Neural Machine Translation

Thang Luong

Lecture @ CS224N

(Thanks to Chris Manning, Abigail See, and Russell Stewart for comments and discussions.)

Sixi roasted husband

Meat Muscle Stupid Bean Sprouts

Sixi roasted husband

Meat Muscle Stupid Bean Sprouts

Let's backtrack

[rescue] [staff] [in] [collapse] [of] [house] [in] [search] [survivors]

救援 人员 在 倒塌 的 房屋 里 寻找 生还者 。

Rescue workers search for survivors in collapsed houses .

(From Chris Manning)

- MT: learn to translate from parallel corpora.

Let's backtrack – Approaches

Probabilistic "Dictionary"
(Brown et al., 1993)

```
" التطورات ||| " developments ||| 1 0.506684 1 0.863476 2
" التطورات في ||| " developments in the ||| 1 0.336365 0.4
" التطورات في ||| " developments in ||| 1 0.336365 0.53846
" التطورات في ميدان ||| " developments in the field of |||
" التطورات في ميدان ||| " developments in the field ||| 0.
" التطورات في ميدان المعلومات ||| " developments in the fie
" التطورات في ميدان ||| " developments in the field of ||
" التطورات في ميدان ||| " developments in the field ||| C
" التطورات في ميدان المعلومات ||| " developments in the fi
" التعاون ||| " cooperation ||| 0.965517 0.619446 0.98245
" التعاون ||| " enhanced ||| 1 0.0100982 0.0175439 0.0007
" التعاون الإقليمي " ||| " enhanced regional " ||| 1 0.003
" تدرج كلمة " التعاون الإقليمي " ||| " enhanced regional "
```

Phrase Table
(Koehn et al., 2003, Och & Ney, 2004)

It's 2015! "**Sentence" Table**?

Can we build a *"sentence" table*?

WHAT THE BRITISH SAY	WHAT THE BRITISH MEAN
❑ With the greatest respect	❑ You are an idiot
❑ That's not bad	❑ That's good
❑ Very interesting	❑ That is clearly nonsense
❑ Quite good	❑ A bit disappointing

- $|V|^N$ combinations in principle.
  - English: average sentence length 14 (~ 70 chars).
  - 10^{14}*70*2 bytes = 14M gigabytes.

Neural Machine Translation to the rescue!

- Store a sentence table implicitly.
- Simple and coherent.

But we need to understand Recurrent Neural Networks first!

Outline

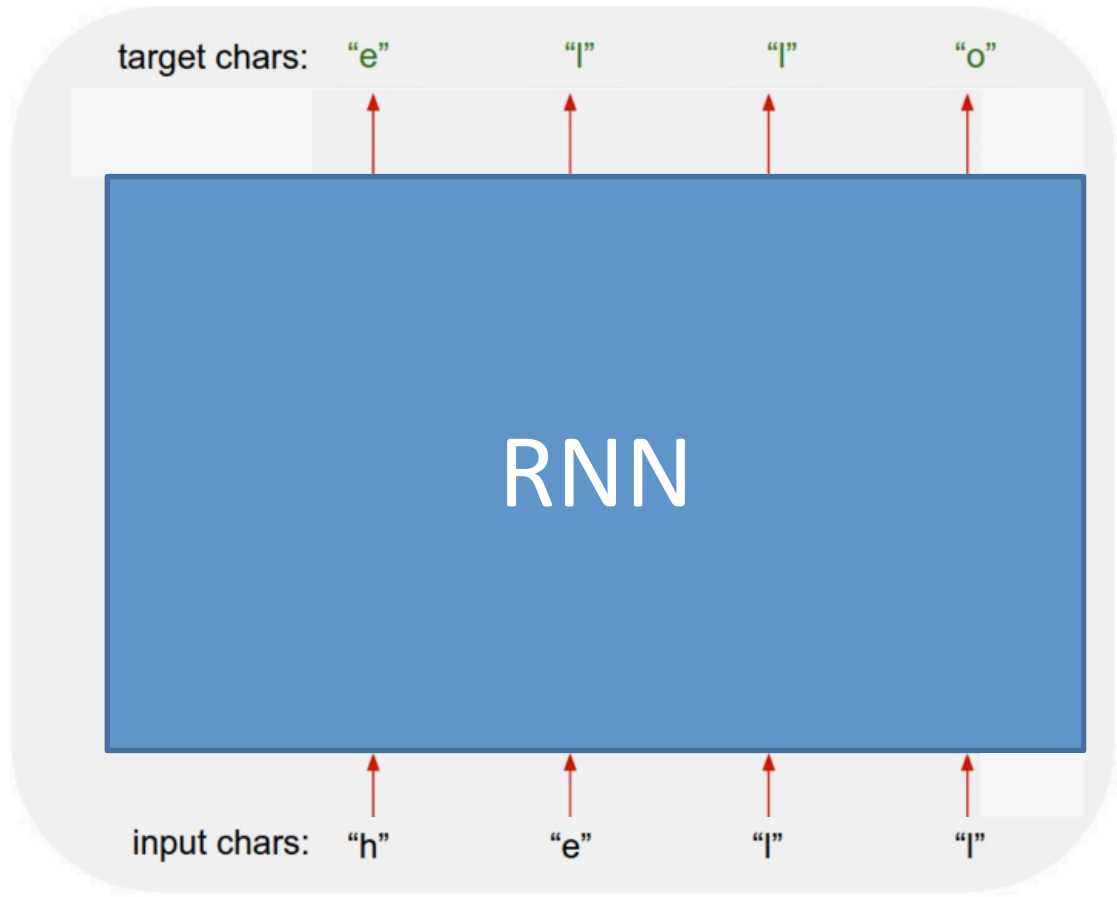
- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
- Attention mechanism (Bahdanau et al., 2015)

Recurrent Neural Networks (RNNs)

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



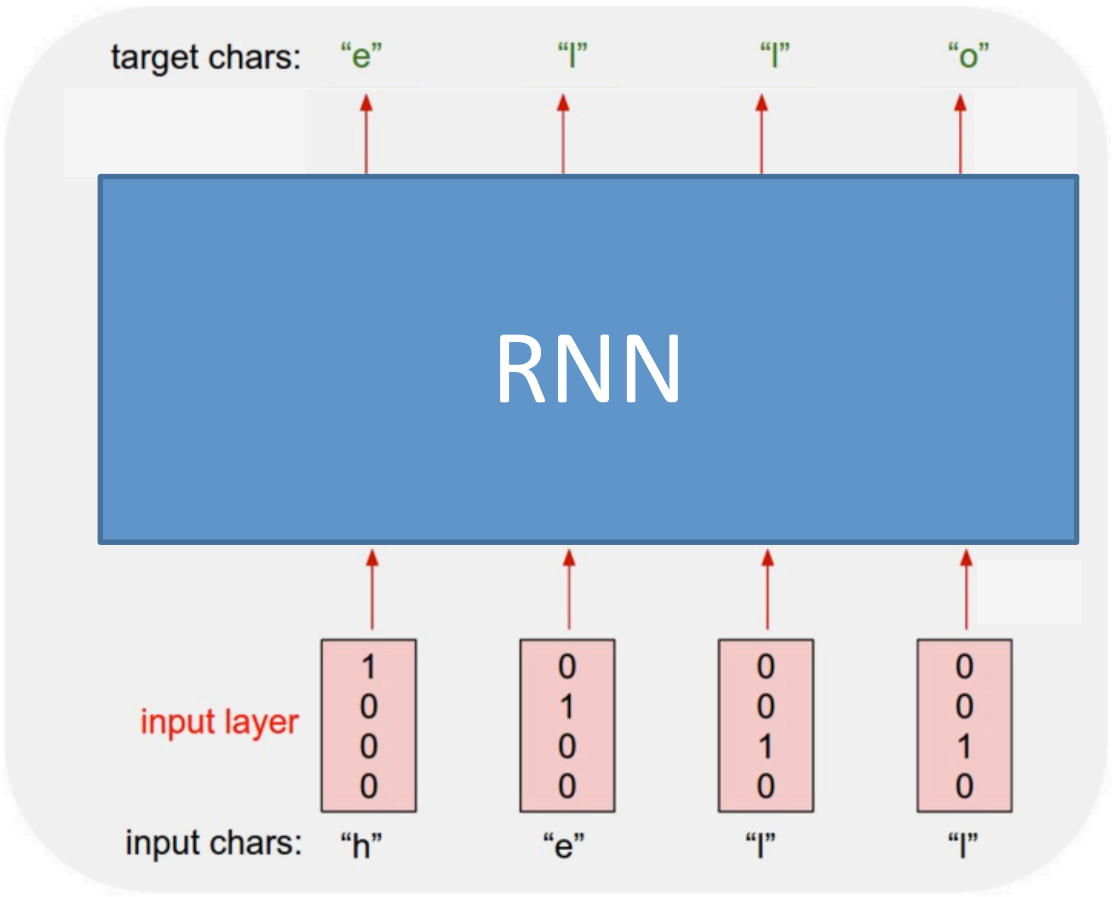
(Picture adapted from Andrej Karparthy)

RNN – *Input Layer*

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



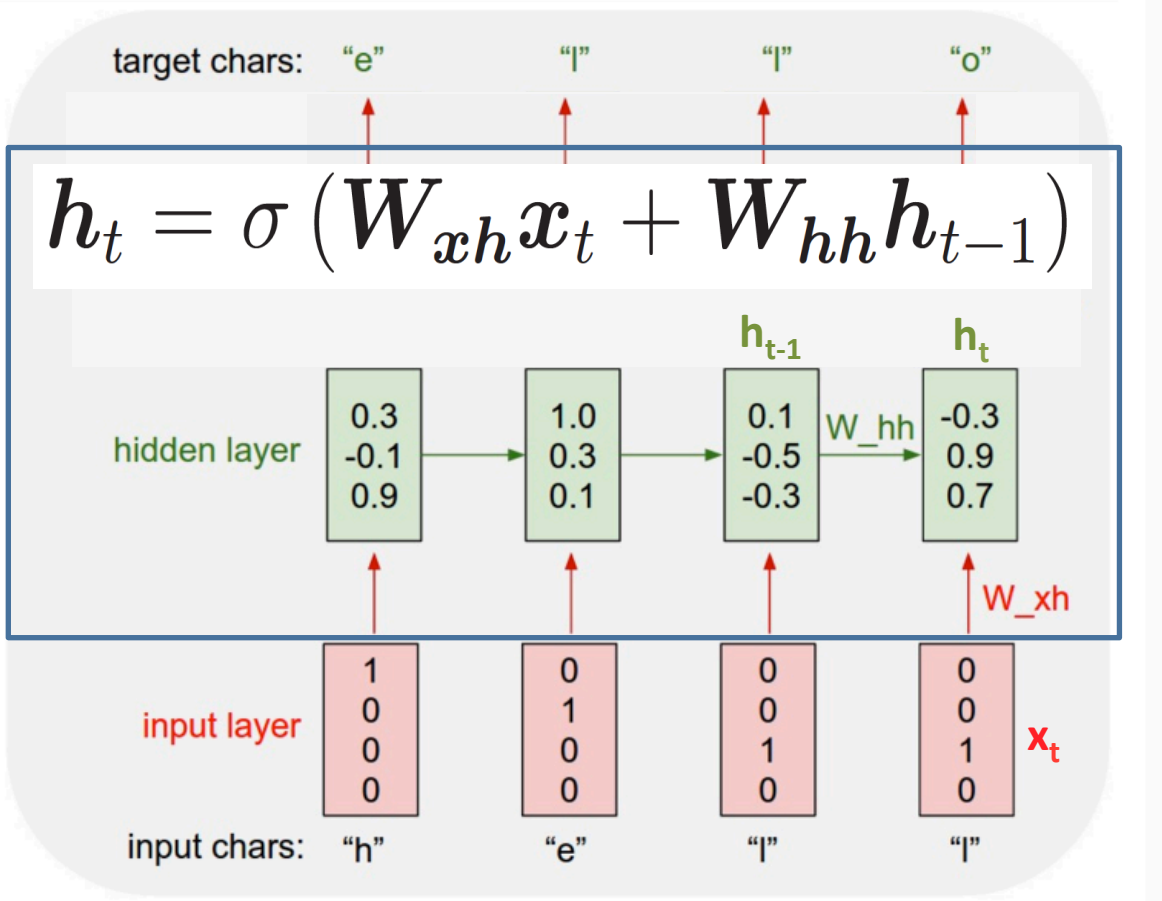
(Picture adapted from Andrej Karparthy)

RNN – *Hidden Layer*

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



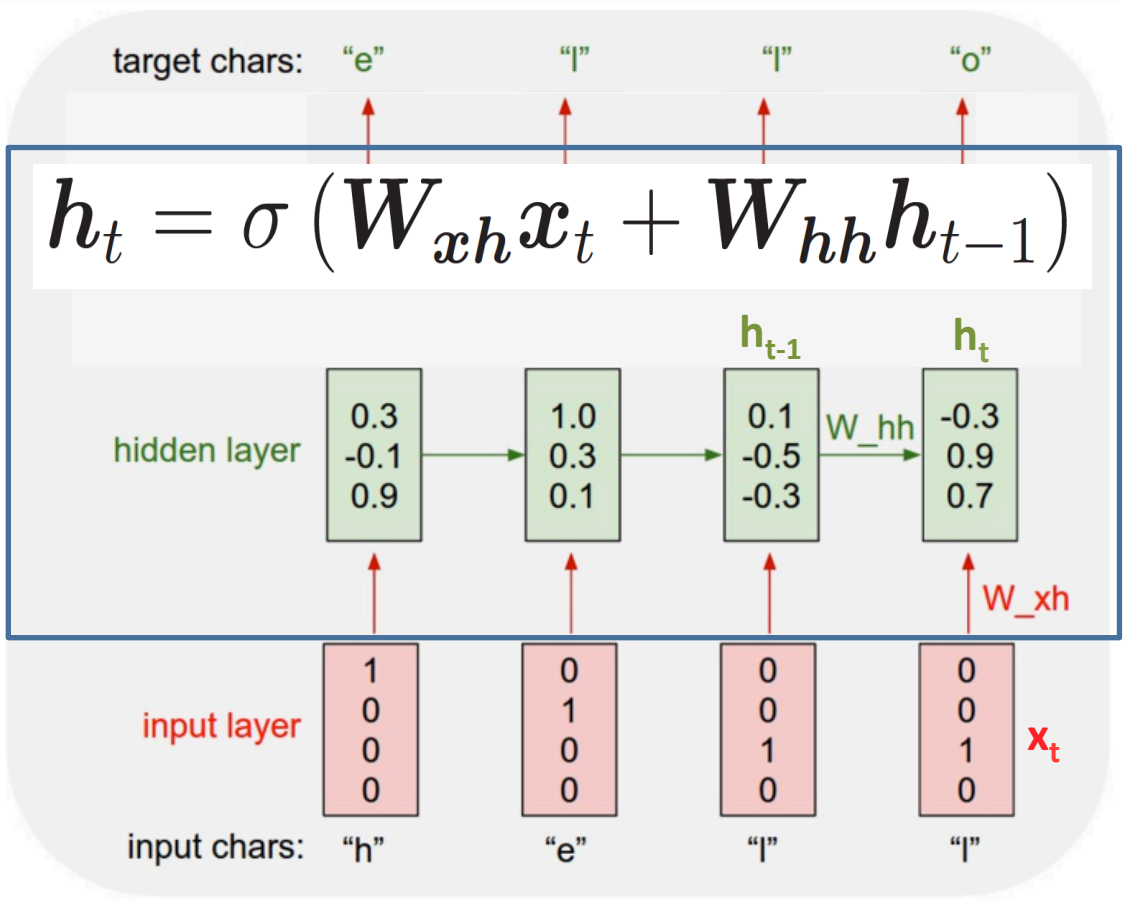
(Picture adapted from Andrej Karparthy)

RNN – *Hidden Layer*

Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



RNNs to represent sequences!

