TTIC 31210:

Advanced Natural Language Processing

Kevin Gimpel

Spring 2017

Lecture 7:

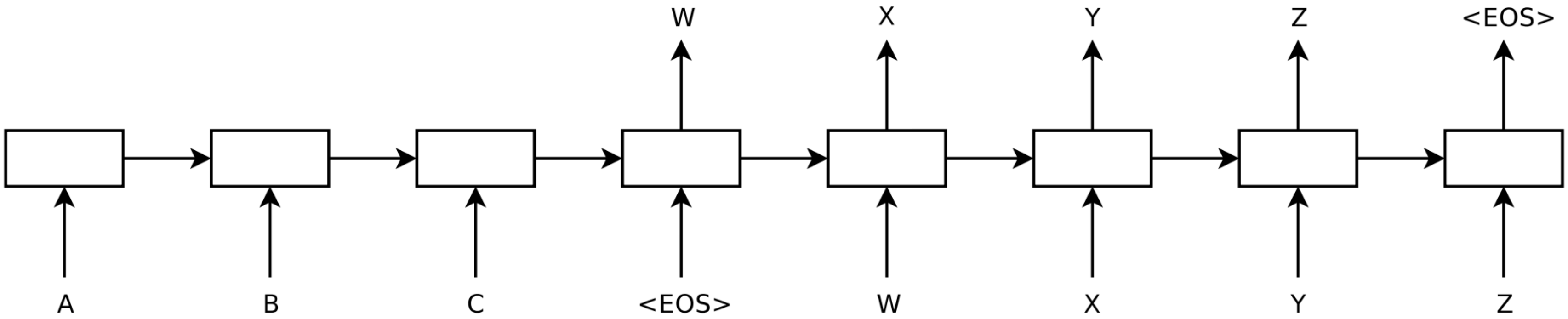
Sequence-to-Sequence Modeling and Attention

Sequence-to-Sequence Modeling

- data: <input sequence, output sequence> pairs
- use one RNN to encode input sequence
- use another RNN to decode the output sequence from the input sequence vector

Application: Machine Translation

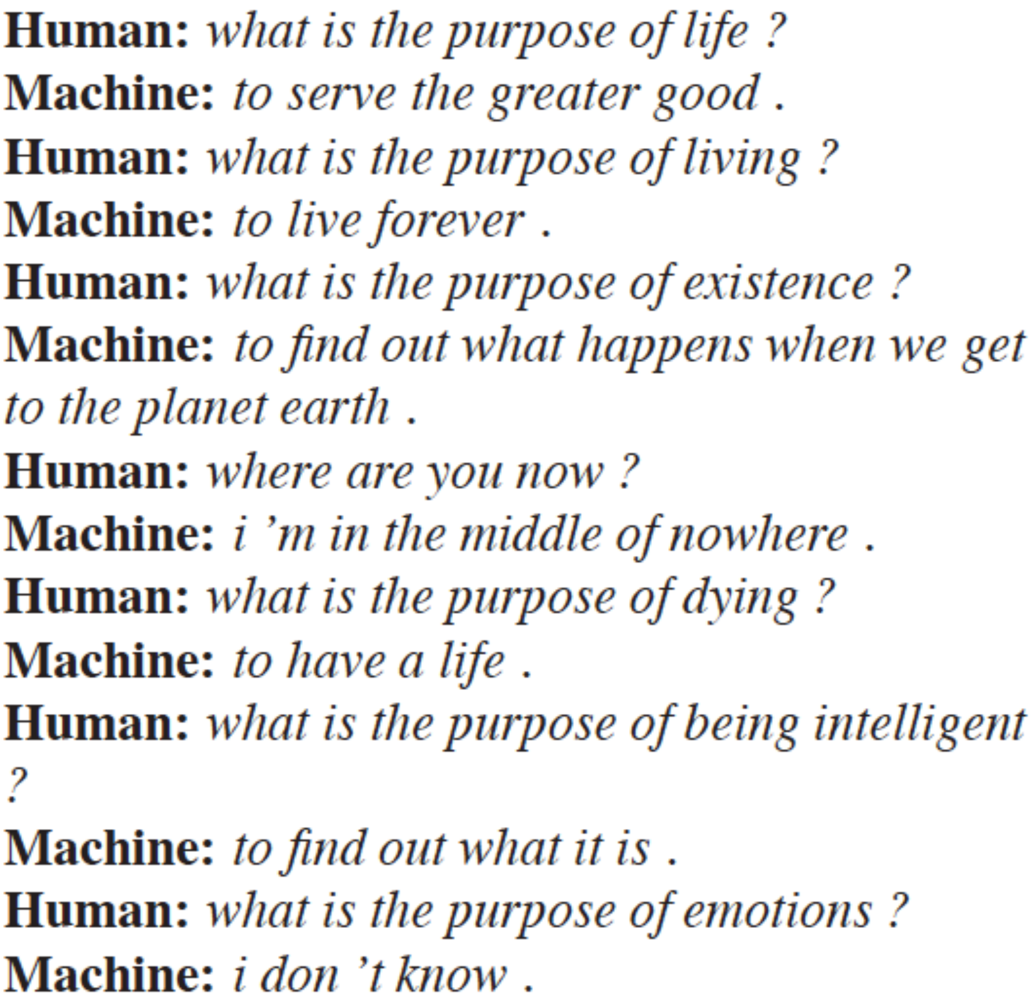
- Encode source sentence, decode translation



Sutskever, Vinyals, Le (2014)

Cho, van Merrienboer, Gulcehre, Bahdanau, Bougares, Schwenk, Bengio (2014)

Application: Chatbots



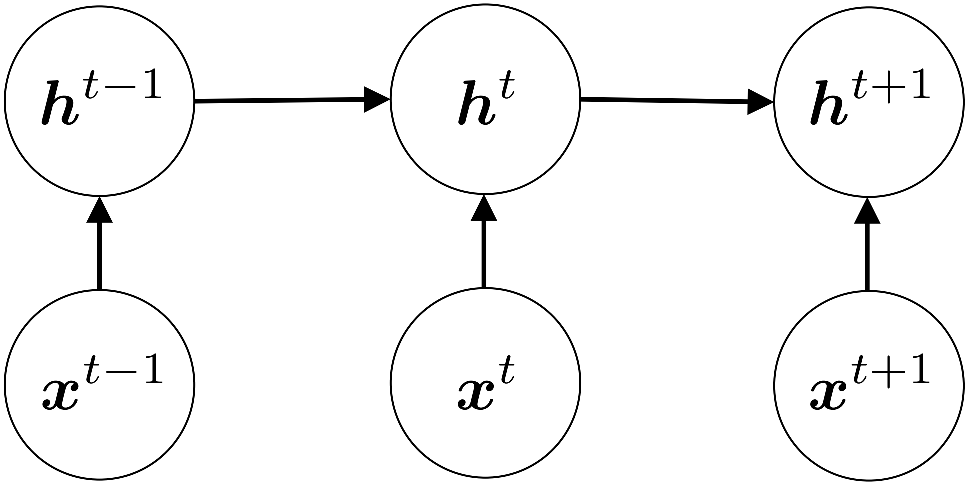
Human: *what is the purpose of life ?*
Machine: *to serve the greater good .*
Human: *what is the purpose of living ?*
Machine: *to live forever .*
Human: *what is the purpose of existence ?*
Machine: *to find out what happens when we get to the planet earth .*
Human: *where are you now ?*
Machine: *i 'm in the middle of nowhere .*
Human: *what is the purpose of dying ?*
Machine: *to have a life .*
Human: *what is the purpose of being intelligent ?*
Machine: *to find out what it is .*
Human: *what is the purpose of emotions ?*
Machine: *i don 't know .*

"A Neural Conversational Model"
Vinyals & Le (2015)

Input RNN ("Encoder")

$$\boldsymbol{h}^{t} = \tanh\left(W^{(x)}\boldsymbol{x}^{t} + W^{(h)}\boldsymbol{h}^{t-1} + \boldsymbol{b}^{(h)}\right)$$

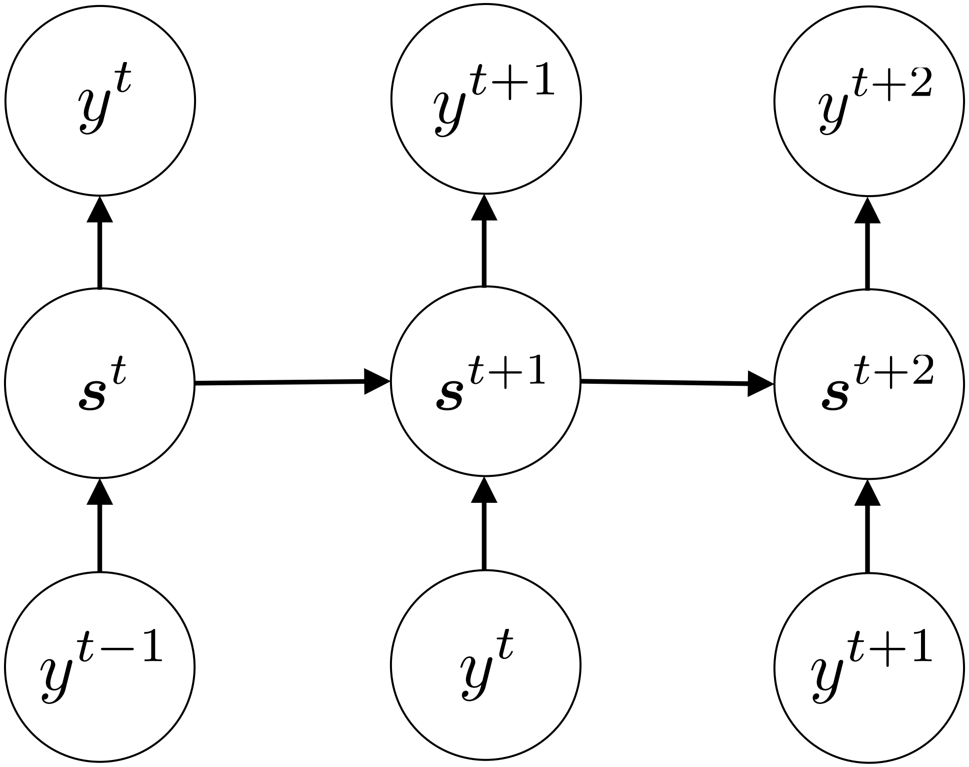
"hidden vector"



