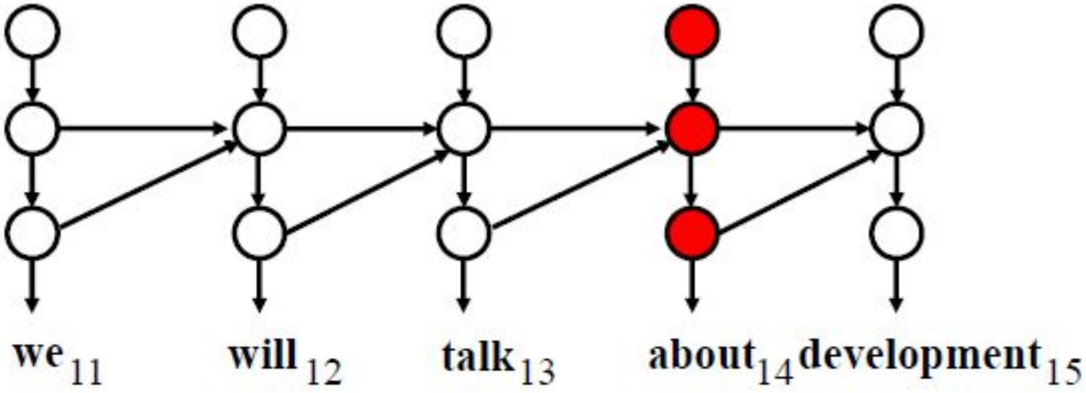

is$_7$
to$_8$
forge$_9$
ahead$_{10}$
·$_{11}$
</s>$_{12}$
$\boldsymbol{\alpha}_9$
$R_{\mathbf{c}_9}$
$R_{\mathbf{s}_9}$
$R_{\boldsymbol{y}_9}$
是$_9$
跨大西洋$_{10}$
关系$_{11}$
。$_{12}$
</s>$_{13}$
is$_6$
to$_7$
is
transatlantic
relations