Outline

- Recurrent Neural Networks (RNNs)

- NMT basics (Sutskever et al., 2014)
  - Encoder-Decoder.
  - Training vs. Testing.
  - Backpropagation.
  - More about RNNs.

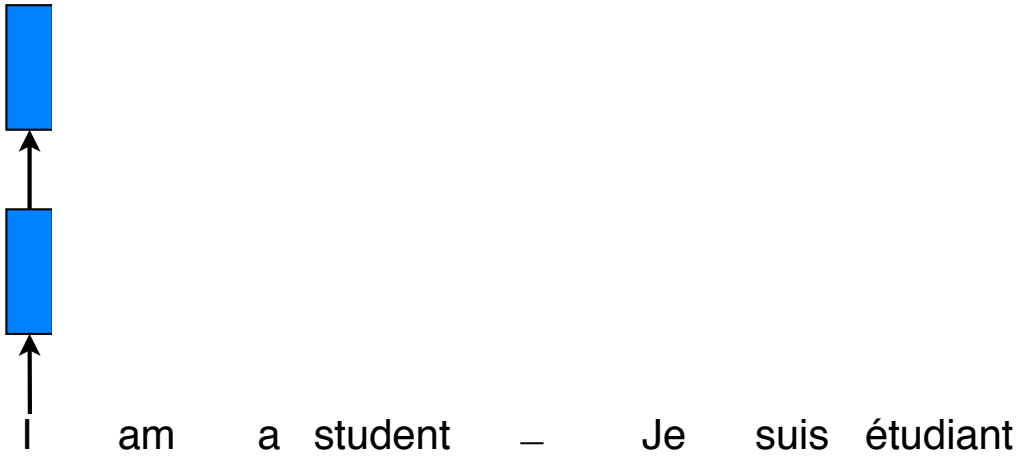
- Attention mechanism (Bahdanau et al., 2015)

Neural Machine Translation (NMT)

I am a student – Je suis étudiant

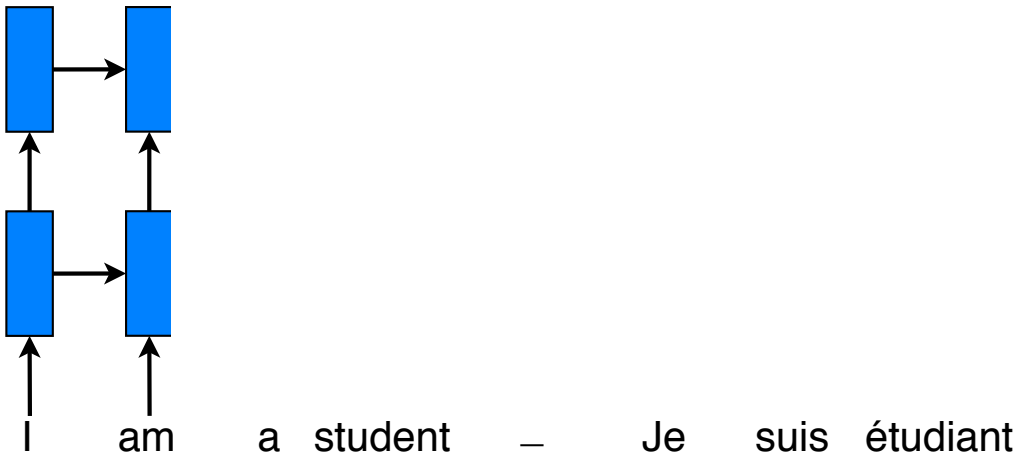
- Model P(target | source) directly.

Neural Machine Translation (NMT)



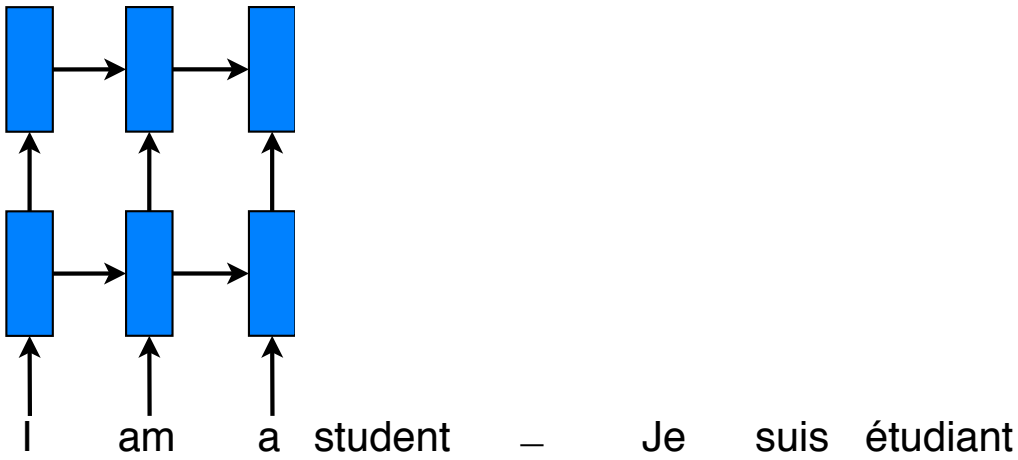
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



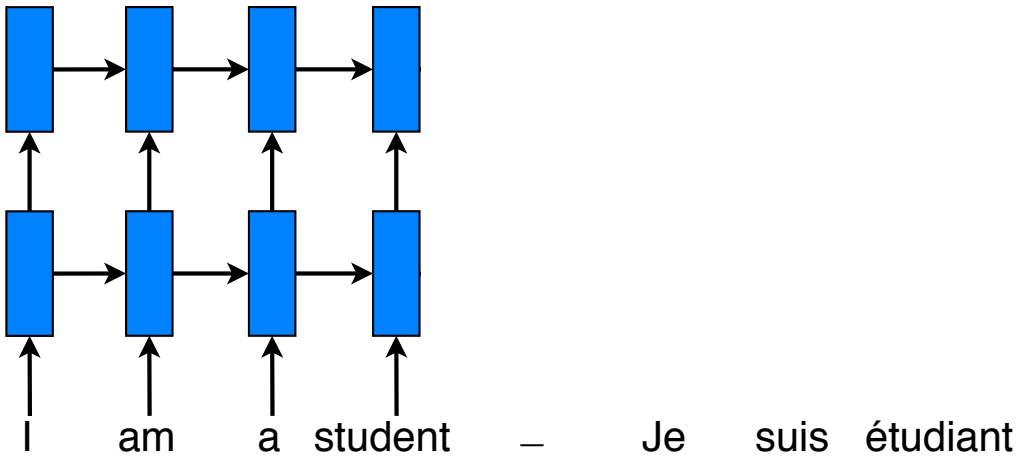
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



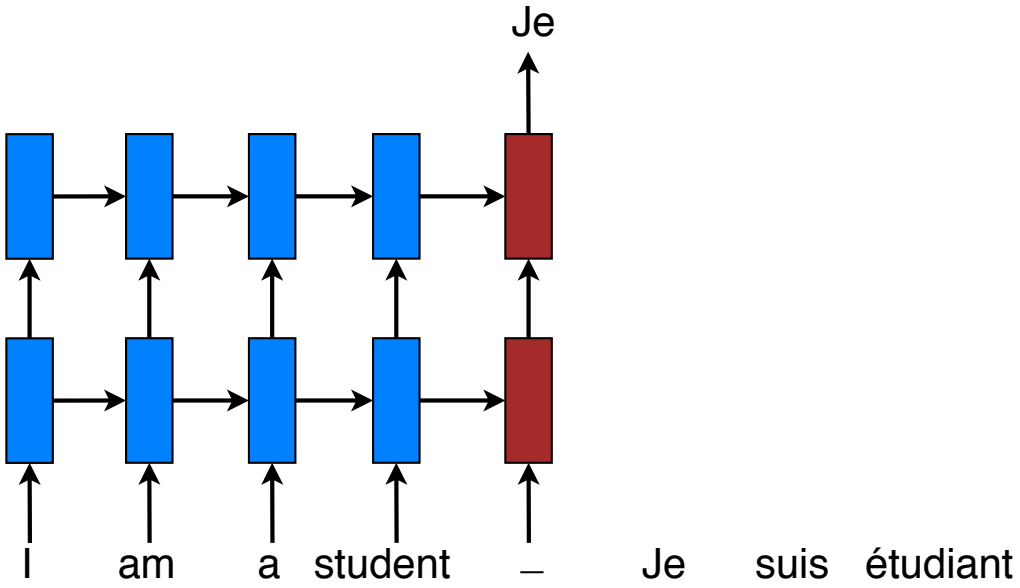
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



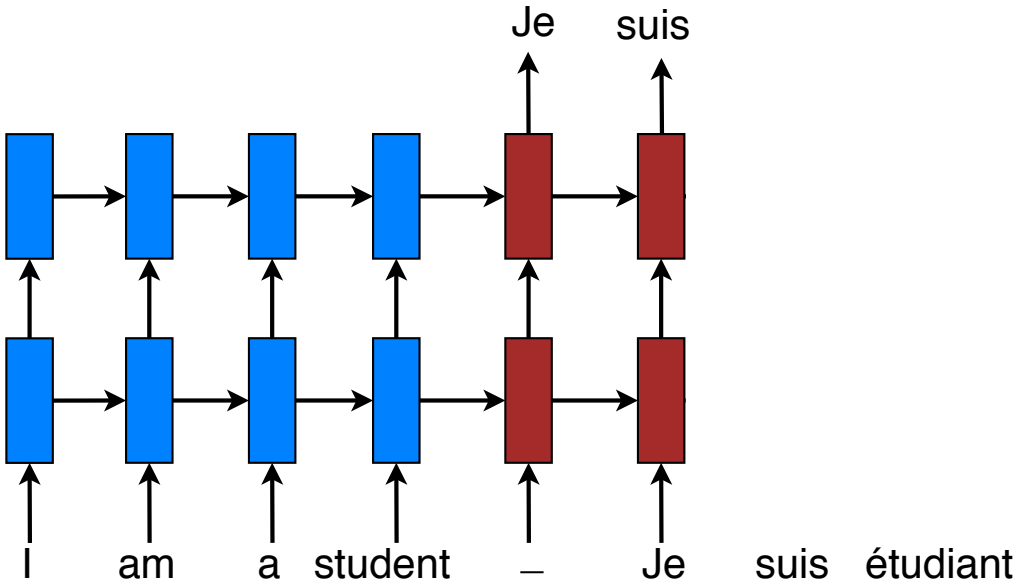
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



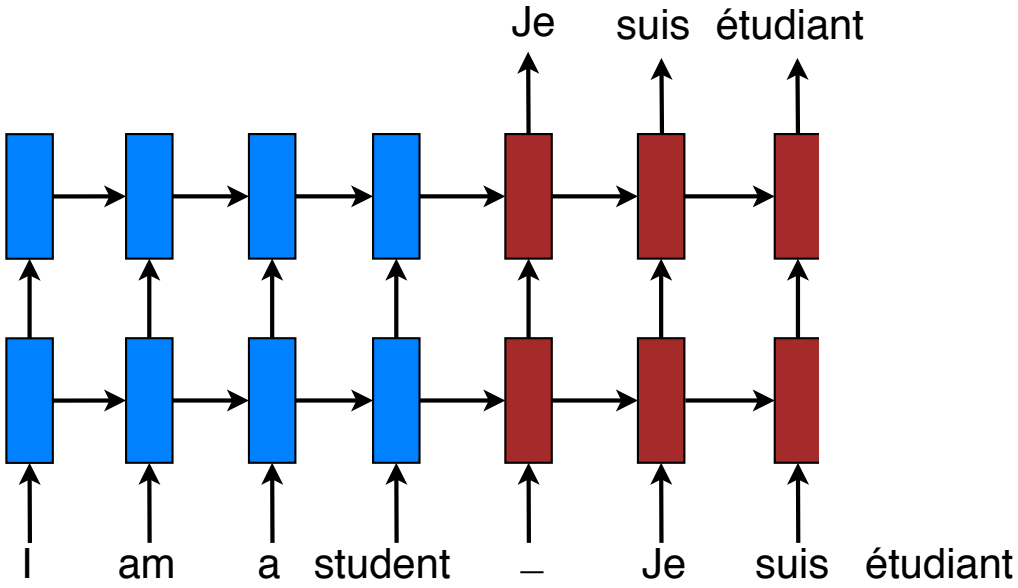
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