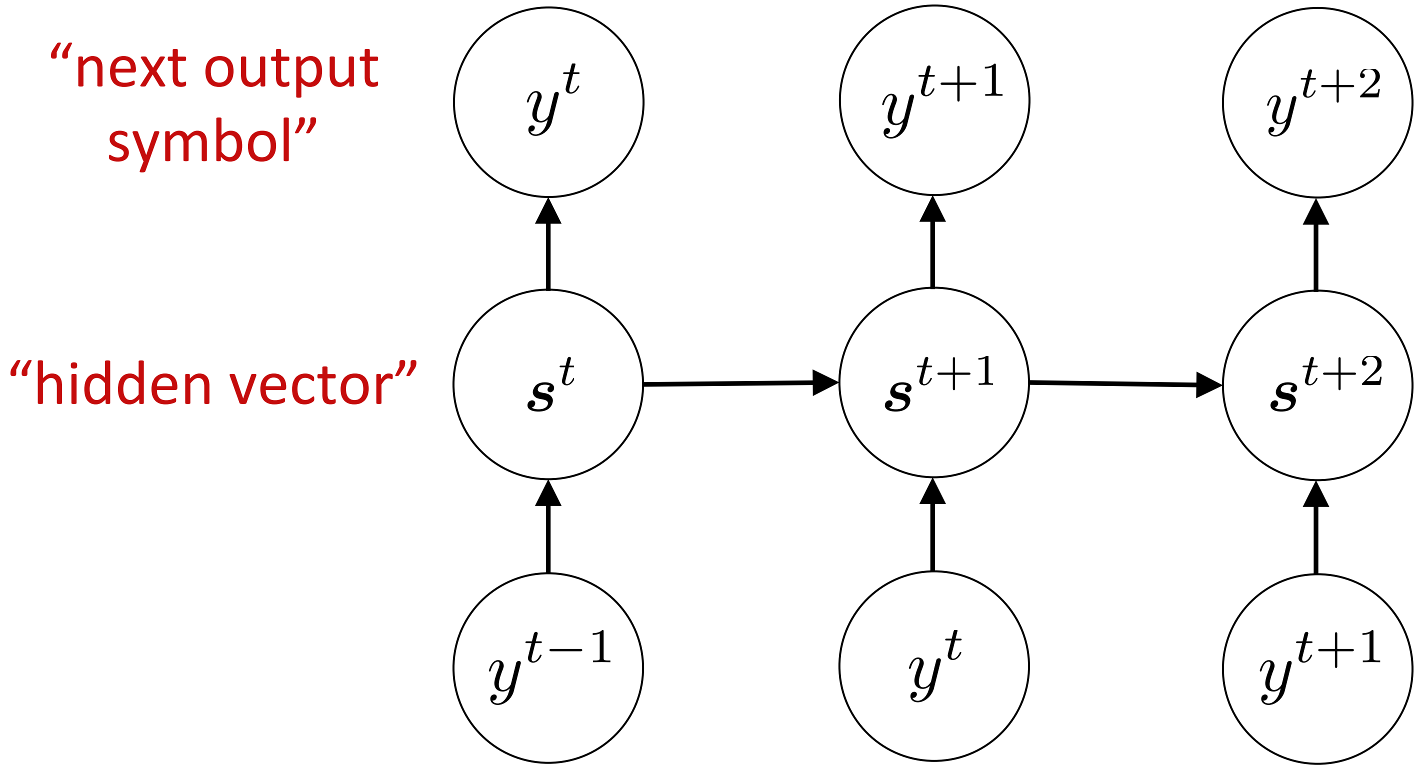
Output RNN ("Decoder")

"next output symbol"

"hidden vector"



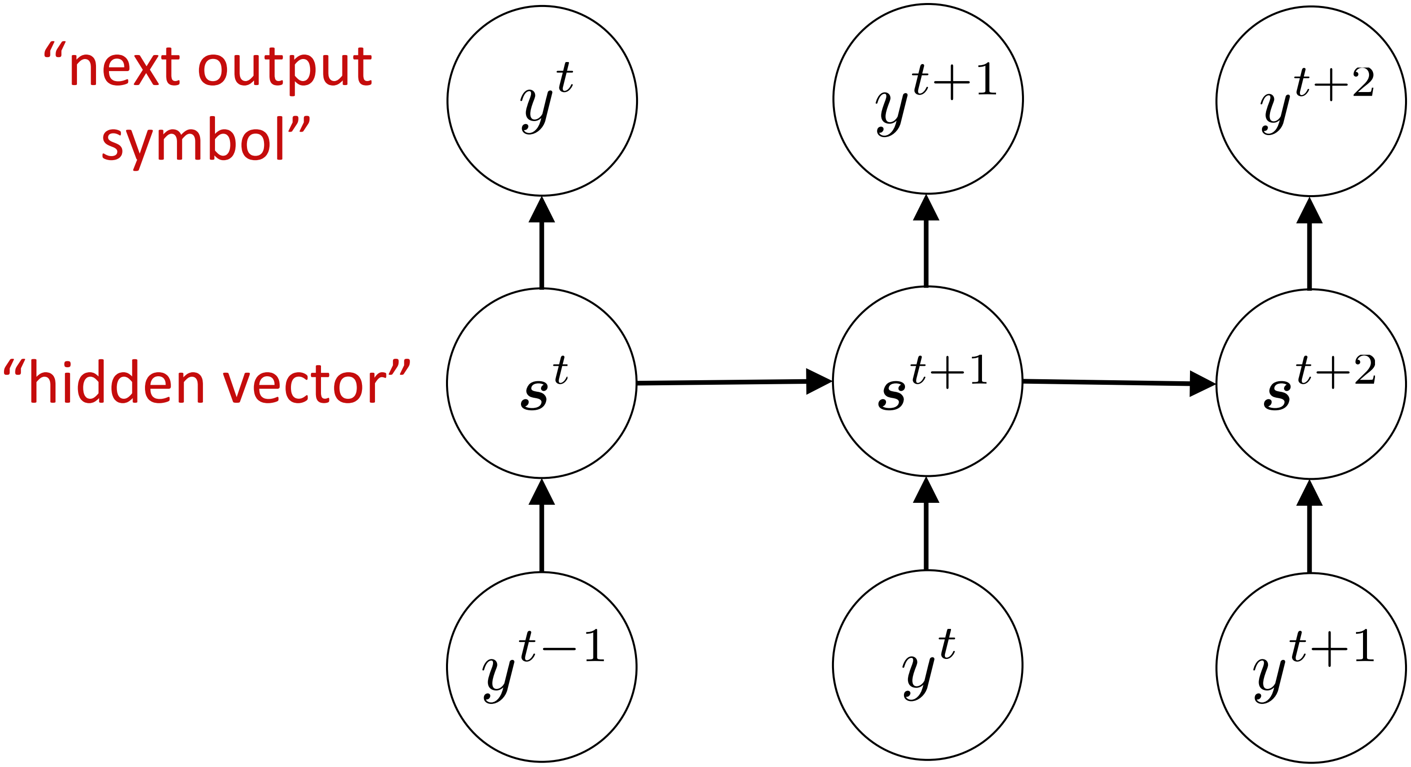
Output RNN ("Decoder")



$$\boldsymbol{s}^t = \tanh\left(W^{(y)}\boldsymbol{y}^{t-1} + W^{(s)}\boldsymbol{s}^{t-1} + \boldsymbol{b}^{(s)}\right)$$

Output RNN ("Decoder")

$$y^t = \operatorname*{argmax}_{y \in \mathcal{O}} \left(emb(y)^\top \boldsymbol{s}^t \right)$$

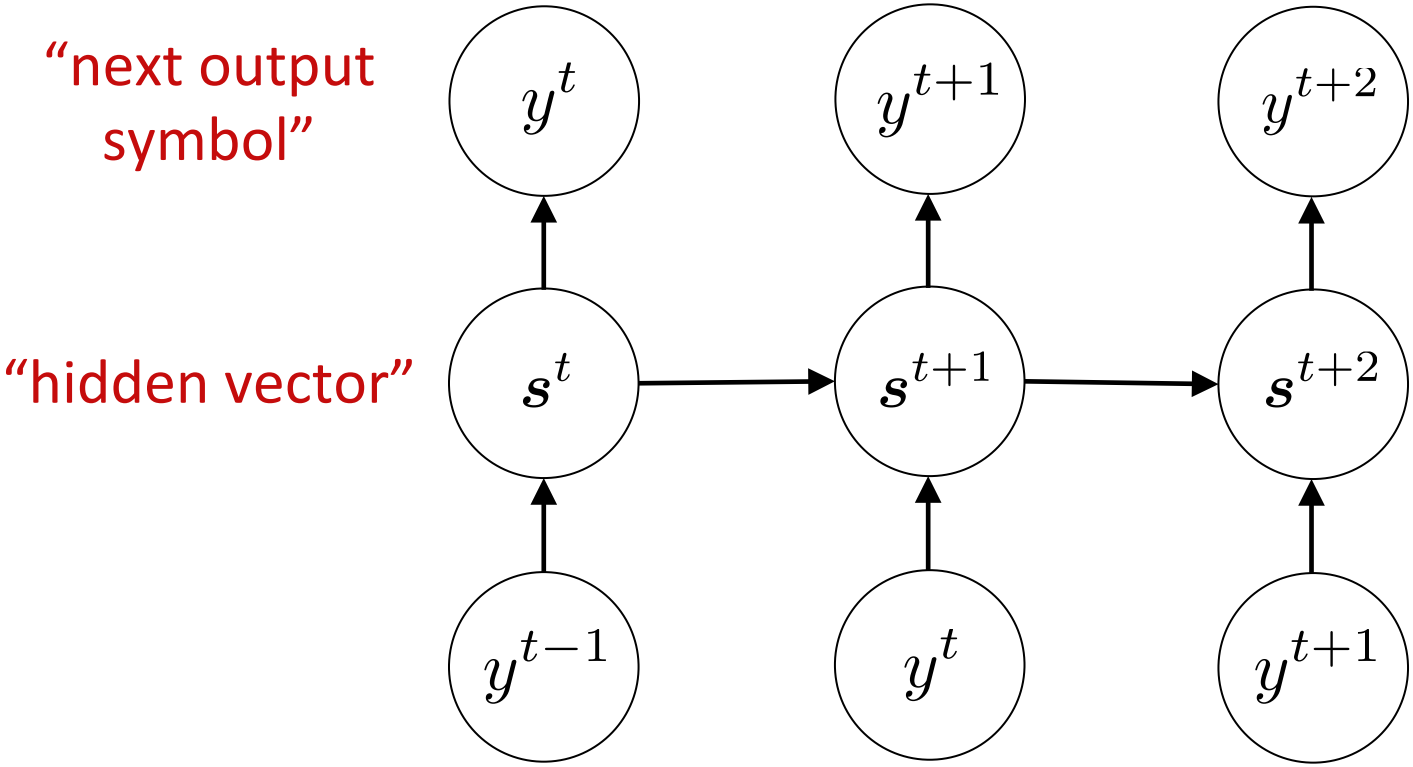


$$\boldsymbol{s}^t = \tanh\left(W^{(y)} \boldsymbol{y}^{t-1} + W^{(s)} \boldsymbol{s}^{t-1} + \boldsymbol{b}^{(s)} \right)$$