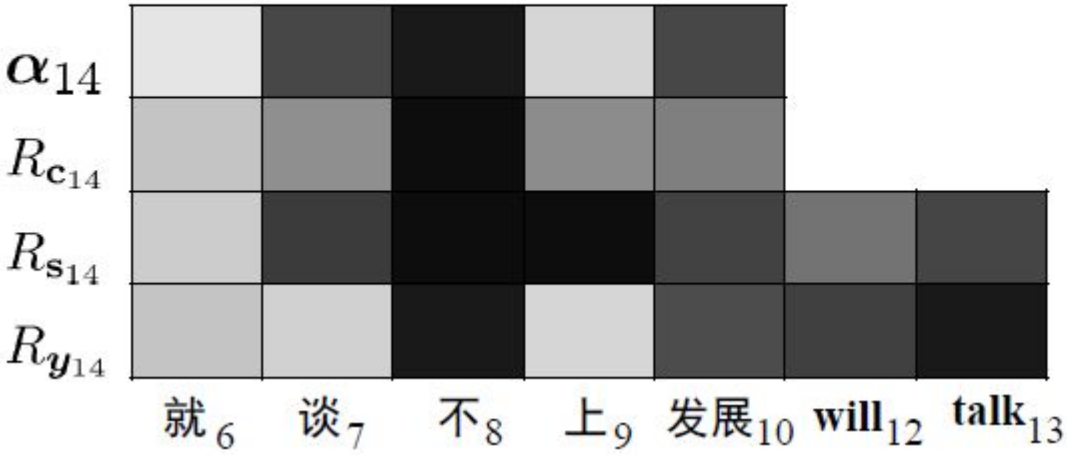


is transatlantic relations

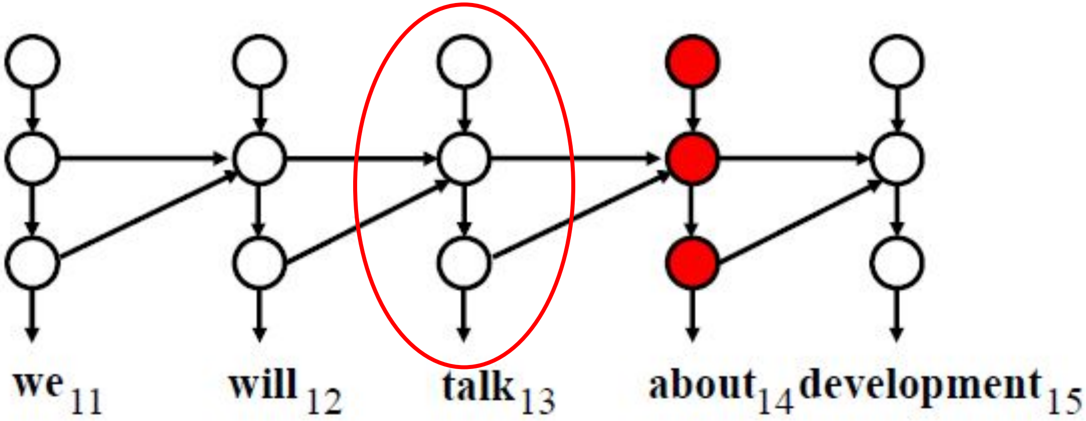
Negation reversion

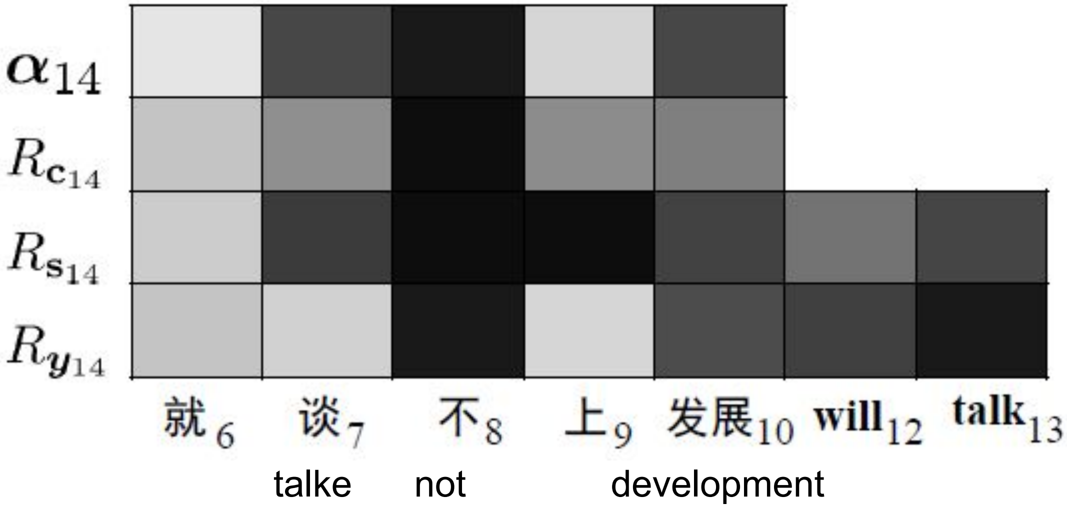
Input	不解决生存问题，就谈不上发展，更谈不上可持续发展
Reference	Without solving the issue of subsistence, there will be no development to speak of , let alone sustainable development
Output	If we do not solve the problem of living , we will talk about development and still less can we talk about sustainable development





talke not development


we_{11}
will_{12}
talk_{13}
about_{14}
development_{15}


α_{14}
$R_{\mathbf{c}_{14}}$
$R_{\mathbf{s}_{14}}$
$R_{\mathbf{y}_{14}}$
就$_6$
谈$_7$
不$_8$
上$_9$
发展$_{10}$
will_{12}
talk_{13}
talke
not
development

Thank you