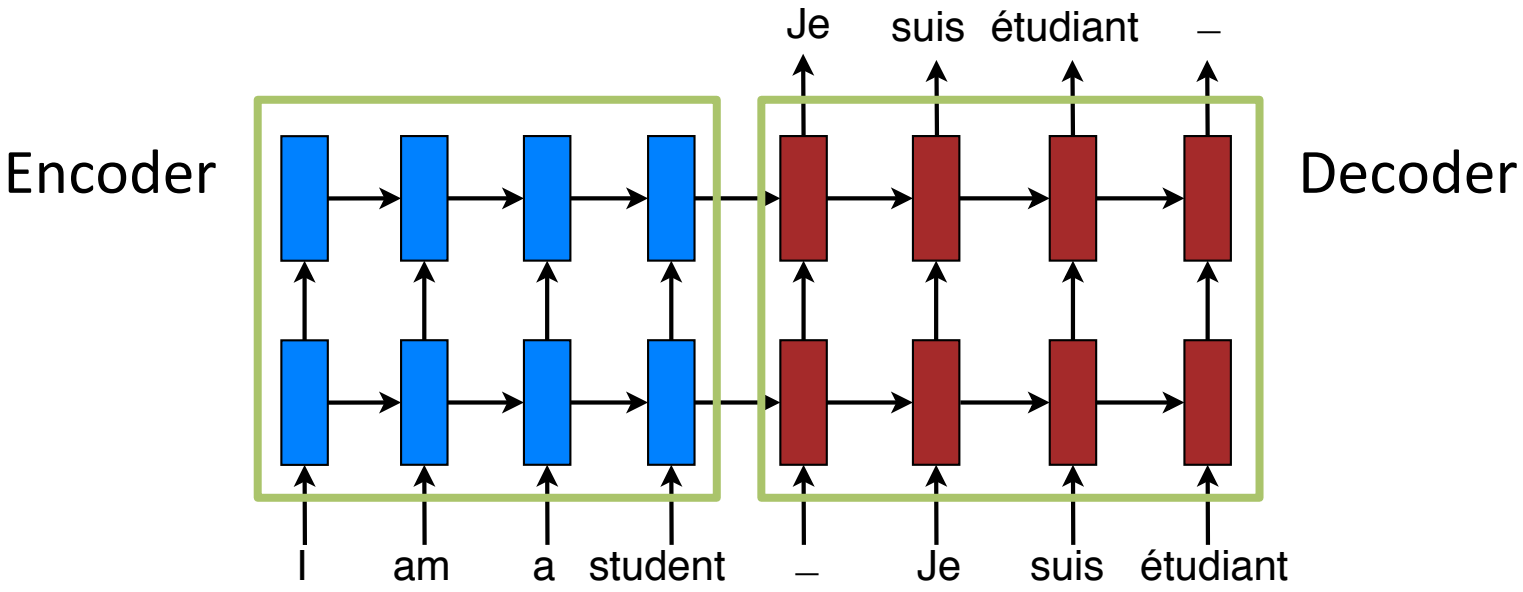
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



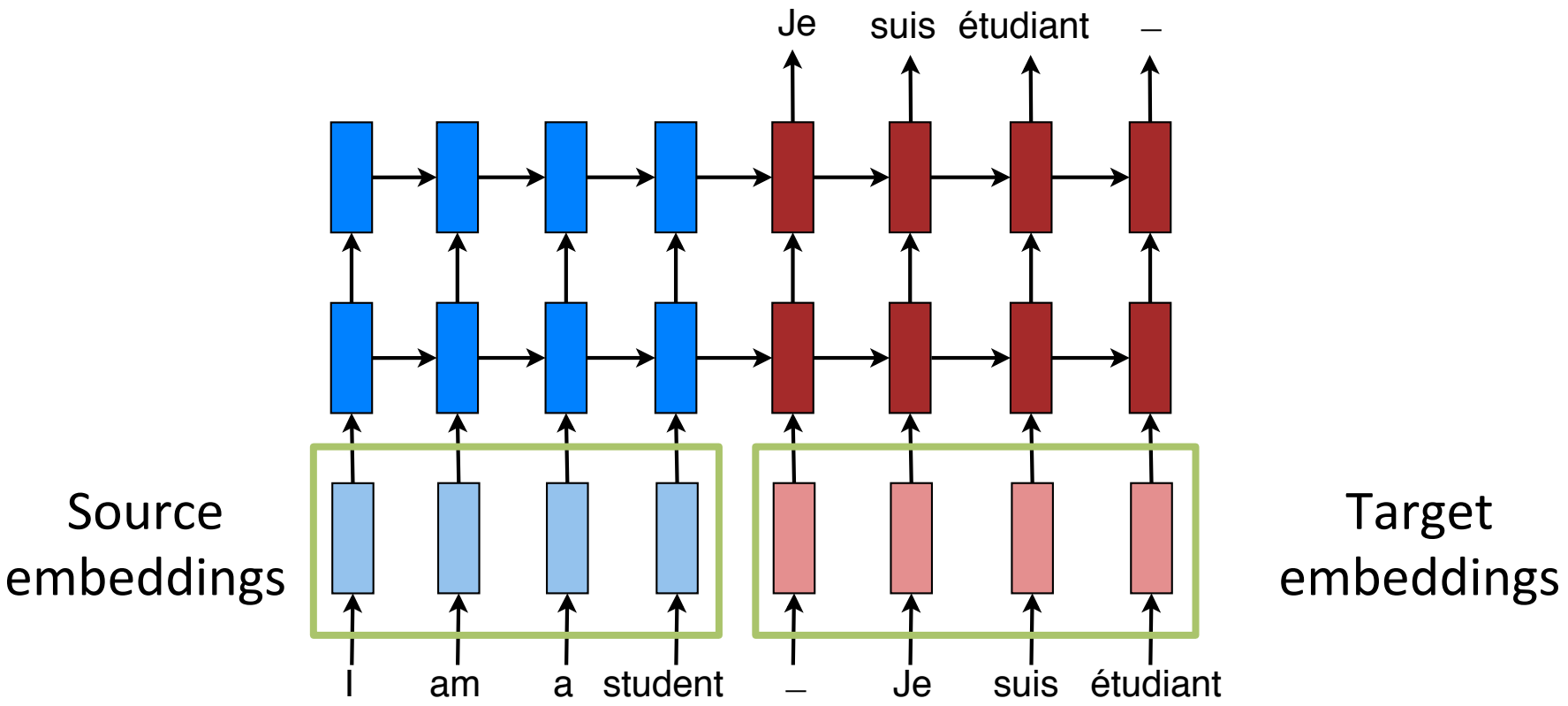
- RNNs trained end-to-end (Sutskever et al., 2014).

Neural Machine Translation (NMT)



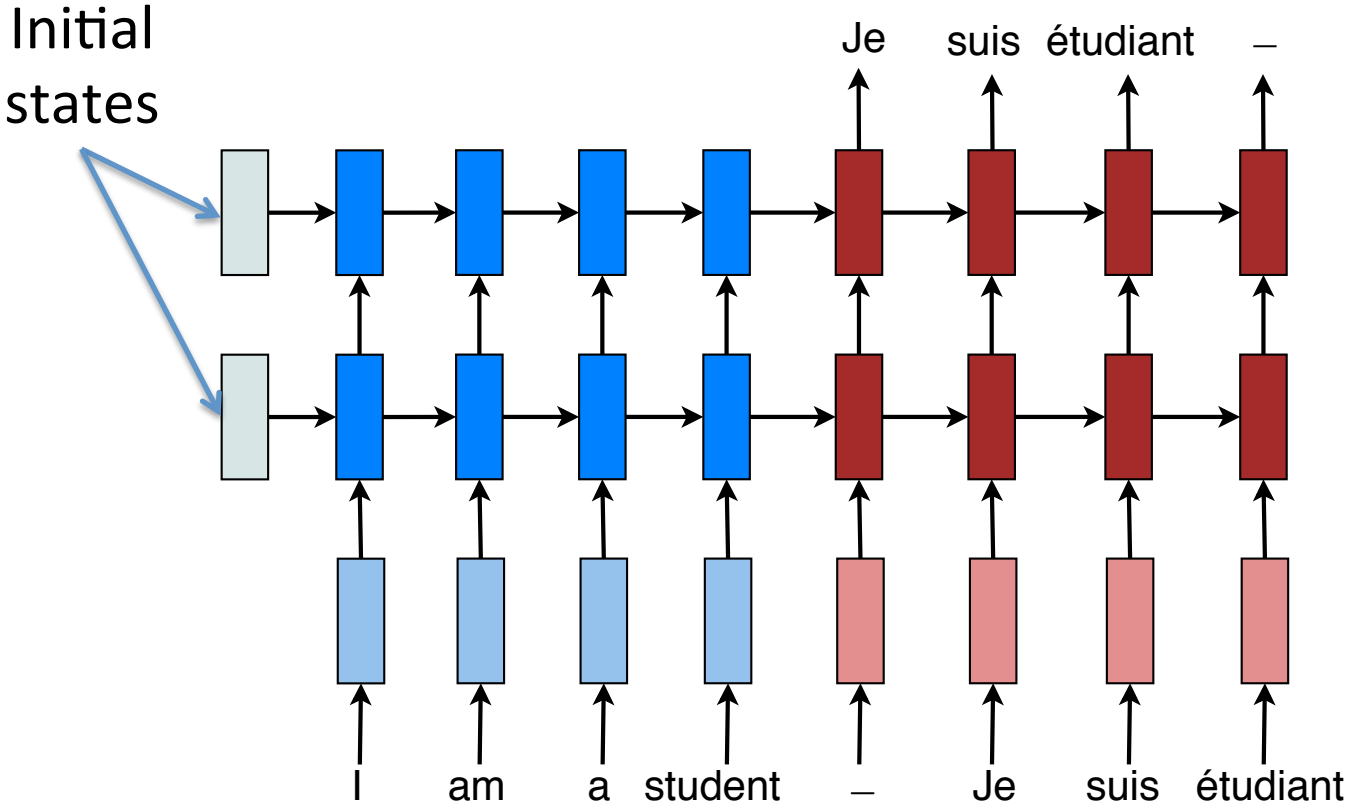
- RNNs trained end-to-end (Sutskever et al., 2014).
- Encoder-decoder approach.

Word Embeddings



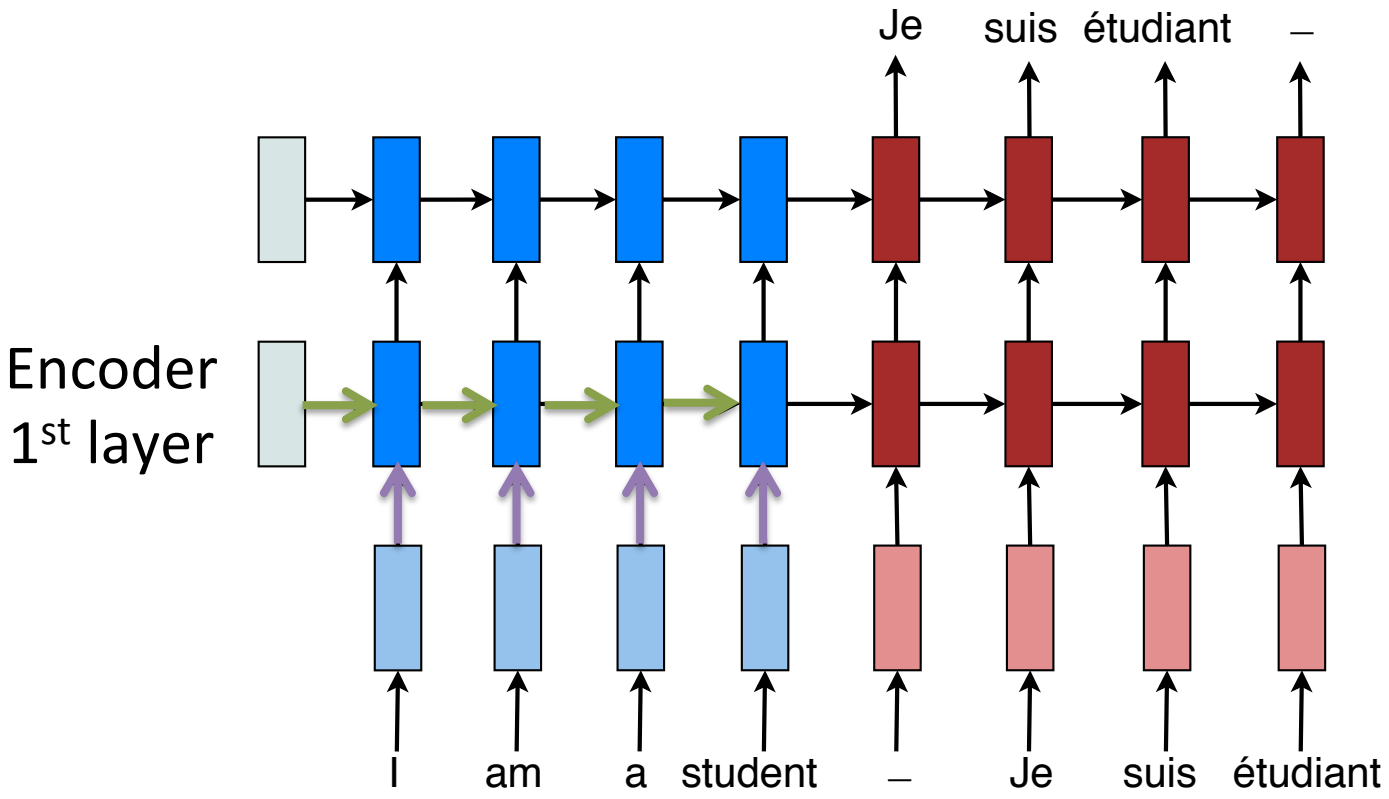
- Randomly initialized, one for each language.
  - Learnable parameters.

Recurrent Connections



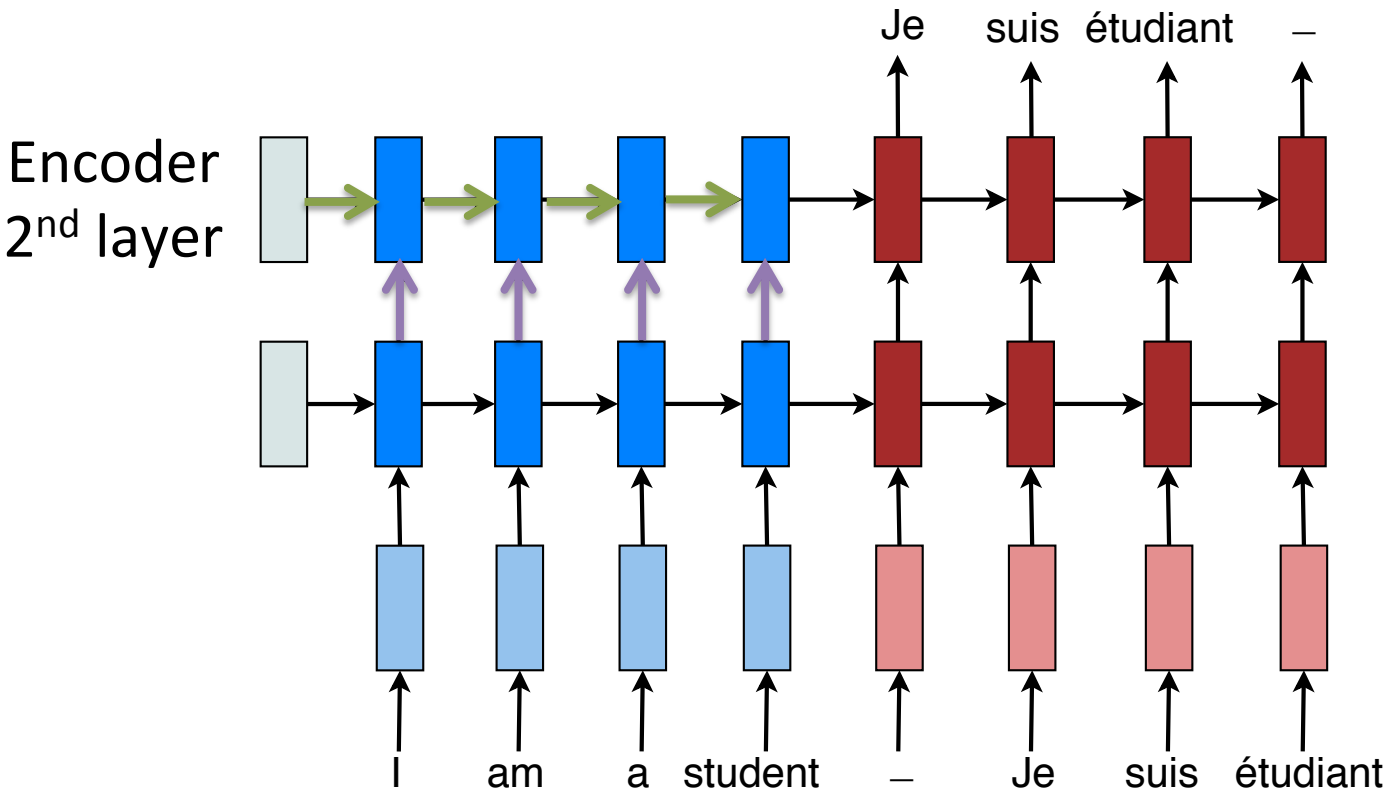
- Often set to 0.