Distribution over next symbols?

$$y^t = \operatorname*{argmax}_{y \in \mathcal{O}} \left(emb(y)^\top \boldsymbol{s}^t \right)$$

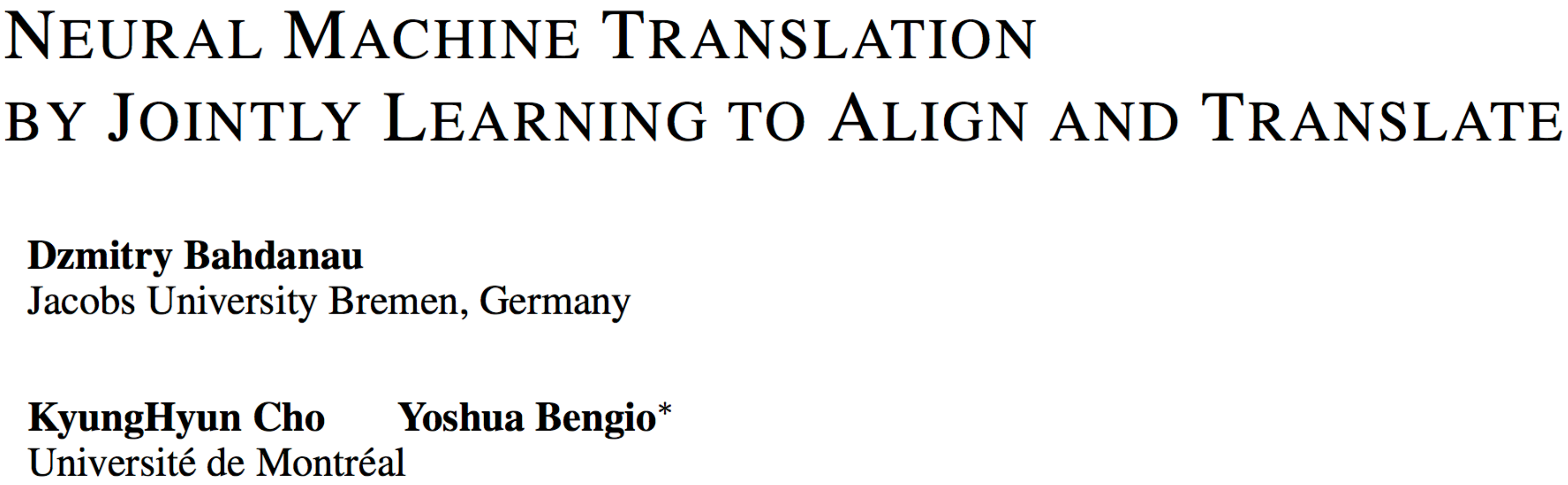
$$P(Y^t) = \text{softmax}\left(W \boldsymbol{s}^t \right)$$



“next output symbol”

“hidden vector”

Extension: Attention



Neural Machine Translation
by Jointly Learning to Align and Translate

Dzmitry Bahdanau
Jacobs University Bremen, Germany

KyungHyun Cho **Yoshua Bengio***
Université de Montréal

Extension: Attention

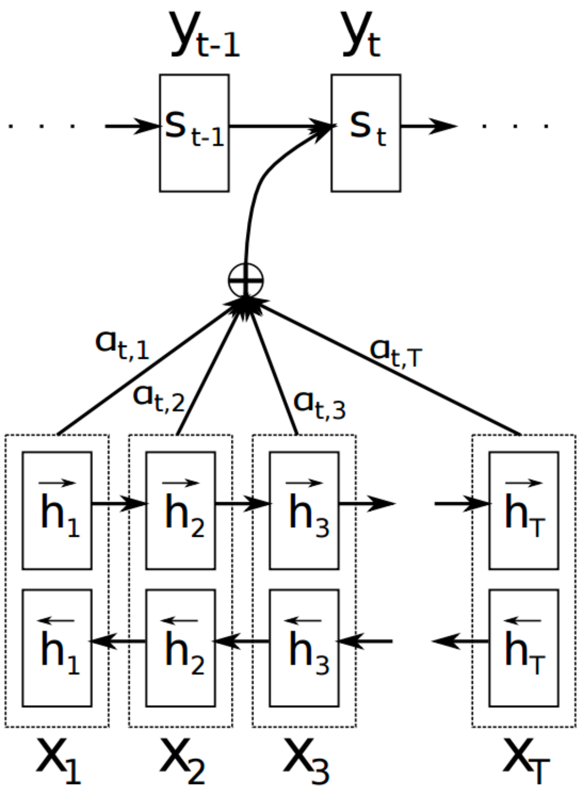
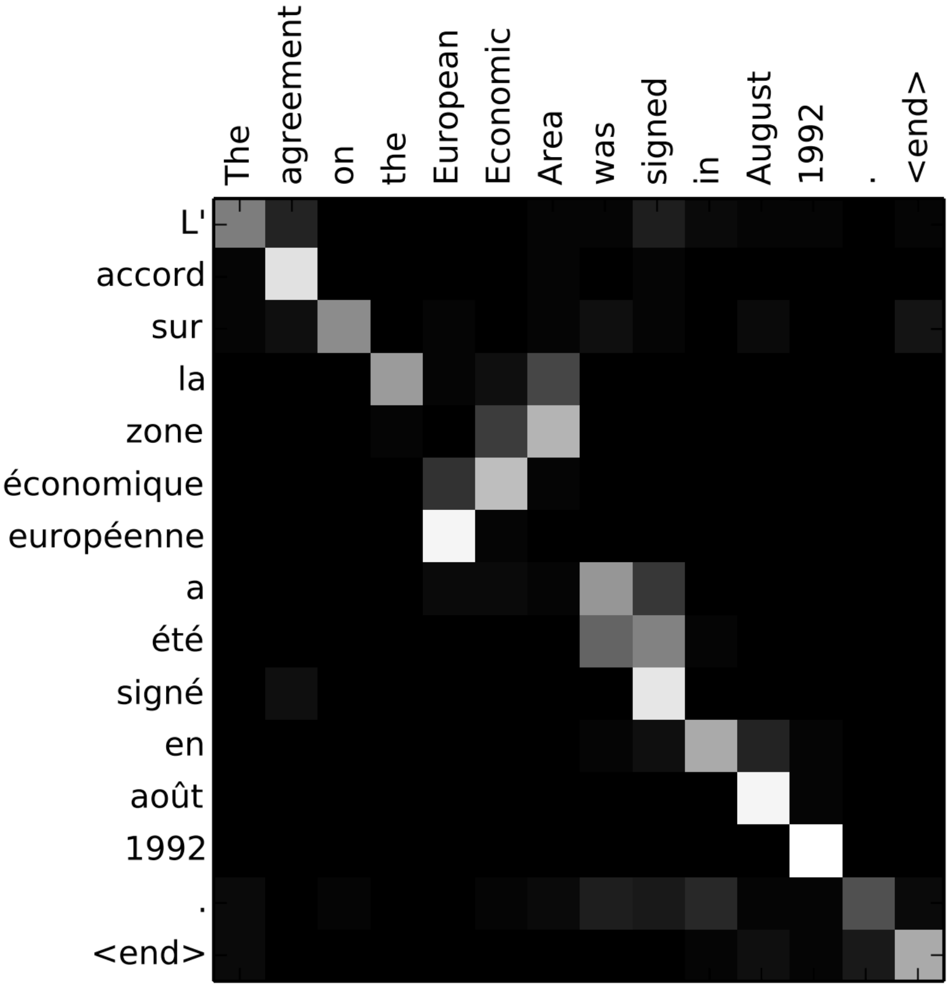


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

"Neural Machine Translation by Jointly Learning to Align and Translate"
Bahdanau et al. (2015)

Extension: Attention



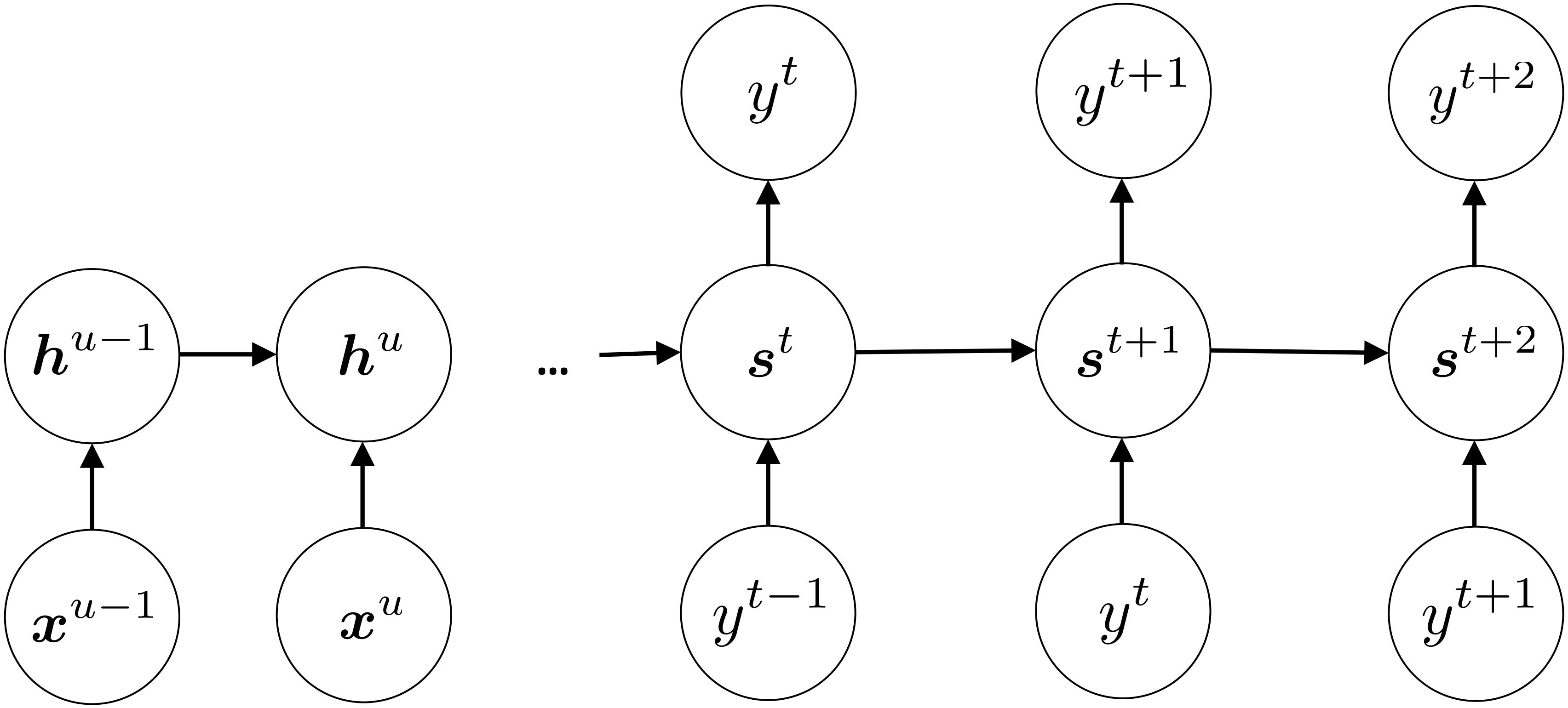
"Neural Machine Translation by Jointly Learning to Align and Translate"
Bahdanau et al. (2015)