Recurrent Connections



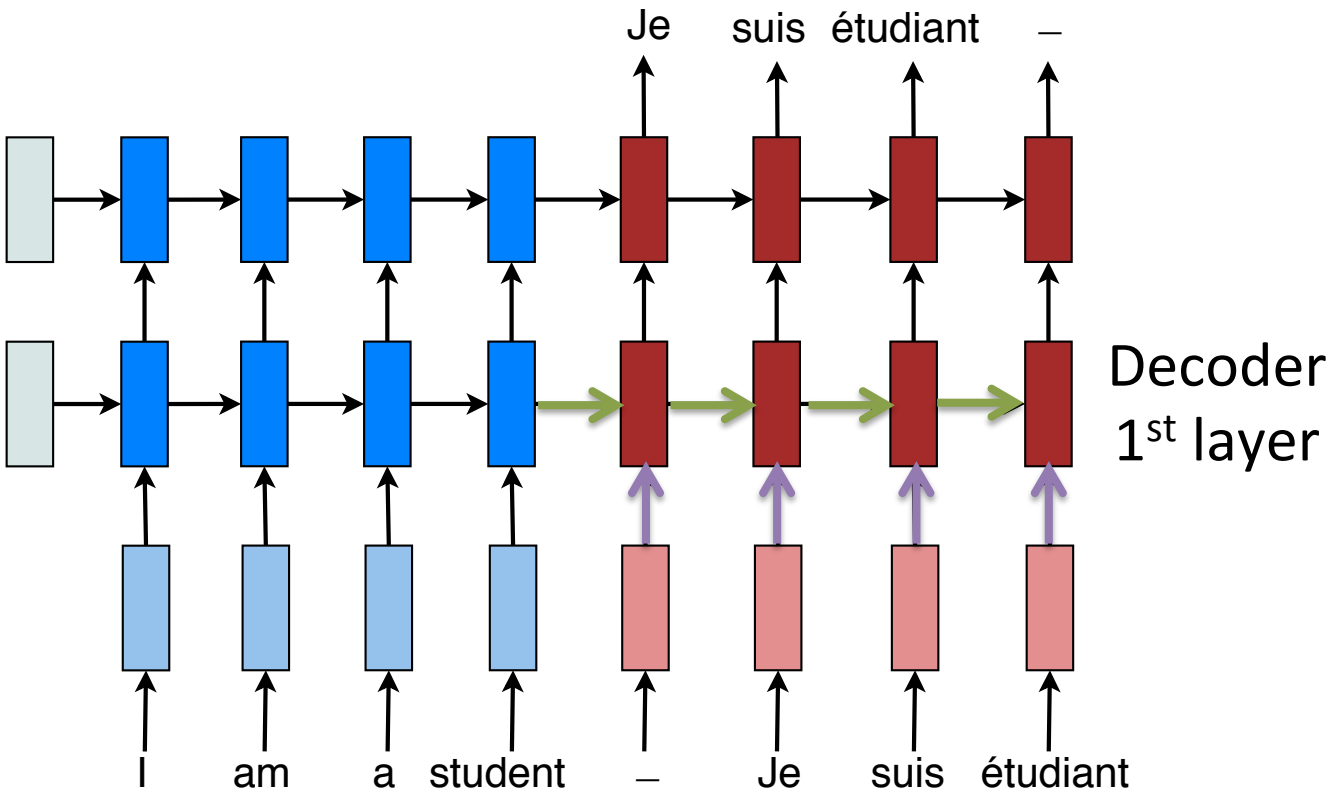
- Different across layers and encoder / decoder.

Recurrent Connections



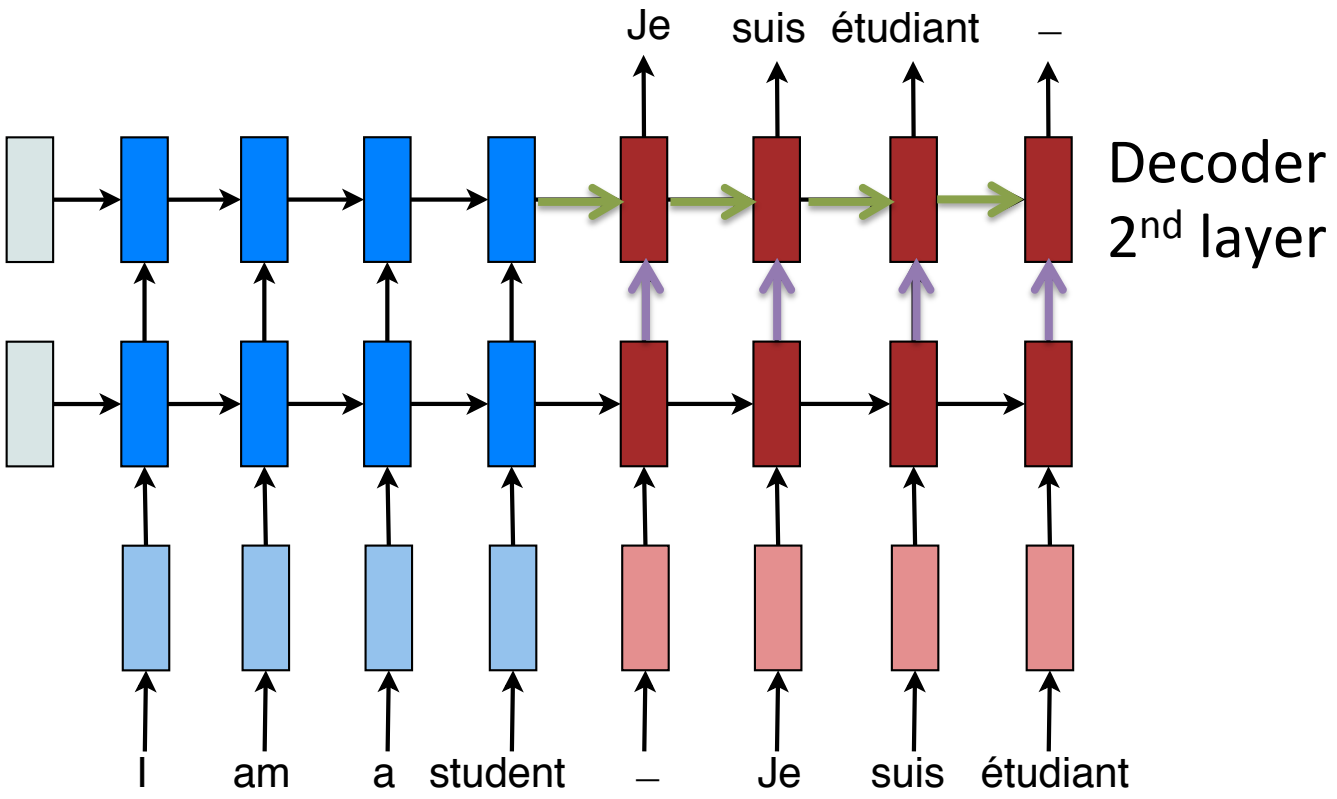
- Different across layers and encoder / decoder.

Recurrent Connections



- Different across layers and encoder / decoder.

Recurrent Connections



- Different across layers and encoder / decoder.

Outline

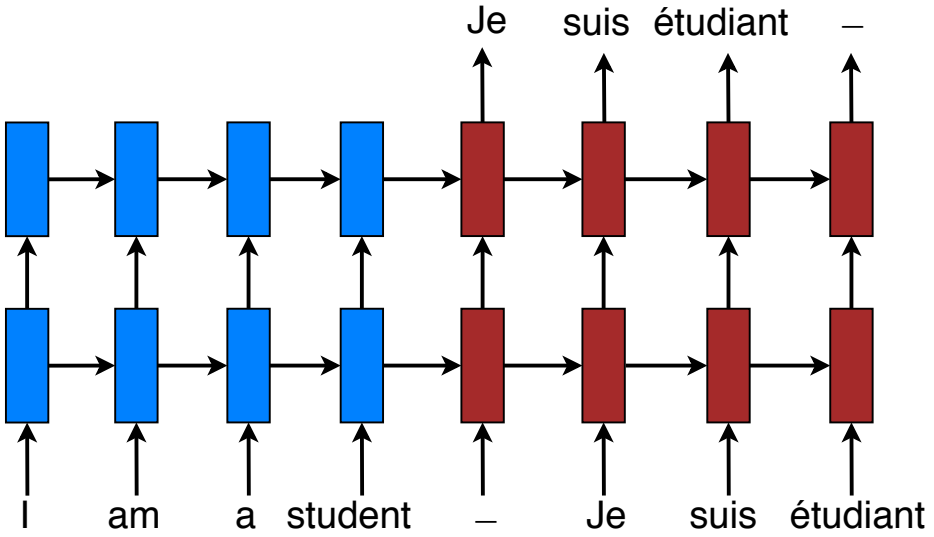
- Recurrent Neural Networks (RNNs)

- NMT basics (Sutskever et al., 2014)
  - Encoder-Decoder.
  - Training vs. Testing.
  - Backpropagation.
  - More about RNNs.

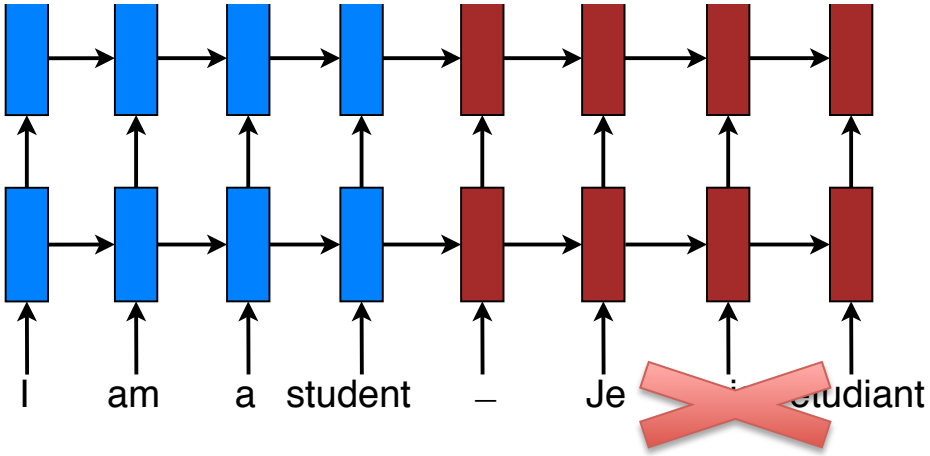
- Attention mechanism (Bahdanau et al., 2015)

Training vs. Testing

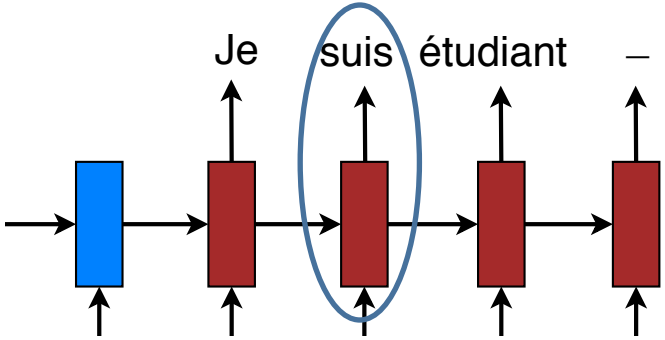
- *Training*
  - Correct translations are available.

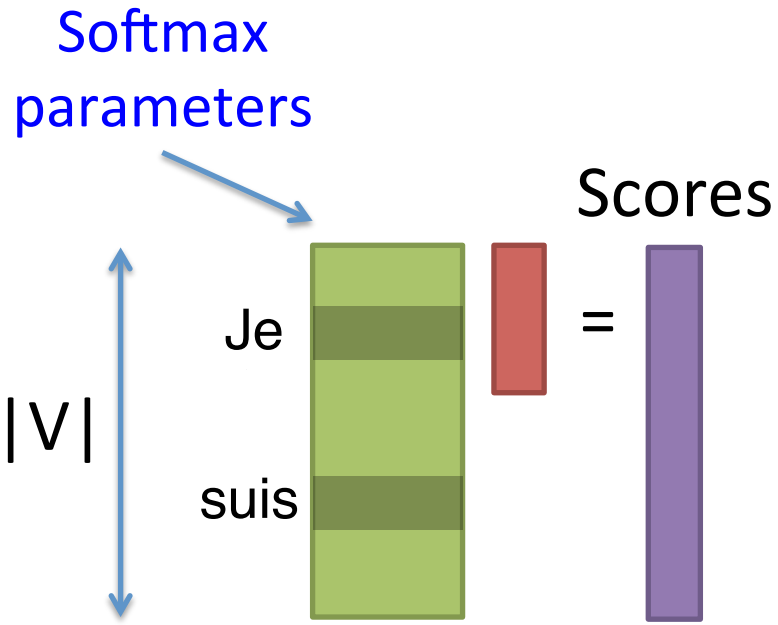


- *Testing*
  - Only source sentences are given.



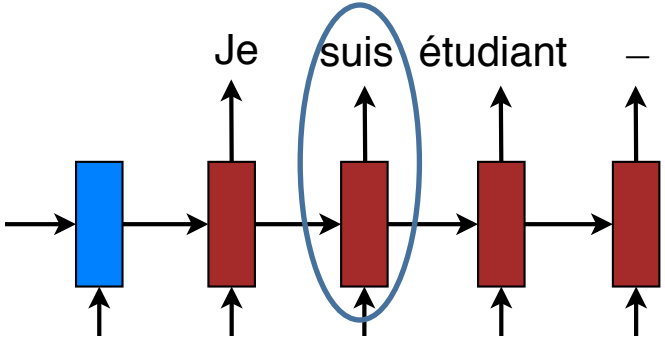
Training – *Softmax*

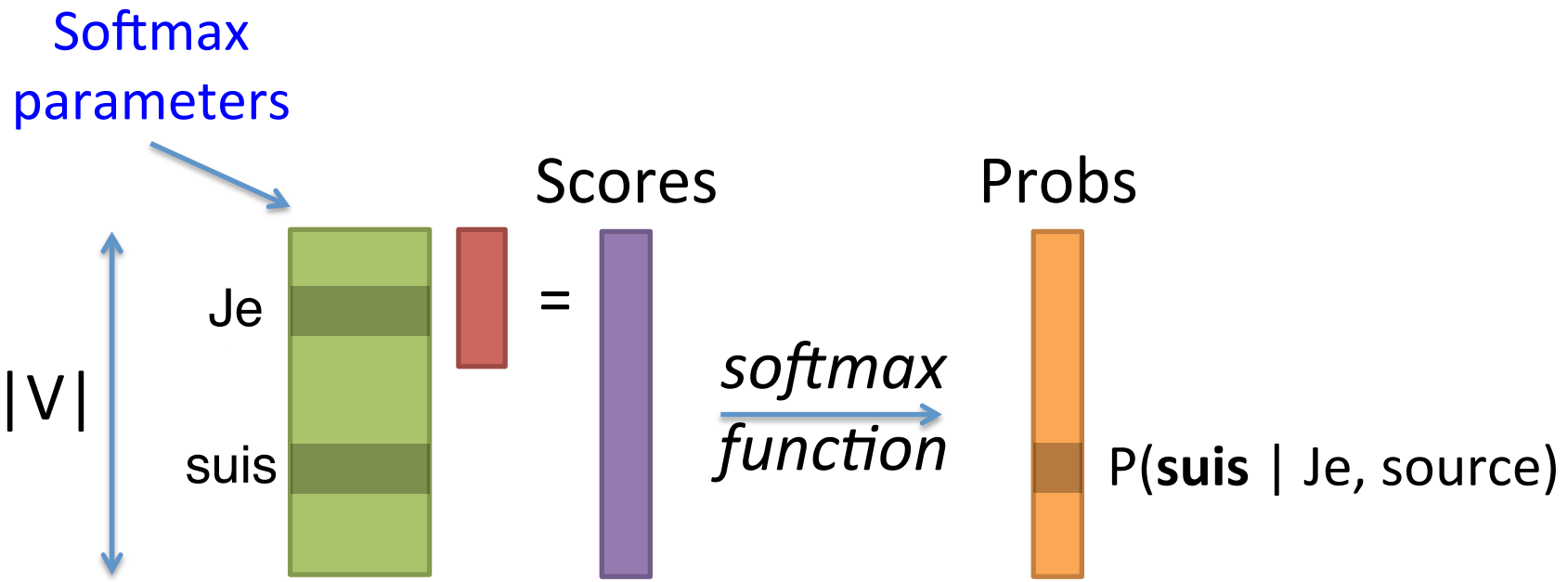




- Hidden states ↦ scores.

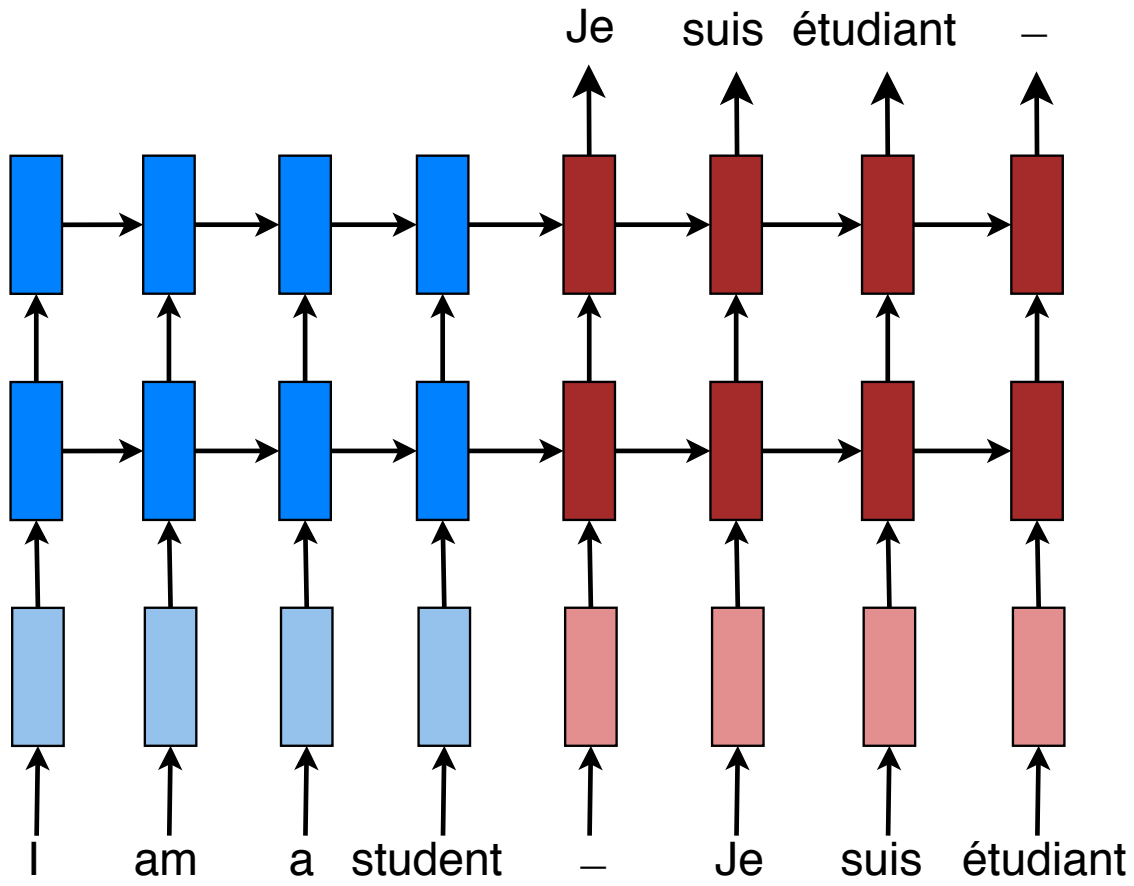
Training – *Softmax*





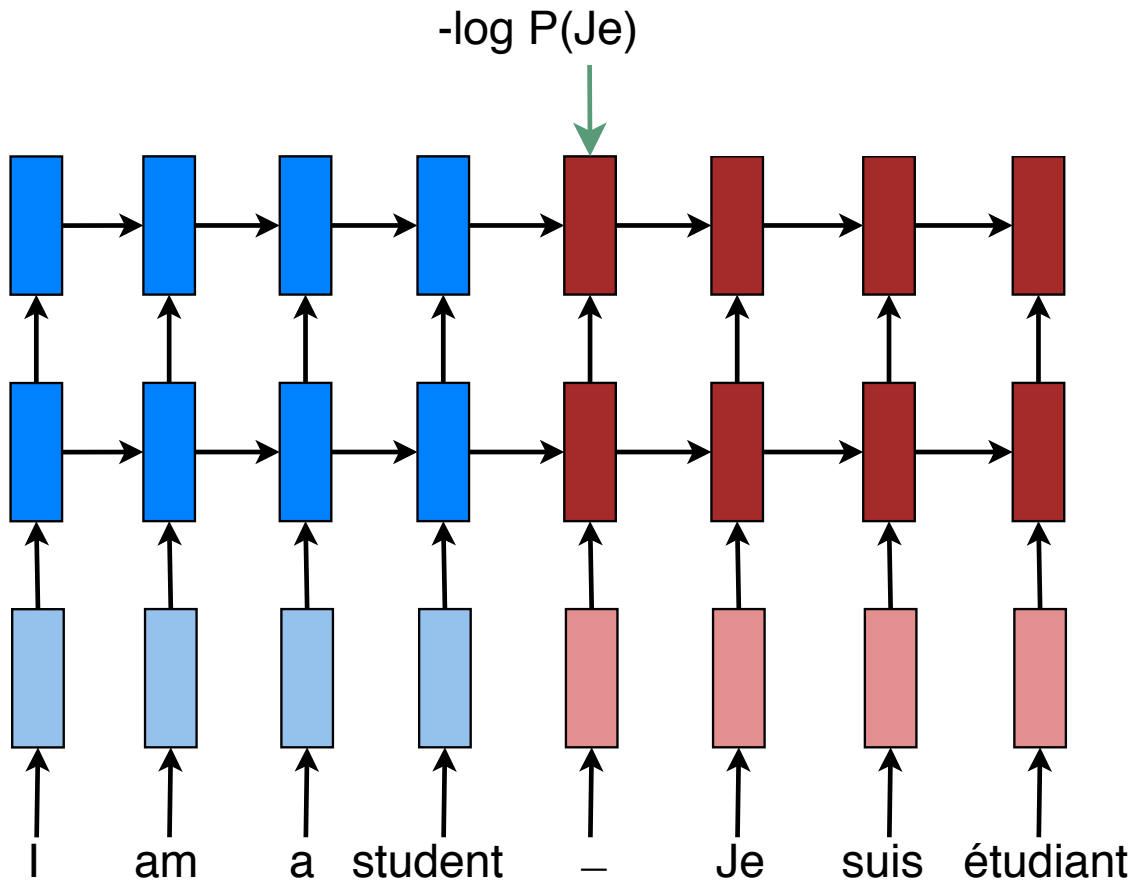
- Scores ↦ probabilities.

Training Loss



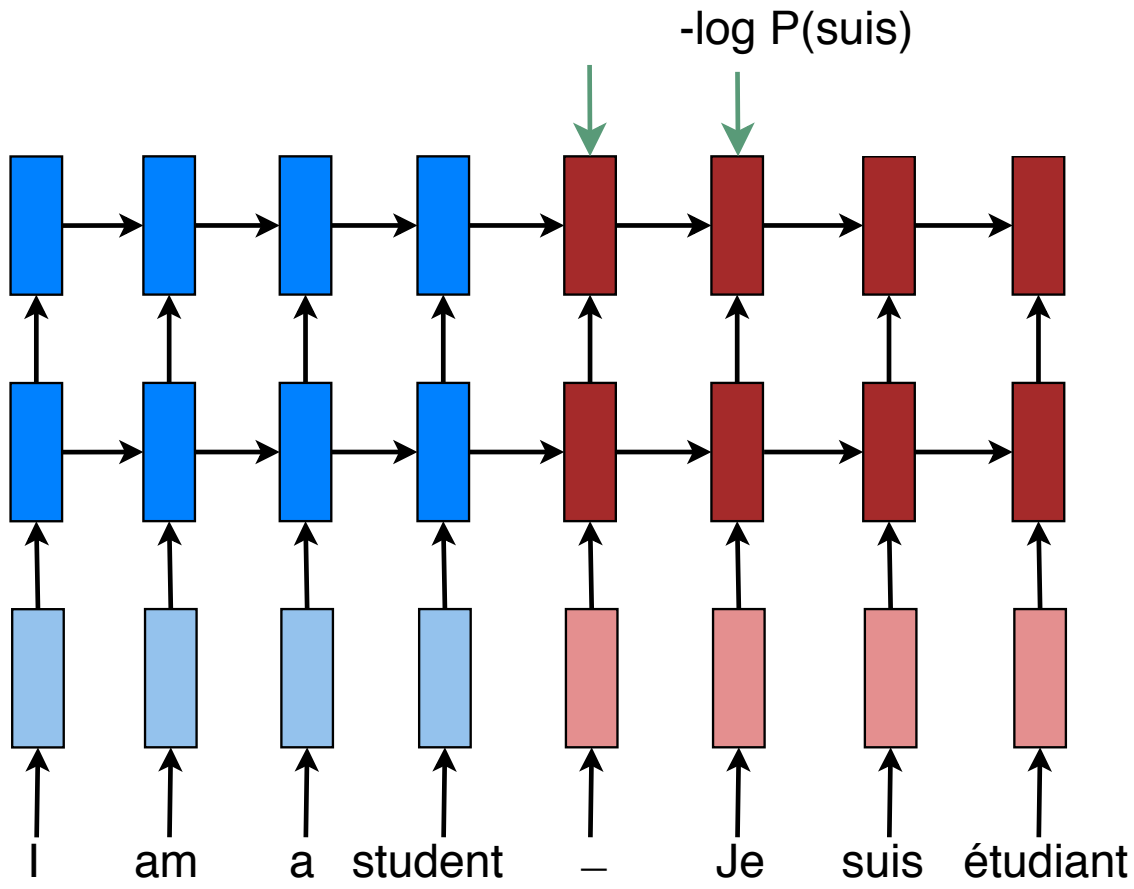
- Maximize P(target | source):
  - Decompose into individual word predictions.