- Disclaimer: the version I will present is a little simpler than the Bahdanau et al version

Adding Attention

$$\alpha^{t,u} \propto \exp\{att(\boldsymbol{s}^{t-1}, \boldsymbol{h}^{u})\}$$

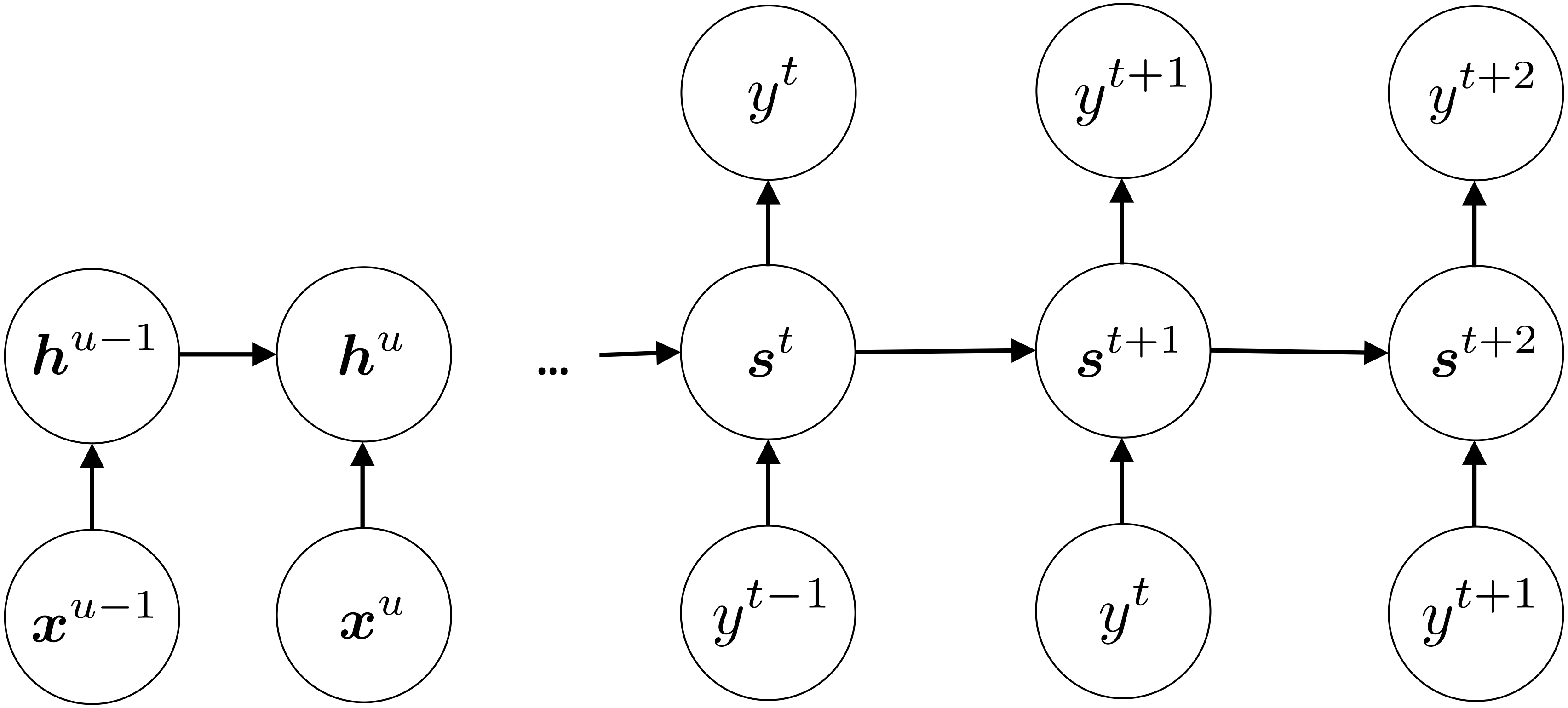
att **function models association between all pairs of hidden vectors in encoder and decoder**



Adding Attention

$$\alpha^{t,u} \propto \exp\{att(\boldsymbol{s}^{t-1}, \boldsymbol{h}^{u})\}$$

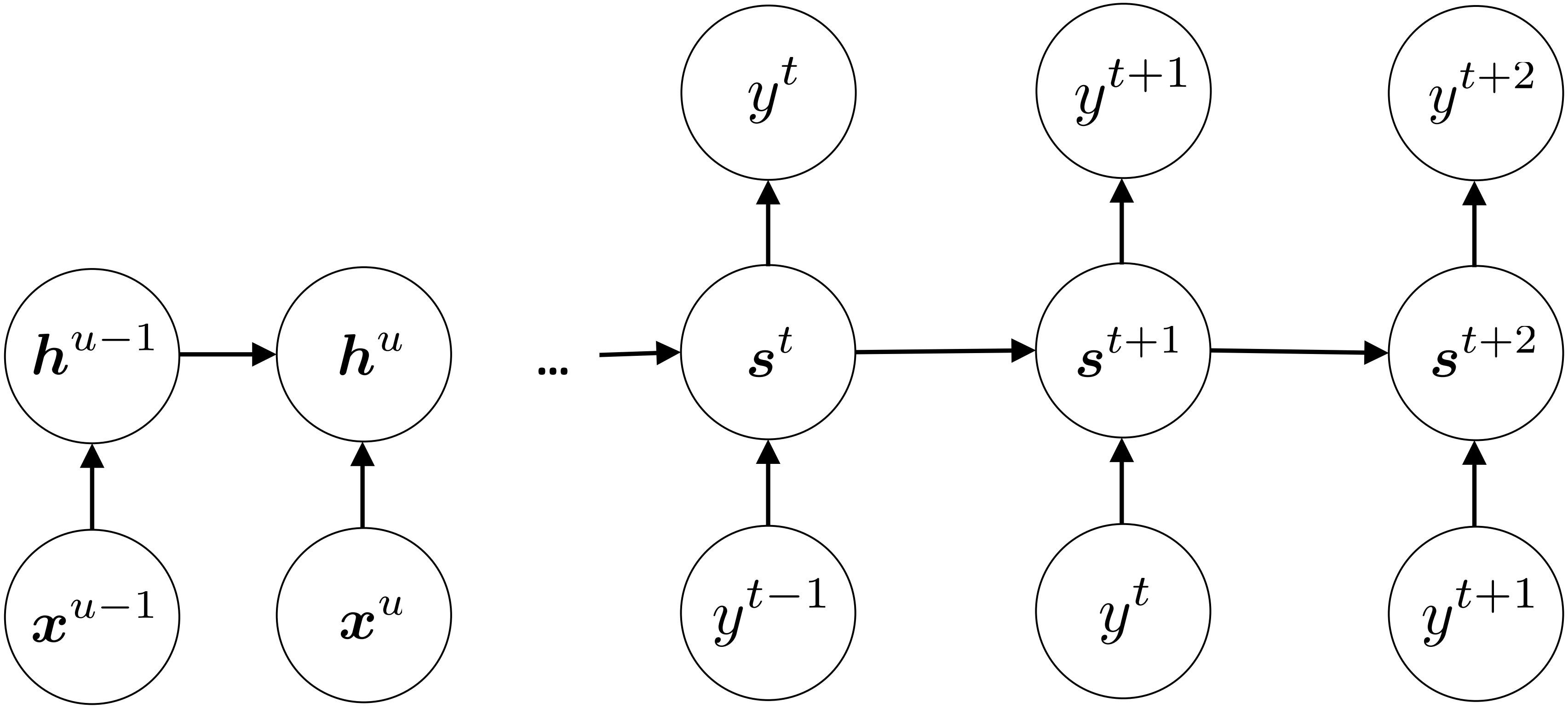
$$\boldsymbol{c}^{t} = \sum_{u=1}^{|\boldsymbol{x}|} \alpha^{t,u} \boldsymbol{h}^{u}$$



Adding Attention

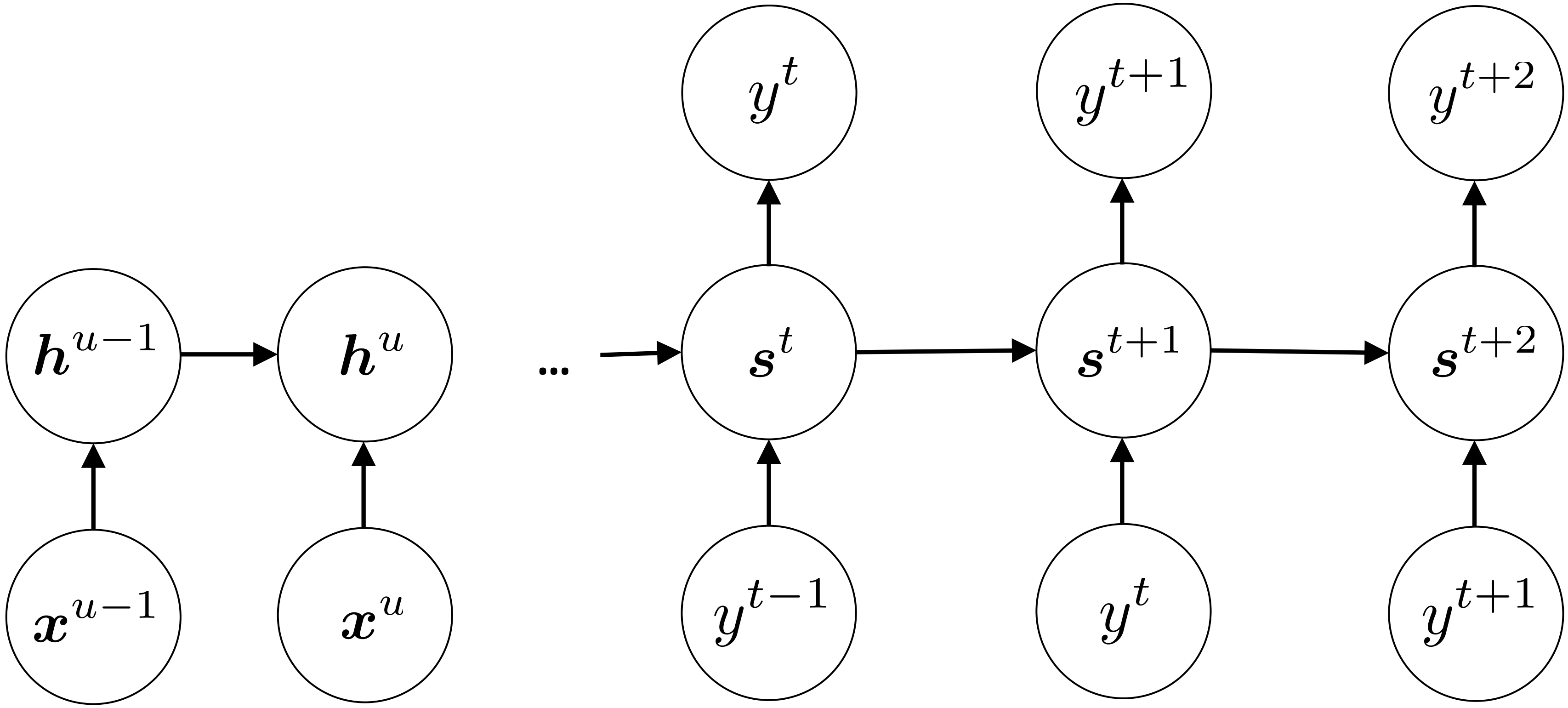
$$\boldsymbol{c}^t = \sum_{u=1}^{|\boldsymbol{x}|} \alpha^{t,u} \boldsymbol{h}^u$$