Training Loss



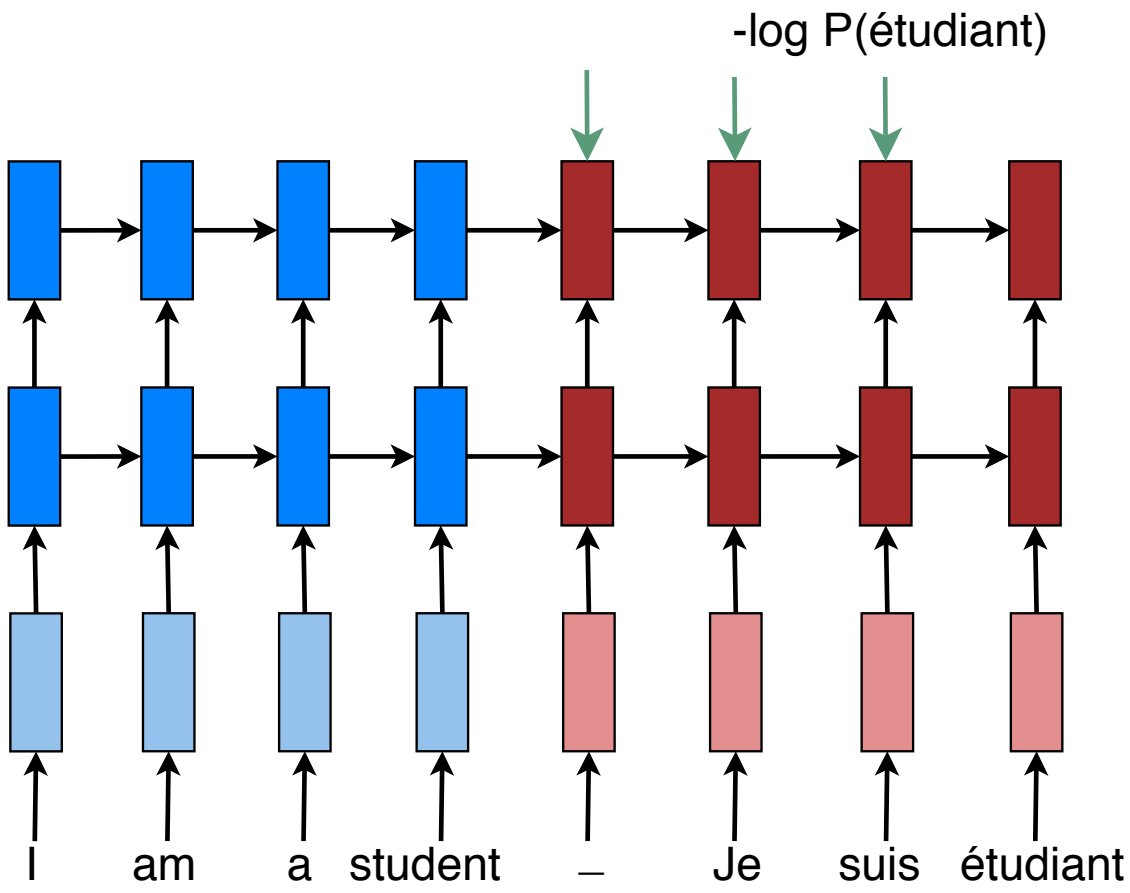
- Sum of all individual losses

Training Loss



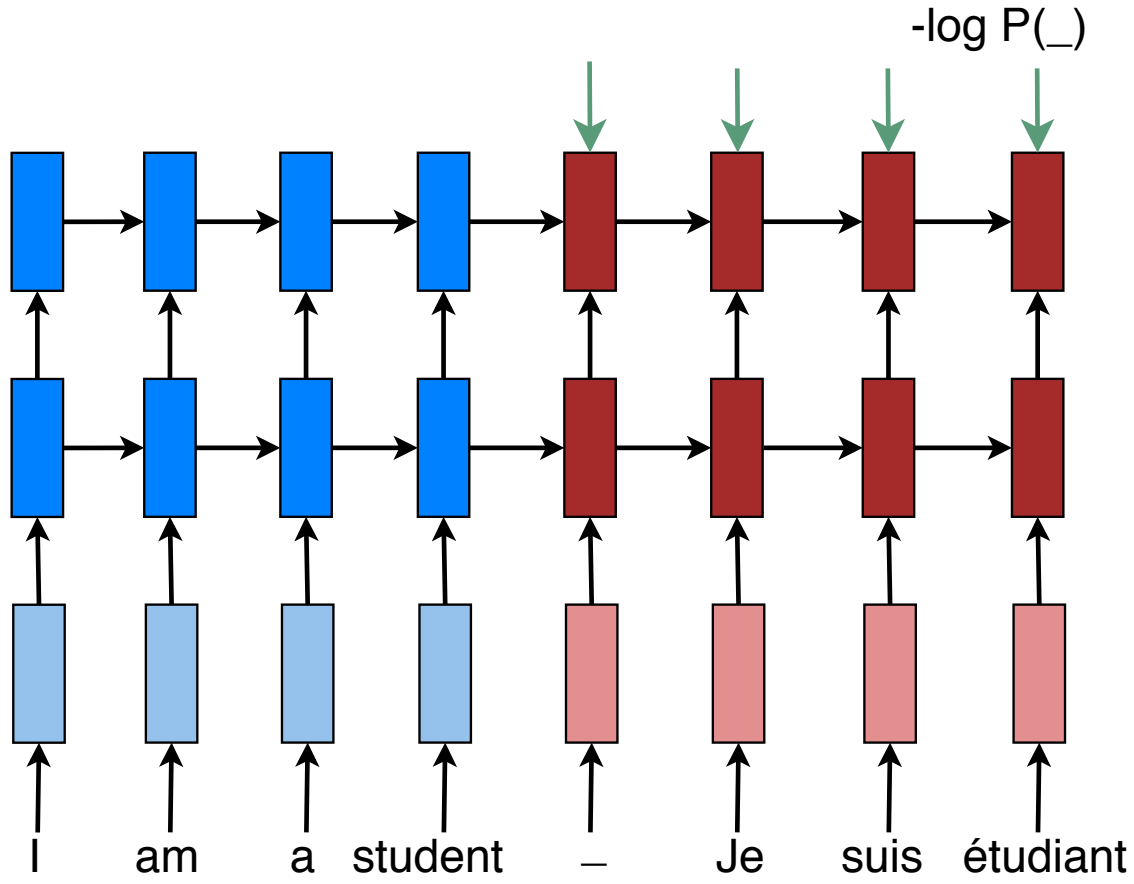
- Sum of all individual losses

Training Loss



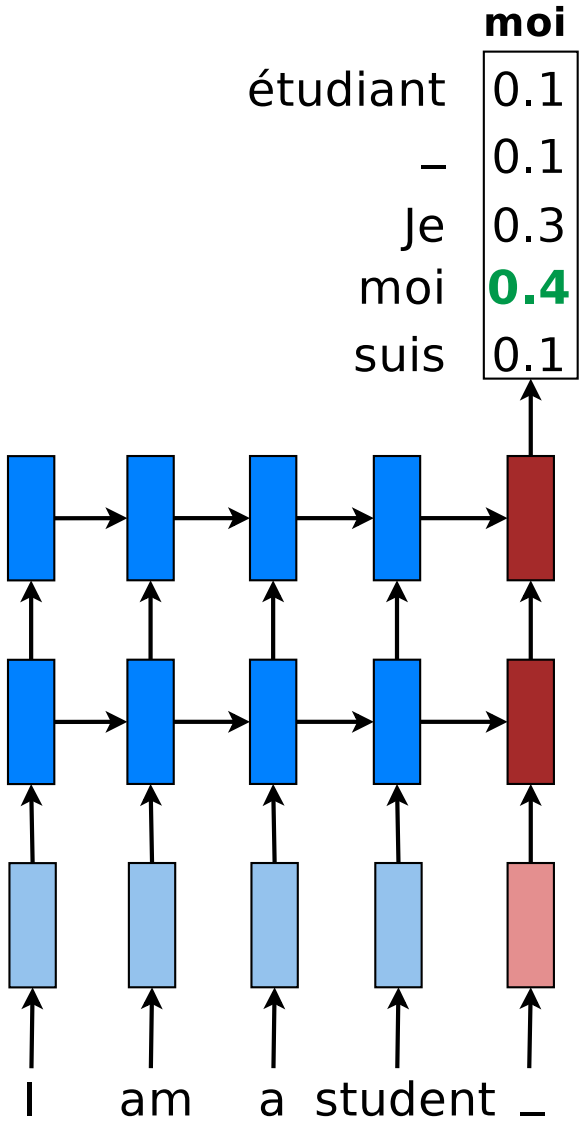
- Sum of all individual losses

Training Loss



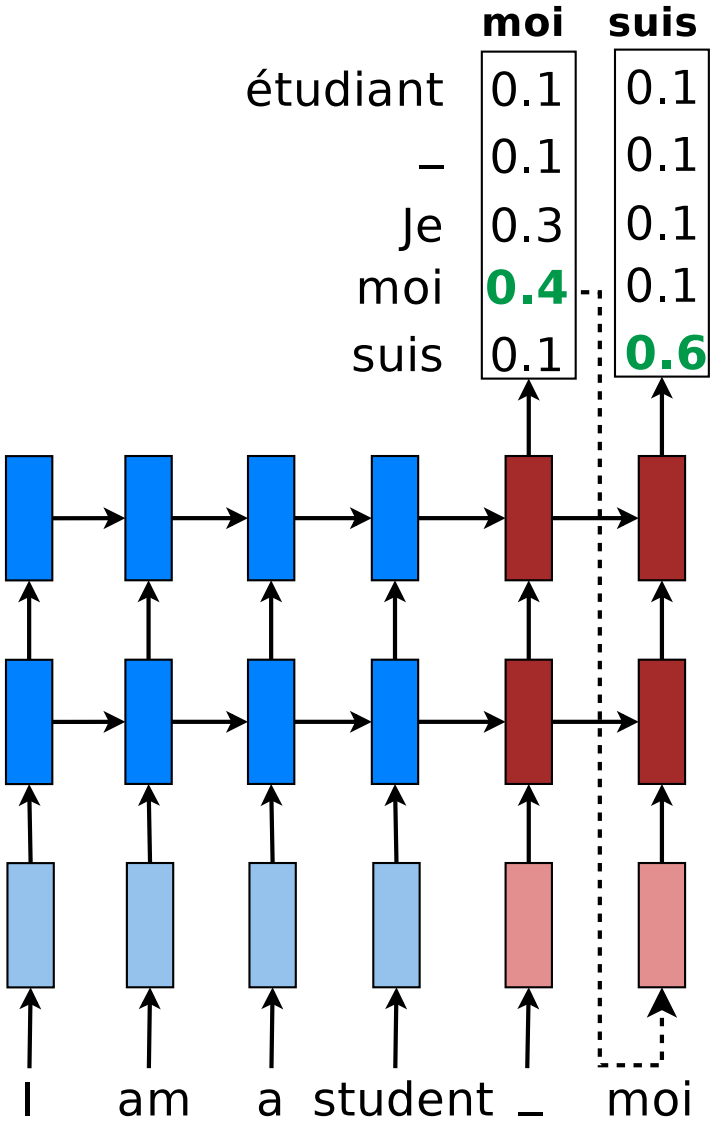
- Sum of all individual losses

Testing



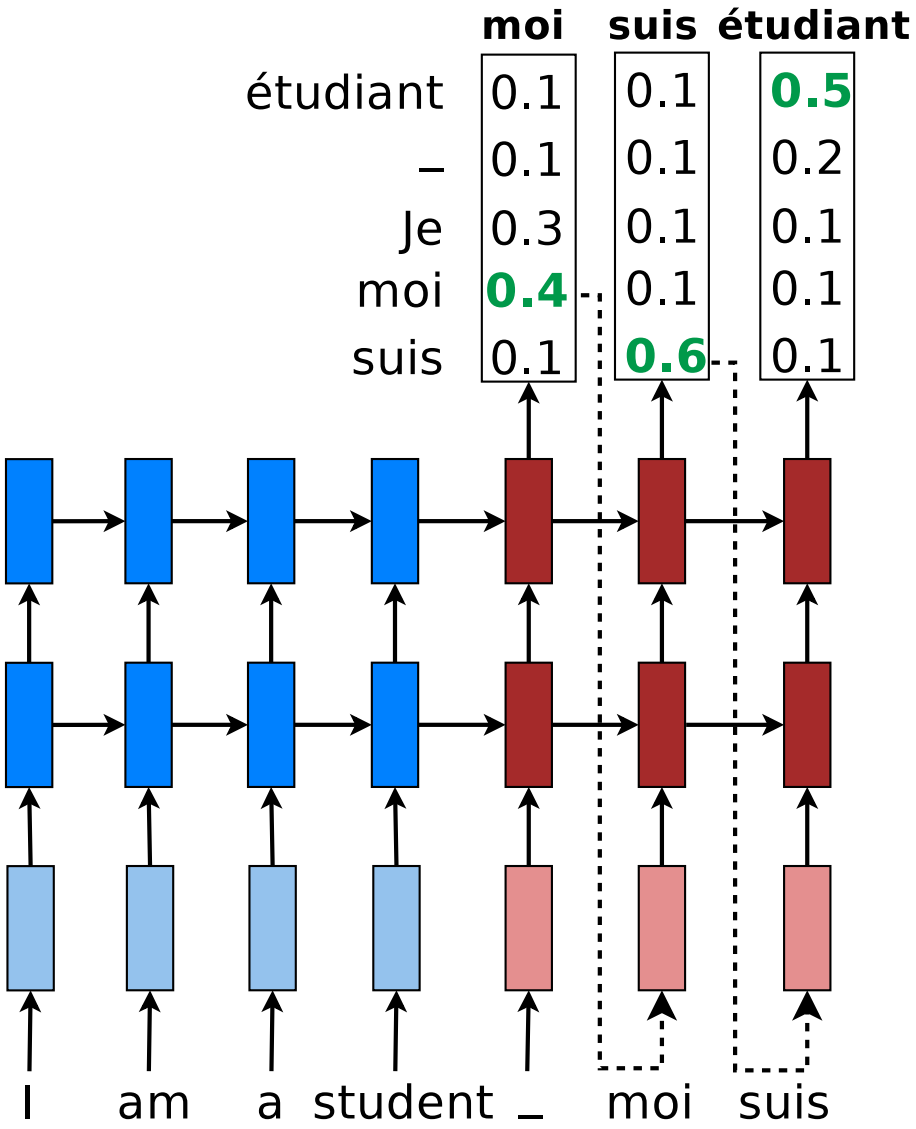
- Feed the most likely word

Testing



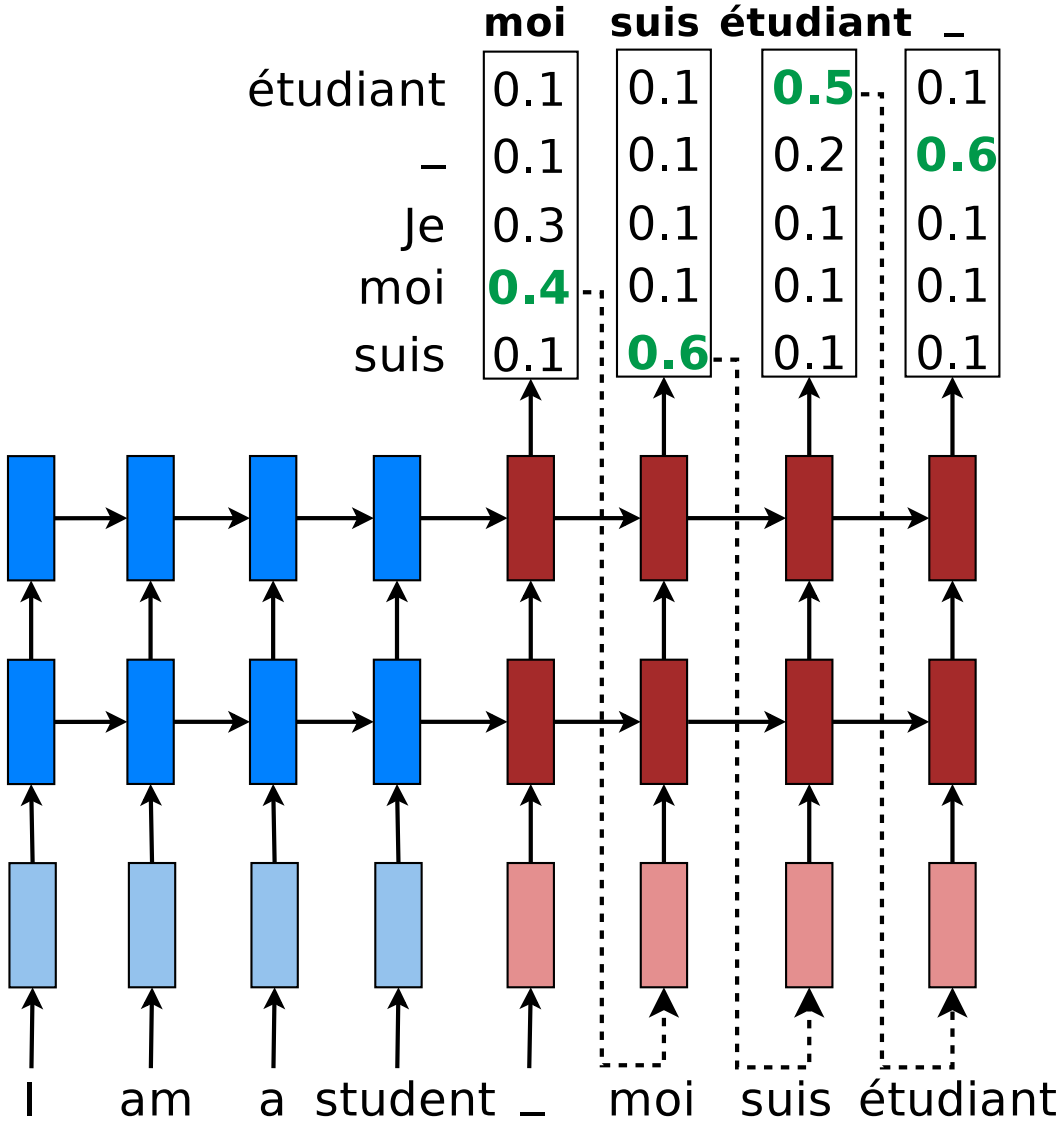
- Feed the most likely word

Testing



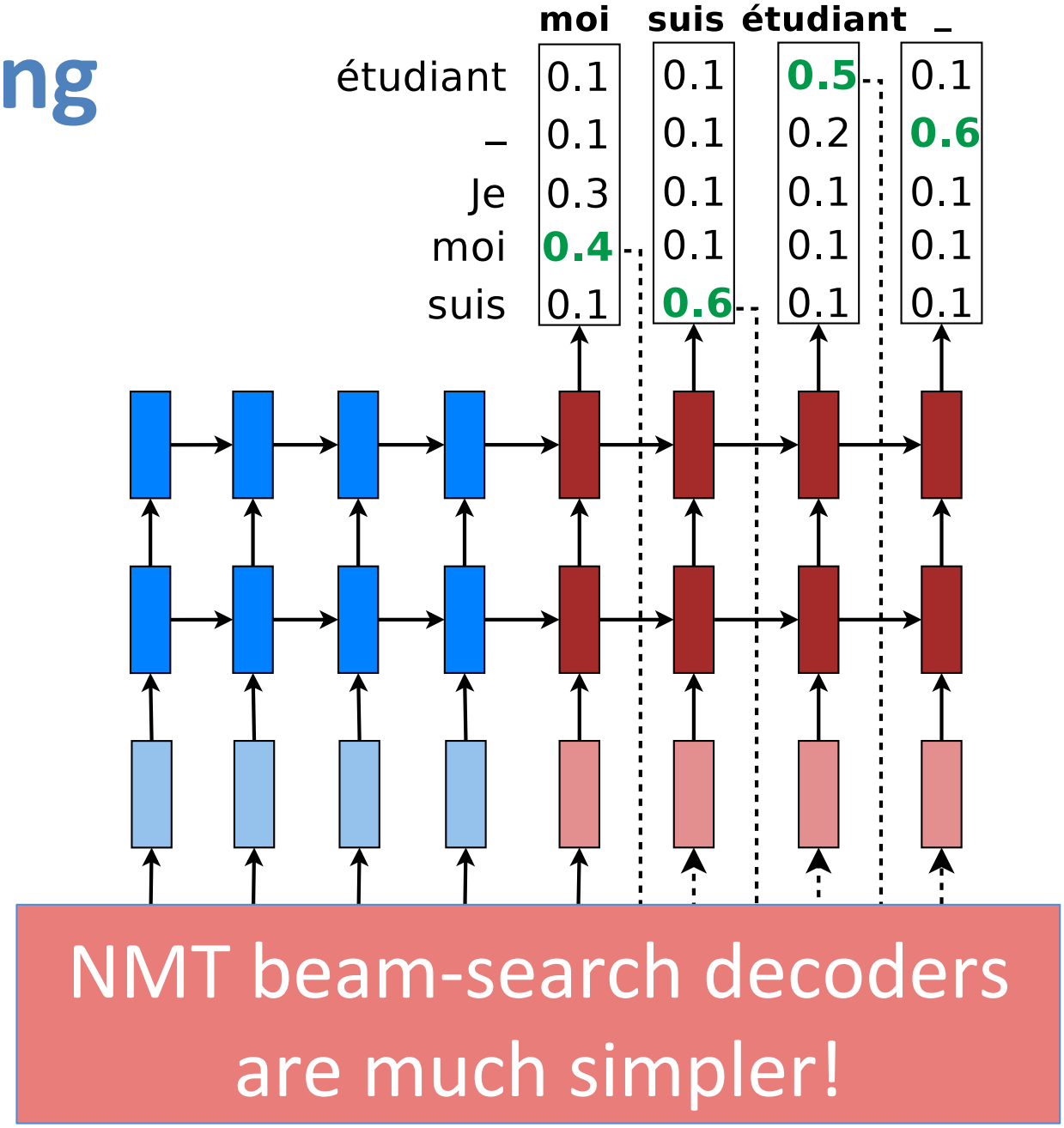
- Feed the most likely word

Testing



- Feed the most likely word

Testing


moi
suis
étudiant
_
étudiant 0.1 0.1 0.5 0.1
_ 0.1 0.1 0.2 0.6
Je 0.3 0.1 0.1 0.1
moi 0.4 0.1 0.1 0.1
suis 0.1 0.6 0.1 0.1
NMT beam-search decoders are much simpler!

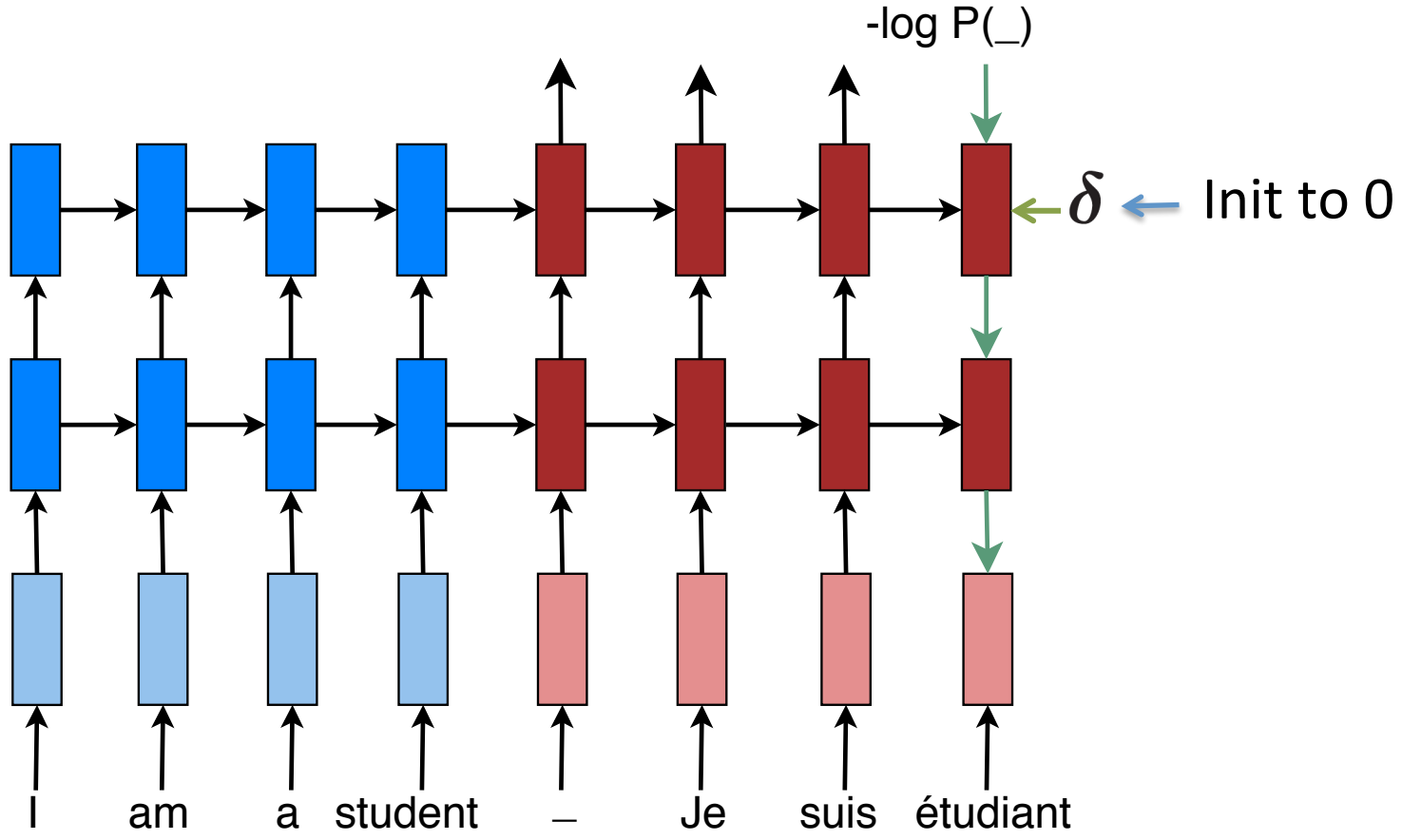
Outline

- Recurrent Neural Networks (RNNs)

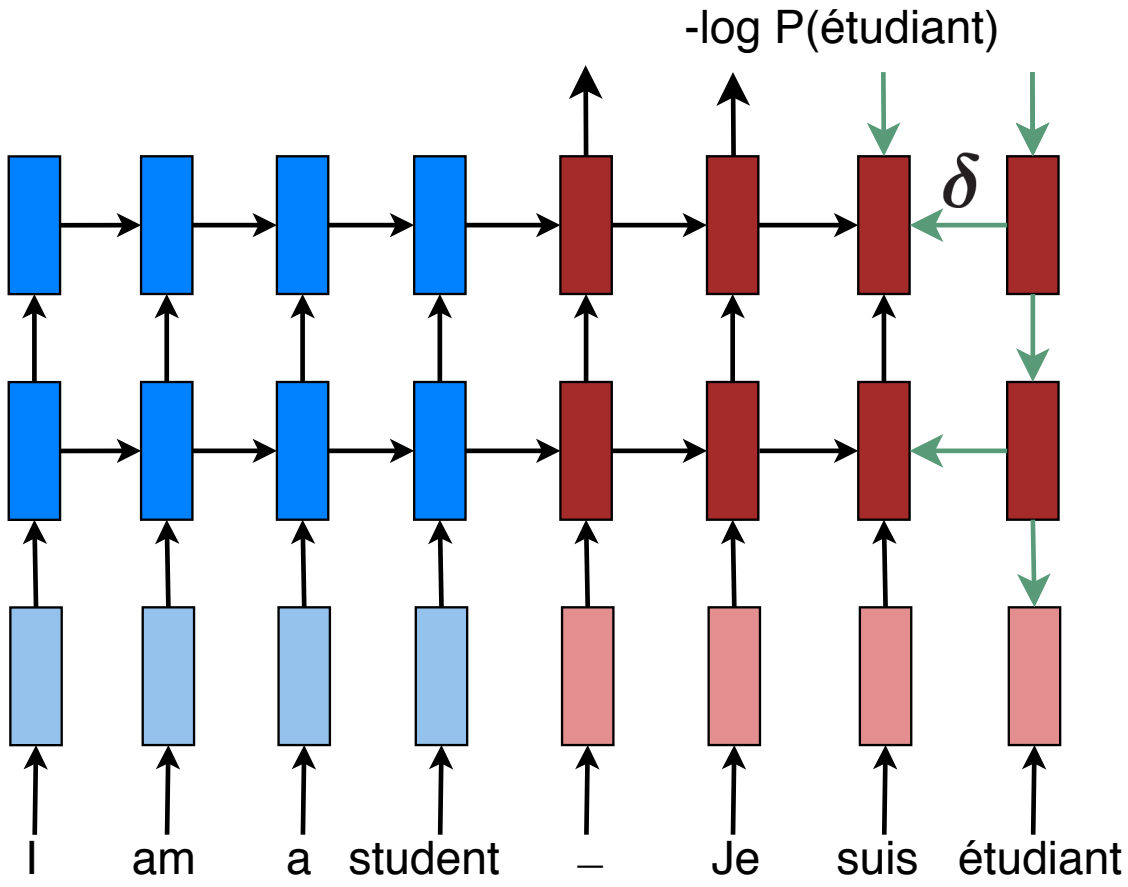
- NMT basics (Sutskever et al., 2014)
  - Encoder-Decoder.
  - Training vs. Testing.
  - Backpropagation.
  - More about RNNs.