$$\boldsymbol{s}^t = \tanh\left(W^{(y)} \boldsymbol{y}^{t-1} + W^{(s)} \boldsymbol{s}^{t-1} + W^{(c)} \boldsymbol{c}^t + \boldsymbol{b}^{(s)}\right)$$

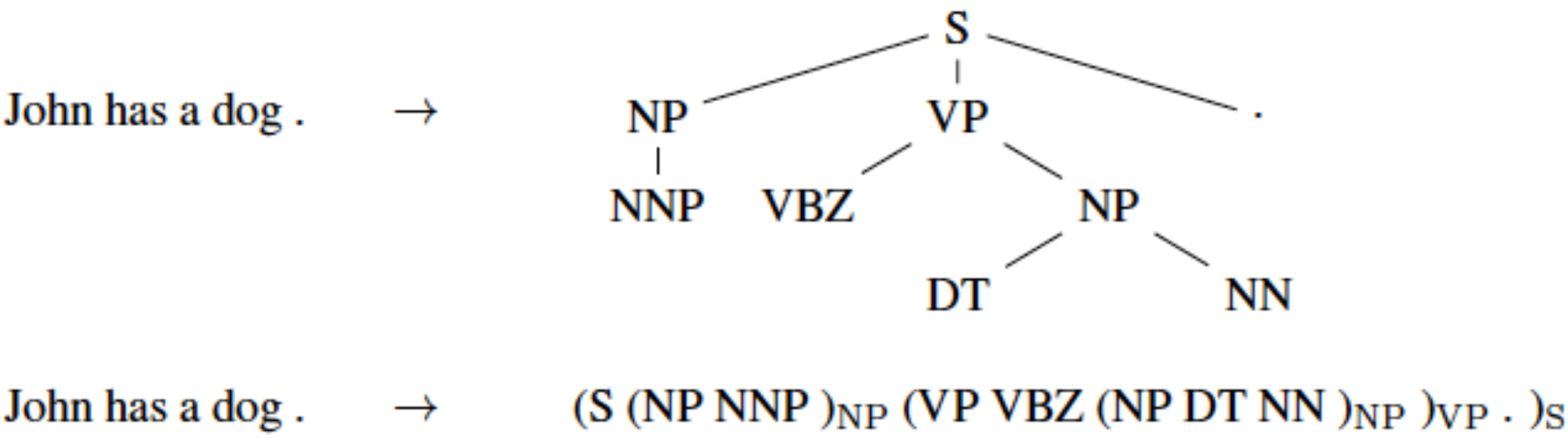


Adding Attention

$$y^t = \underset{y \in \mathcal{O}}{\operatorname{argmax}} \left(emb(y)^\top [\boldsymbol{s}^t ; \boldsymbol{c}^t] \right)$$



Application: Parsing



John has a dog . →

John has a dog . → (S (NP NNP)$_{NP}$ (VP VBZ (NP DT NN)$_{NP}$)$_{VP}$.)$_{S}$

Figure 2: Example parsing task and its linearization.

"Grammar as a Foreign Language"
{Vinyals, Kaiser} et al. (2015)

Application: Parsing

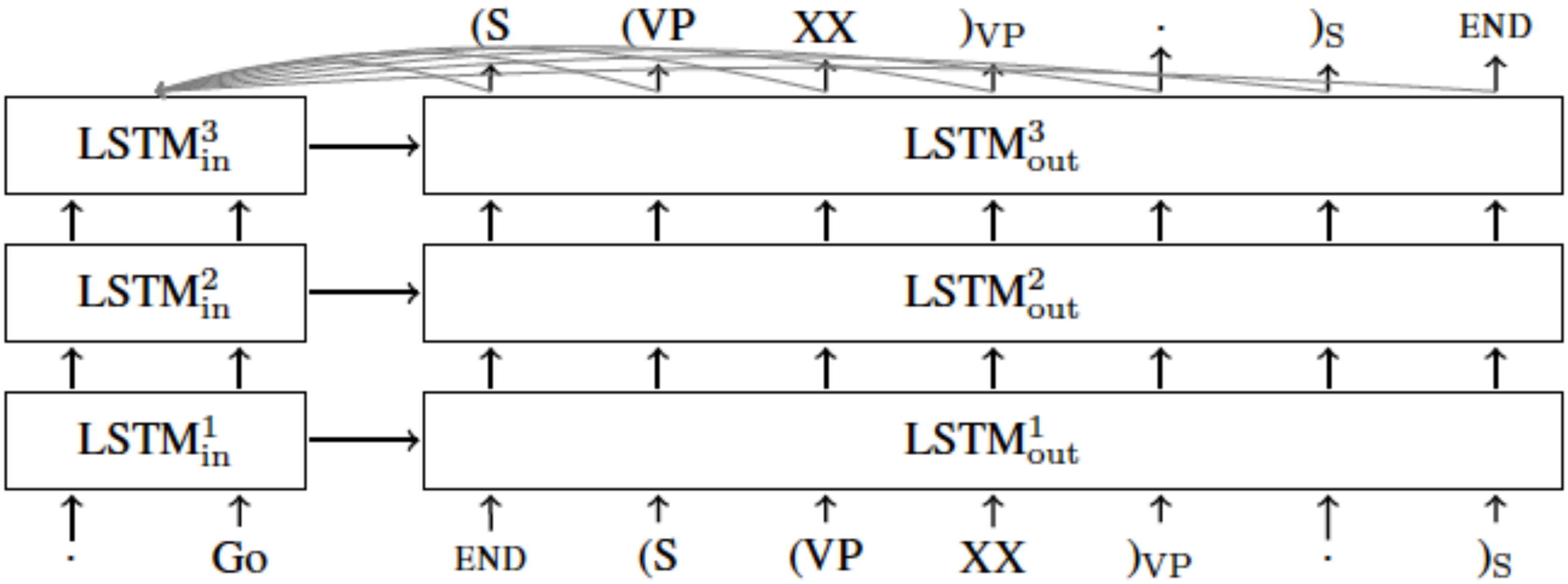
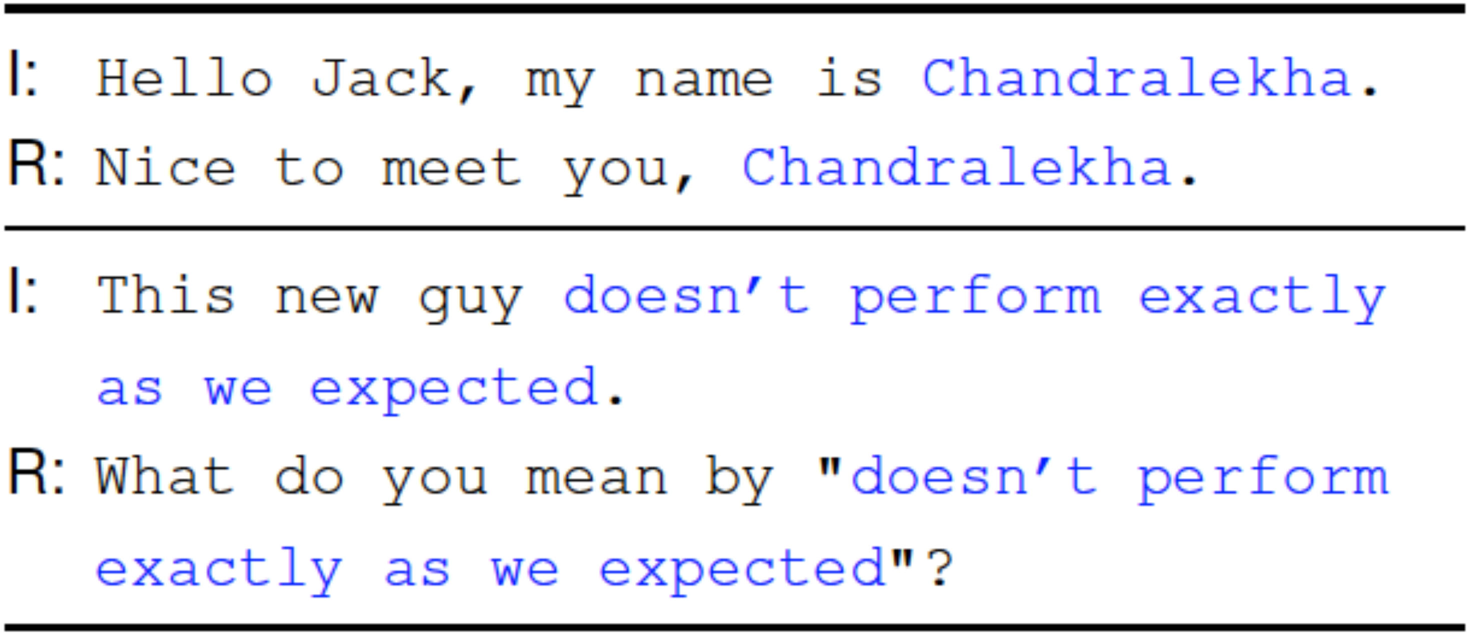


Figure 1: A schematic outline of a run of our LSTM+A model on the sentence "Go.". See text for details.

"Grammar as a Foreign Language"
{Vinyals, Kaiser} et al. (2015)

Extension: Copy Mechanism



I: Hello Jack, my name is Chandralekha.

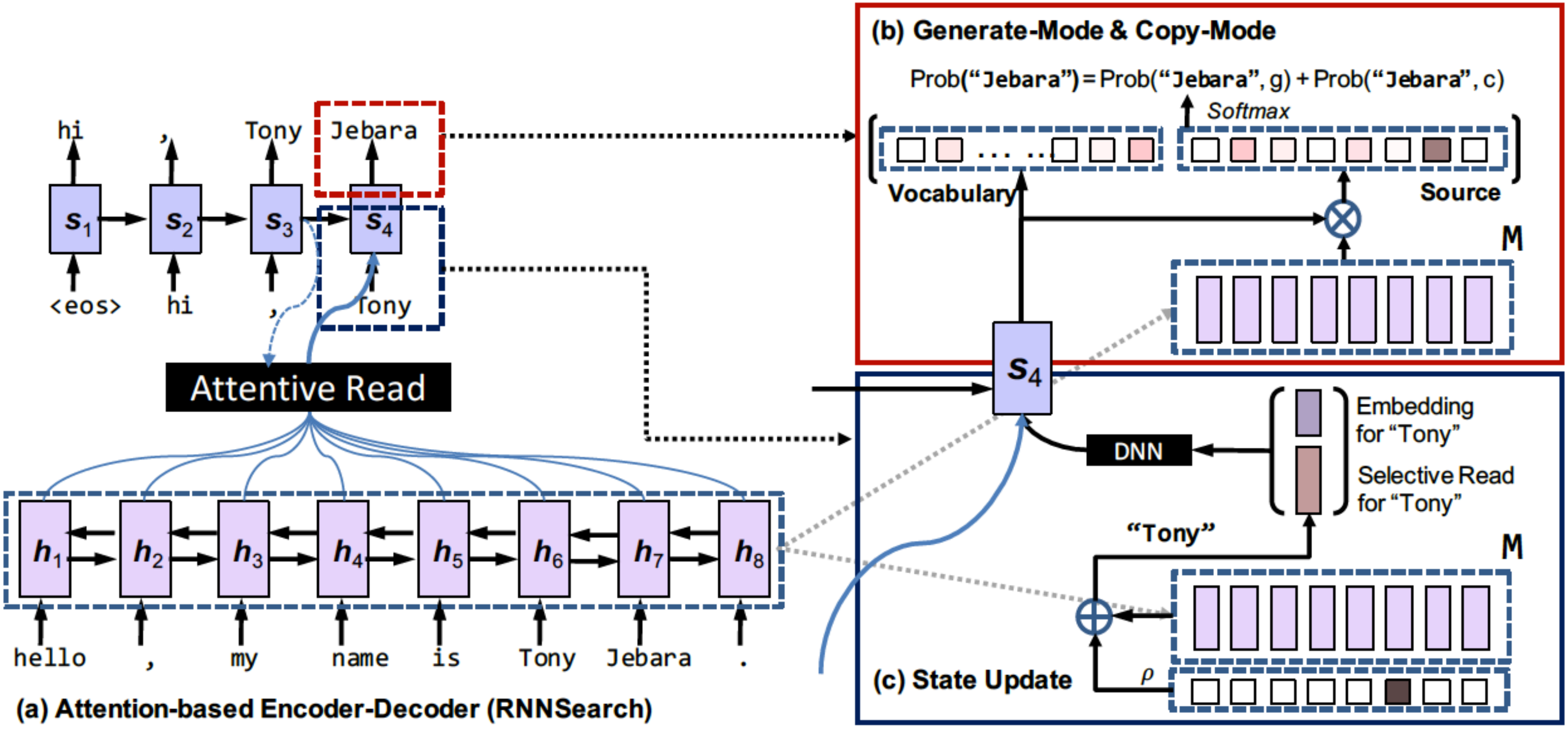
R: Nice to meet you, Chandralekha.

I: This new guy doesn't perform exactly as we expected.

R: What do you mean by "doesn't perform exactly as we expected"?

"Incorporating Copying Mechanism in Sequence-to-Sequence Learning"
Gu et al. (2016)

Extension: Copy Mechanism



"Incorporating Copying Mechanism in Sequence-to-Sequence Learning"
Gu et al. (2016)

Extension: Copy Mechanism

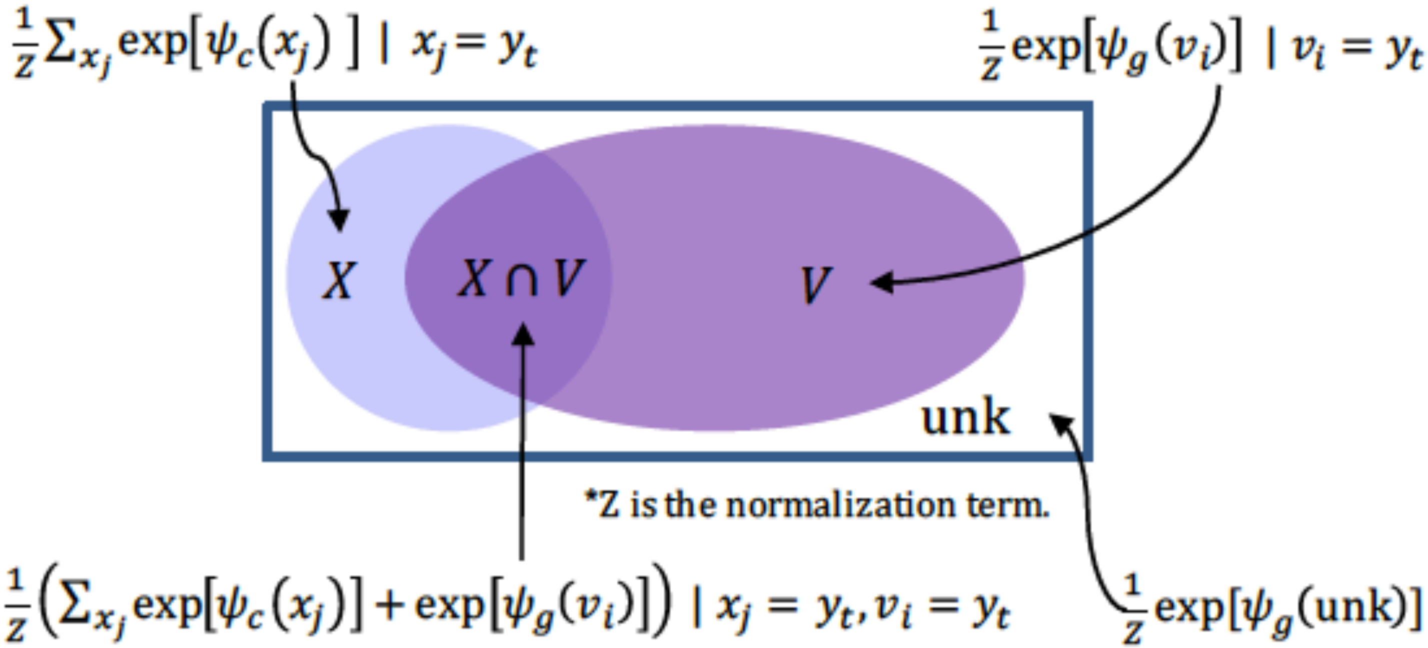

$\frac{1}{Z}\Sigma_{x_j}\exp[\psi_c(x_j)\,] \mid x_j = y_t$
$\frac{1}{Z}\exp[\psi_g(v_i)] \mid v_i = y_t$
X
$X \cap V$
V
unk
*Z is the normalization term.
$\frac{1}{Z}\left(\Sigma_{x_j}\exp[\psi_c(x_j)] + \exp[\psi_g(v_i)]\right) \mid x_j = y_t, v_i = y_t$
$\frac{1}{Z}\exp[\psi_g(\text{unk})]$

Figure 2: The illustration of the decoding probability $p(y_t|\cdot)$ as a 4-class classifier.

“Incorporating Copying Mechanism in Sequence-to-Sequence Learning”
Gu et al. (2016)

Reading Comprehension

Dataset: MCTest

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

What did Fritz draw first?

A) the toothpaste

B) his mama

C) cereal and milk

D) his bicycle

“MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text” Richardson et al. (2013)

Dataset: MCTest

Once there was a boy named Fritz who loved to draw. He drew everything. In the morning, he drew a picture of his cereal with milk. His papa said, “Don’t draw your cereal. Eat it!”

After school, Fritz drew a picture of his bicycle. His uncle said, “Don't draw your bicycle. Ride it!”

…

What did Fritz draw first?

A) the toothpaste

B) his mama

C) cereal and milk

D) his bicycle

“MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text” Richardson et al. (2013)

Dataset: MCTest

- 660 fictional stories, written at a 4th grade reading level
- 4 multiple choice questions per story

“MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text” Richardson et al. (2013)

Dataset: CNN/Daily Mail Comprehension Tasks

Document:

actress @entity1 has entered a rehab facility for her addictions , a spokesman said . " @entity1 has valiantly battled substance abuse over the years and whenever she has needed to seek treatment she has done so , " said spokesman @entity5 … @entity1 won an @entity15 in 1973 for her performance in " cabaret . " …

Question:

XXXXX won an @entity15 for her performance in " cabaret "

“Teaching Machines to Read and Comprehend”
Hermann et al. (2015)

Dataset: CNN/Daily Mail Comprehension Tasks

Document:

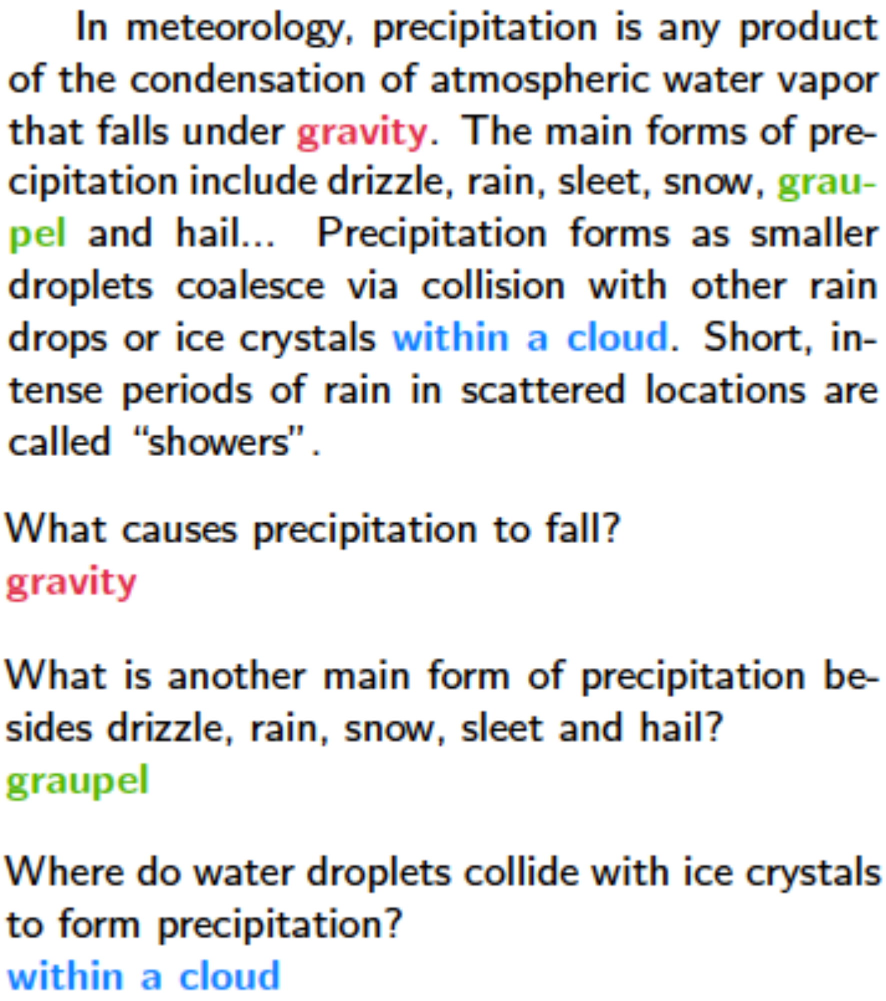
actress @entity1 has entered a rehab facility for her addictions , a spokesman said . " @entity1 has valiantly battled substance abuse over the years and whenever she has needed to seek treatment she has done so , " said spokesman @entity5 … **@entity1 won an @entity15 in 1973 for her performance in " cabaret . "** …

Question:

XXXXX won an @entity15 for her performance in " cabaret "

"Teaching Machines to Read and Comprehend"
Hermann et al. (2015)

Dataset: SQuAD



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

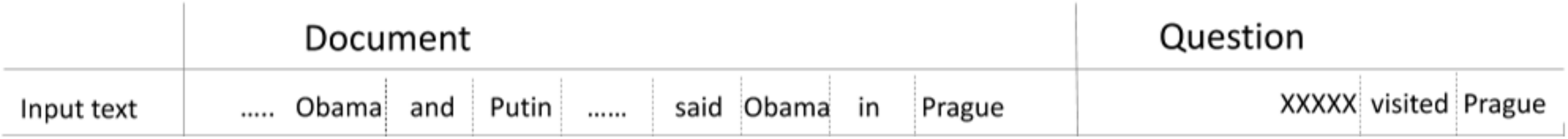
"SQuAD: 100,000+ Questions for Machine Comprehension of Text"
Rajpurkar et al. (2016)

Neural Models for Comprehension

- lots of recent activity here!
  - Hermann et al. (2015)
  - Hill et al. (2016)
  - Chen et al. (2016)
  - Kadlec et al. (2016)
  - Dhingra et al. (2016)
  - *inter alia*
- we will describe the Attention Sum Reader (Kadlec et al., 2016) because it is simple and works well

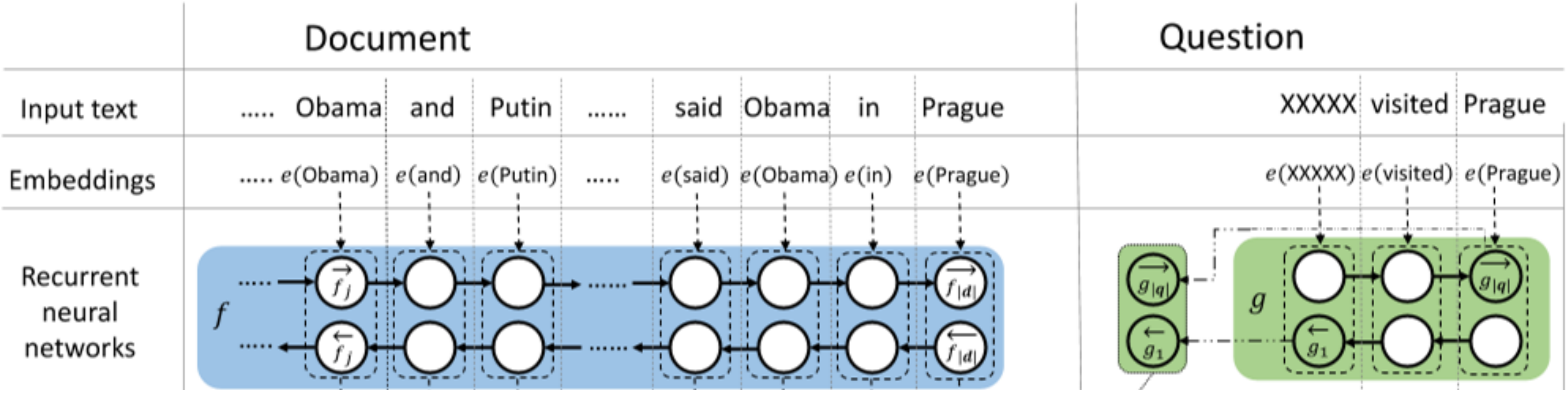
Attention Sum Reader

(Kadlec et al., 2016)



	Document									Question		
Input text	.....	Obama	and	Putin	......	said	Obama	in	Prague	XXXXX	visited	Prague

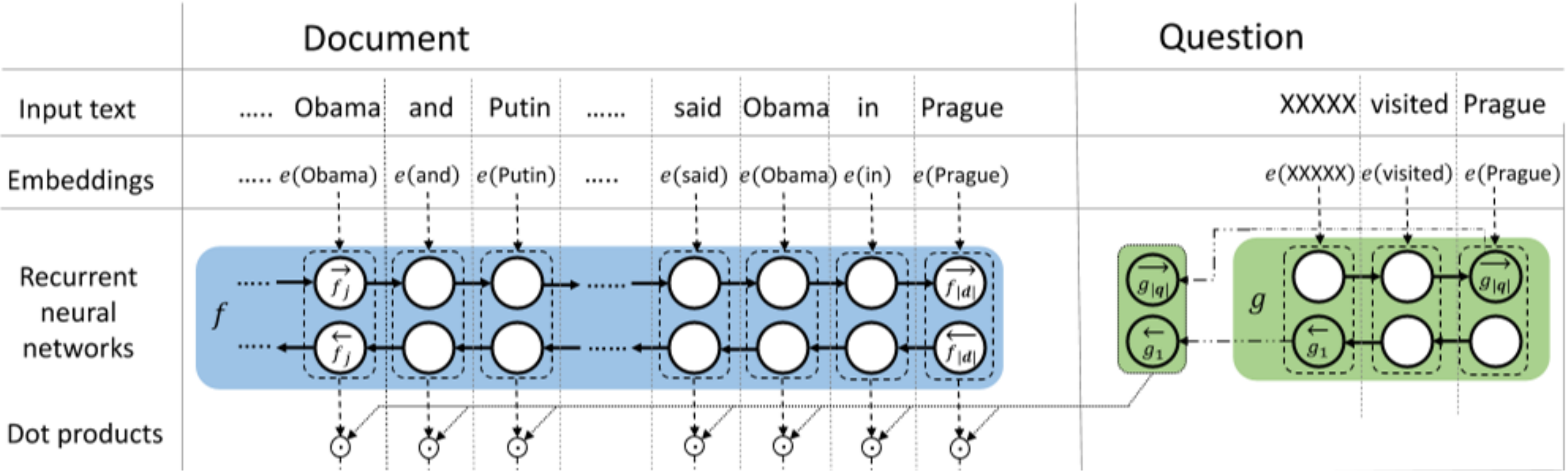
Attention Sum Reader
(Kadlec et al., 2016)



- Encode document using bidirectional RNN
- Encode question using another bidirectional RNN

Attention Sum Reader

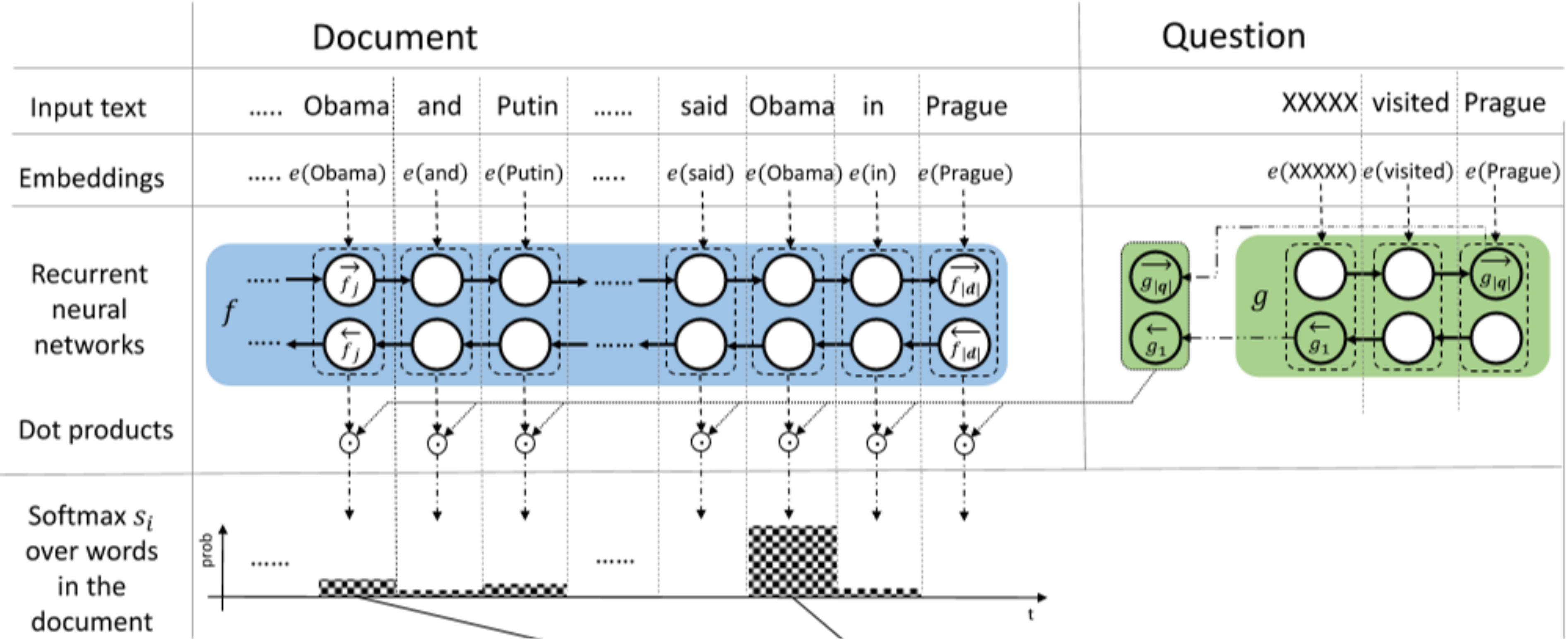
(Kadlec et al., 2016)


Document
Question
Input text
..... Obama and Putin ...... said Obama in Prague
XXXXX visited Prague
Embeddings
..... e(Obama) e(and) e(Putin) ..... e(said) e(Obama) e(in) e(Prague)
e(XXXXX) e(visited) e(Prague)
Recurrent neural networks
f
$\overrightarrow{f_j}$
$\overleftarrow{f_j}$
$\overrightarrow{f_{|d|}}$
$\overleftarrow{f_{|d|}}$
g
$\overrightarrow{g_{|q|}}$
$\overleftarrow{g_1}$
Dot products

- Compute attention over positions of document using question representation

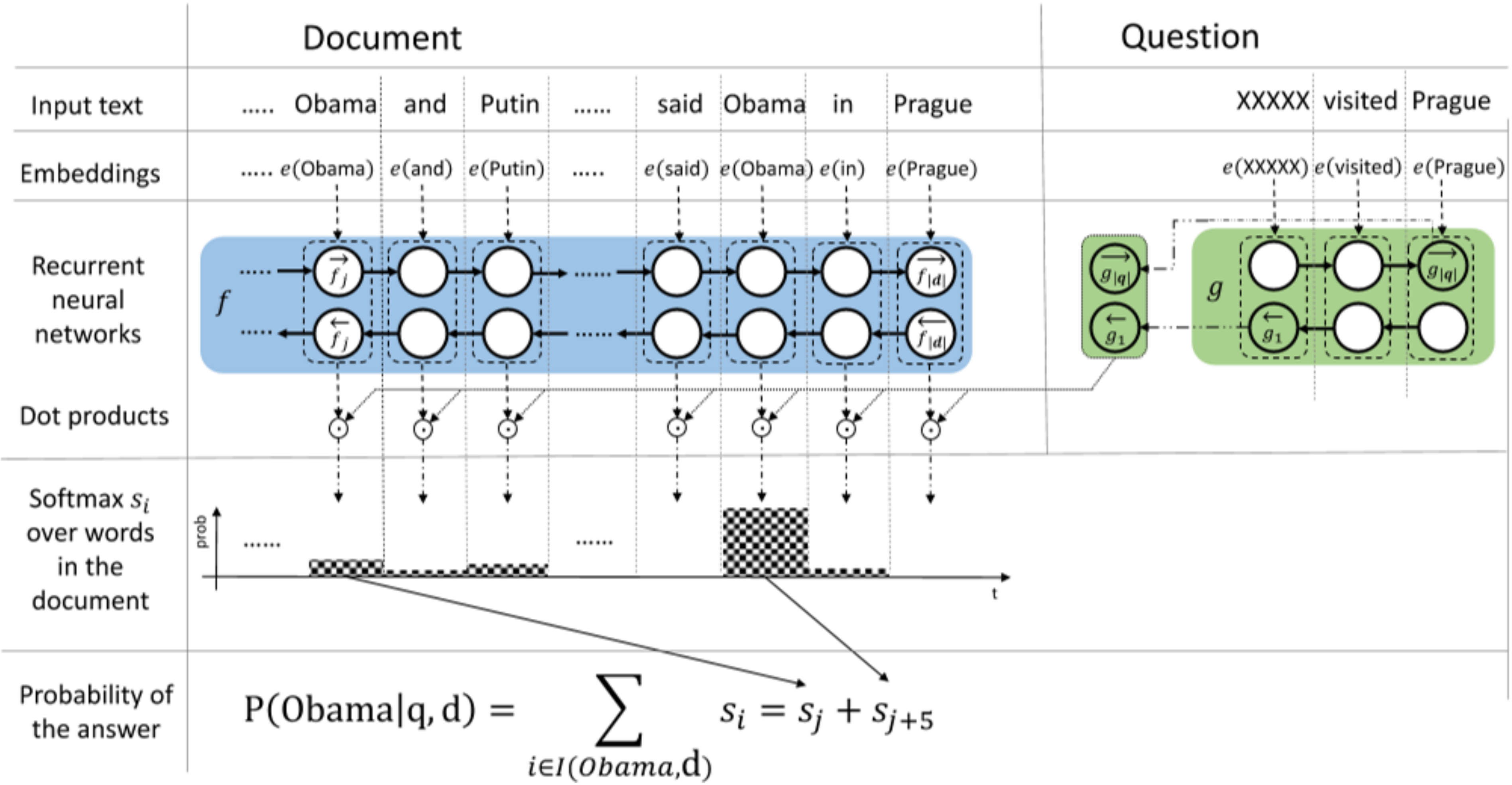
Attention Sum Reader

(Kadlec et al., 2016)


Document
Question
Input text
..... Obama and Putin ...... said Obama in Prague
XXXXX visited Prague
Embeddings
..... e(Obama) e(and) e(Putin) ..... e(said) e(Obama) e(in) e(Prague)
e(XXXXX) e(visited) e(Prague)
Recurrent neural networks
f
$\overrightarrow{f_j}$
$\overleftarrow{f_j}$
$\overrightarrow{f_{|d|}}$
$\overleftarrow{f_{|d|}}$
g
$\overrightarrow{g_{|q|}}$
$\overleftarrow{g_1}$
Dot products
Softmax s_i over words in the document
prob
t

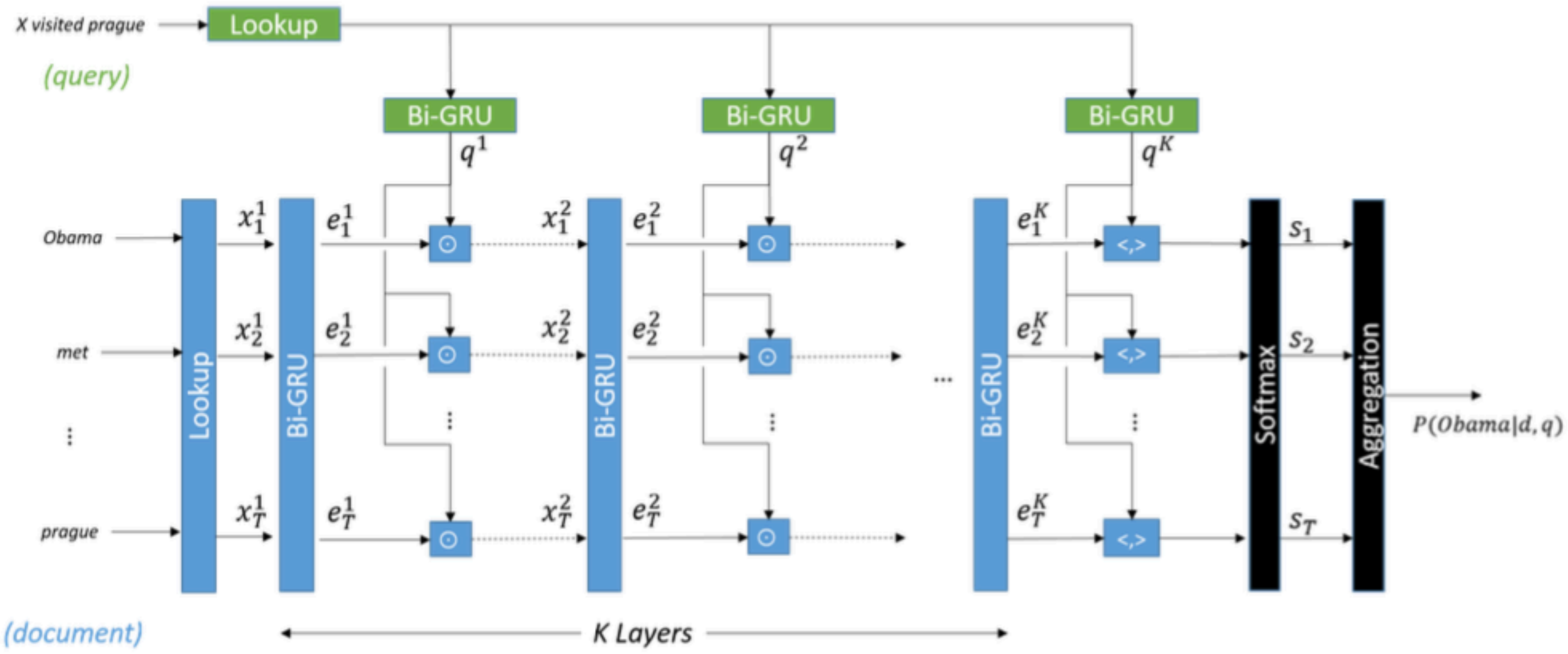
- Normalize over positions of document

Attention Sum Reader
(Kadlec et al., 2016)


Document
Question
Input text
..... Obama and Putin ..... said Obama in Prague
XXXXX visited Prague
Embeddings
..... e(Obama) e(and) e(Putin) ..... e(said) e(Obama) e(in) e(Prague)
e(XXXXX) e(visited) e(Prague)
Recurrent neural networks
f
$\overrightarrow{f_j}$
$\overleftarrow{f_j}$
$\overrightarrow{f_{|d|}}$
$\overleftarrow{f_{|d|}}$
g
$\overrightarrow{g_{|q|}}$
$\overleftarrow{g_1}$
Dot products
Softmax s_i over words in the document
prob
t
Probability of the answer

$$\mathrm{P}(\mathrm{Obama}|\mathrm{q},\mathrm{d}) = \sum_{i \in I(Obama,\mathrm{d})} s_i = s_j + s_{j+5}$$

Gated Attention Reader
(Dhingra et al., 2016)


X visited prague
Lookup
(query)
Bi-GRU
q^1
Bi-GRU
q^2
Bi-GRU
q^K
Obama
met
⋮
prague
Lookup
x_1^1
x_2^1
x_T^1
Bi-GRU
e_1^1
e_2^1
e_T^1
⊙
x_1^2
x_2^2
x_T^2
Bi-GRU
e_1^2
e_2^2
e_T^2
⊙
...
Bi-GRU
e_1^K
e_2^K
e_T^K
<,>
Softmax
s_1
s_2
s_T
Aggregation
$P(Obama|d,q)$
(document)
K Layers