- Attention mechanism (Bahdanau et al., 2015)

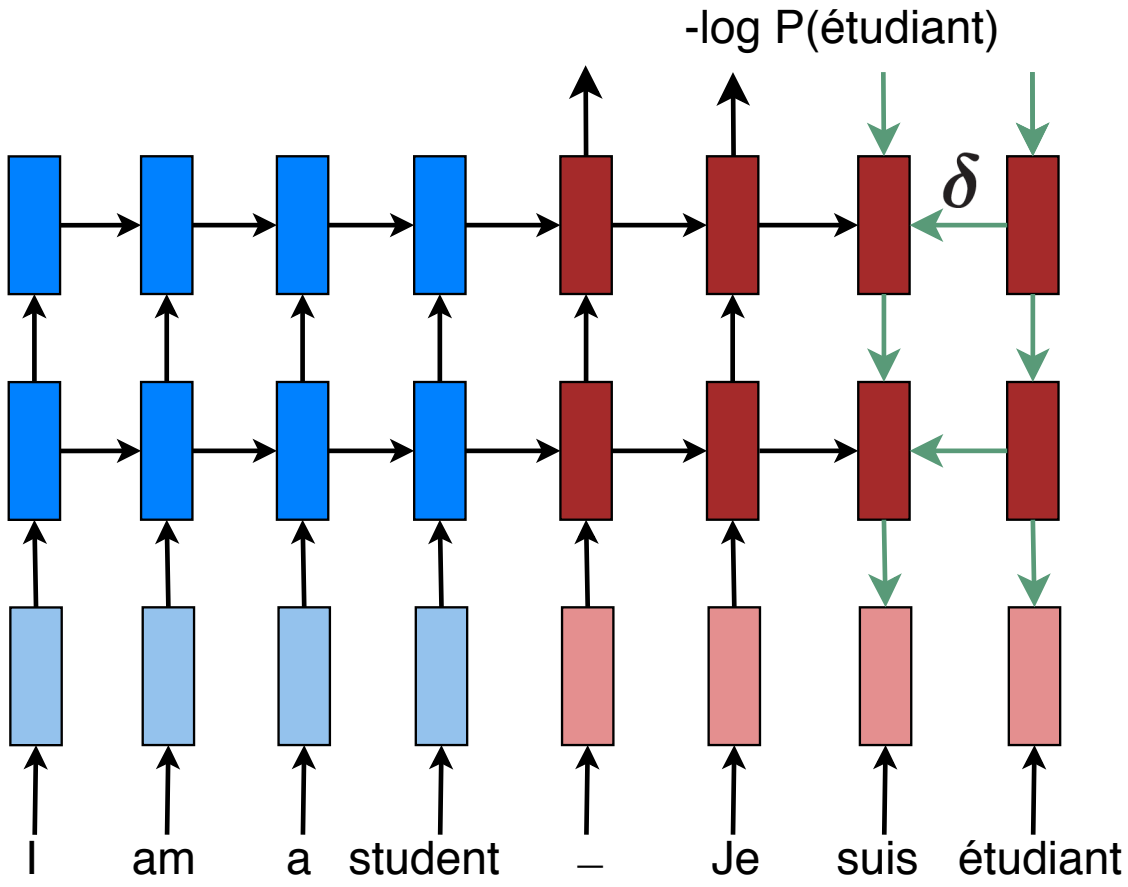
Backpropagation Through Time



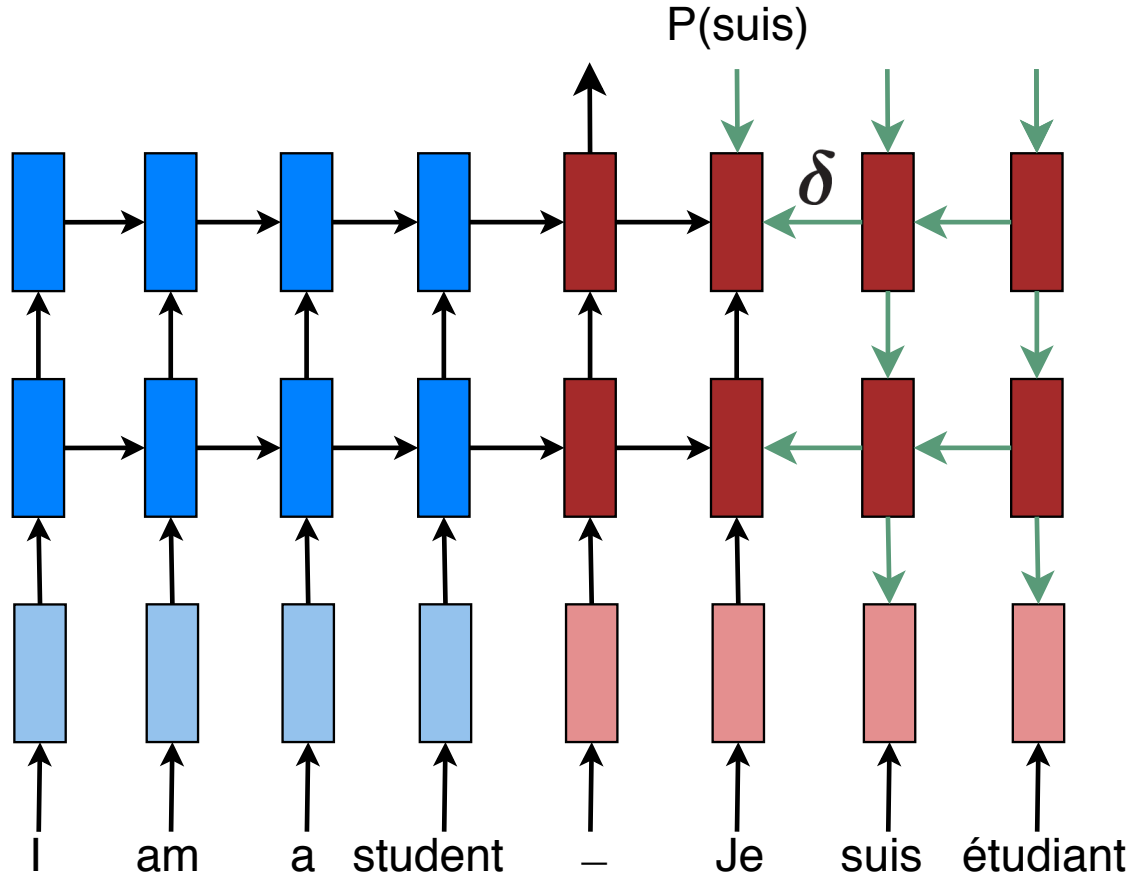
Backpropagation Through Time



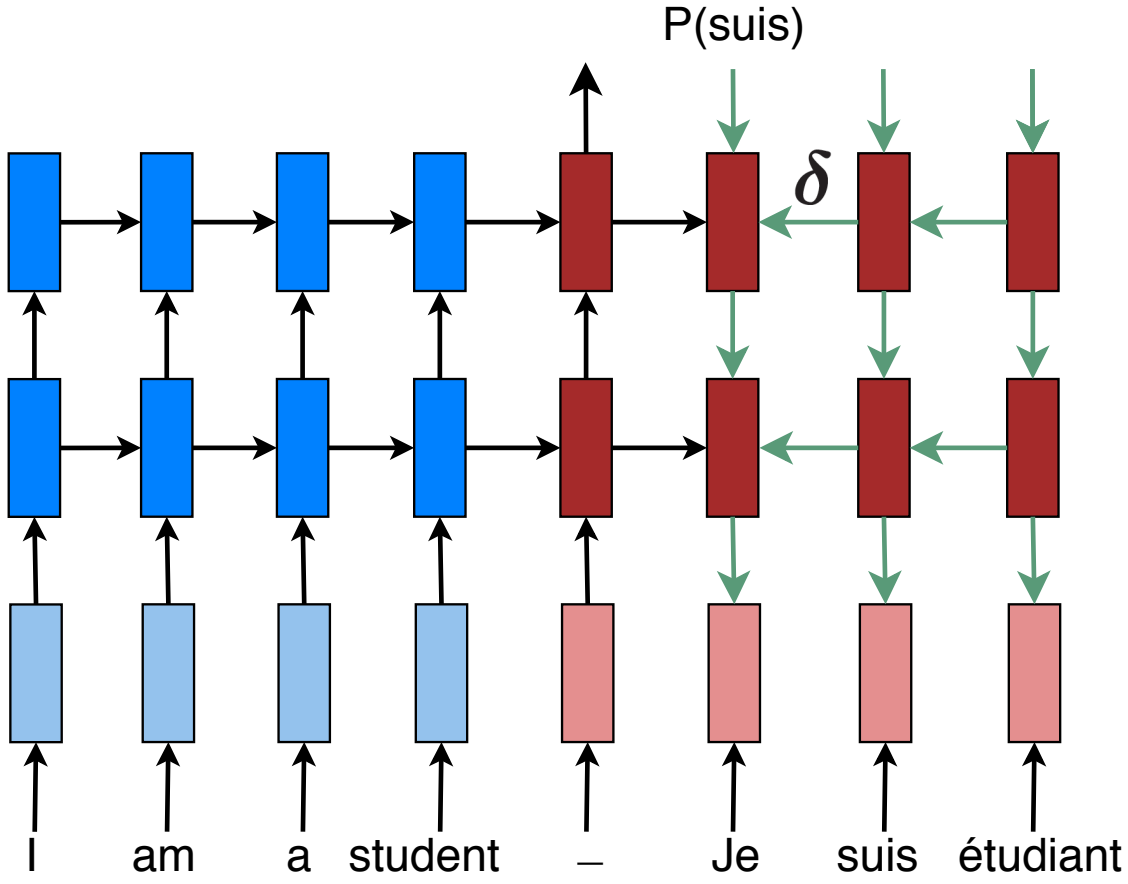
Backpropagation Through Time



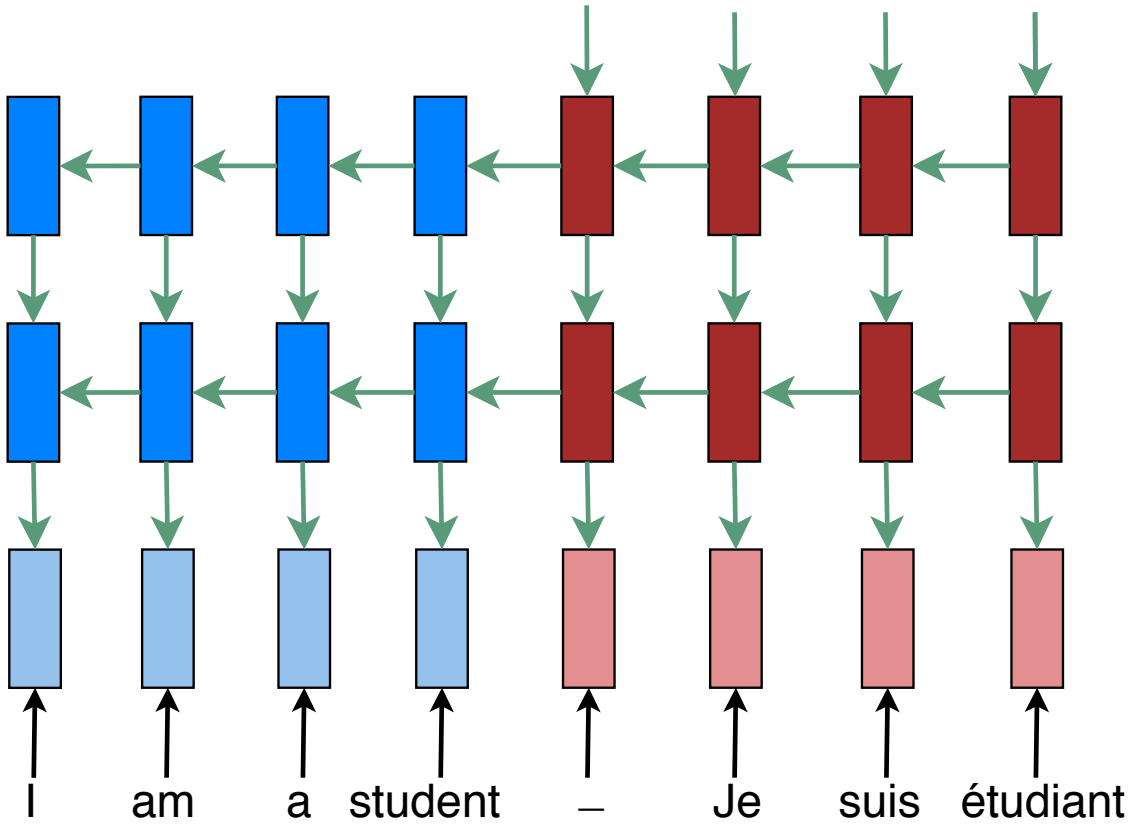
Backpropagation Through Time



Backpropagation Through Time



Backpropagation Through Time



RNN gradients are accumulated.

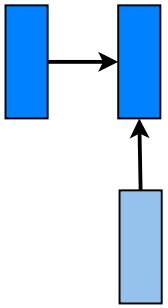
Outline

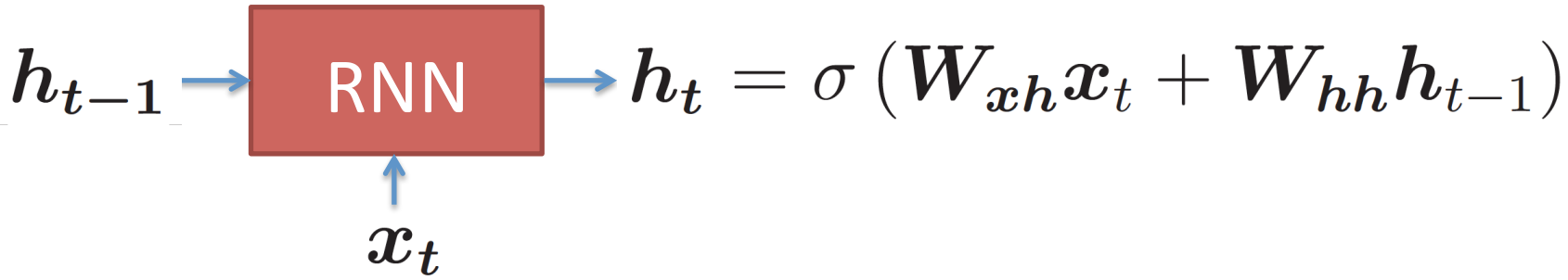
- Recurrent Neural Networks (RNNs)

- NMT basics (Sutskever et al., 2014)
  - Encoder-Decoder.
  - Training vs. Testing.
  - Backpropagation.
  - More about RNNs.

- Attention mechanism (Bahdanau et al., 2015)

Recurrent types – vanilla RNN





Vanishing gradient problem!

Vanishing gradients

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}_{\boldsymbol{xh}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{hh}}\boldsymbol{h}_{t-1}\right)$$

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \operatorname{diag}\left(\sigma'(\ldots)\right) \boldsymbol{W}_{\boldsymbol{hh}}{}^{\top}$$

Chain Rule

$$\left\| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \right\| \leq \left\| \operatorname{diag}\left(\sigma'(\ldots)\right) \right\| \left\| \boldsymbol{W}_{\boldsymbol{hh}}{}^{\top} \right\|$$

Bound Rules

$$\leq \gamma \lambda_1$$

Bound $\left\| \operatorname{diag}\left(\sigma'(\ldots)\right) \right\|$ Largest singular value $\boldsymbol{W}_{\boldsymbol{hh}}{}^{\top}$

(Pascanu et al., 2013)

Vanishing gradients

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}_{\boldsymbol{xh}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{hh}}\boldsymbol{h}_{t-1}\right)$$

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \operatorname{diag}\left(\sigma'(\ldots)\right) \boldsymbol{W}_{\boldsymbol{hh}}{}^{\top}$$

Chain Rule

$$\left\| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \right\| \leq \gamma\lambda_1$$

Bound Rules

$$\left\| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-k}} \right\| \leq (\gamma\lambda_1)^k$$

Chain Rule

(Pascanu et al., 2013)

Vanishing gradients

$$\boldsymbol{h}_t = \sigma\left(\boldsymbol{W}_{\boldsymbol{xh}}\boldsymbol{x}_t + \boldsymbol{W}_{\boldsymbol{hh}}\boldsymbol{h}_{t-1}\right)$$

$$\frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} = \operatorname{diag}\left(\sigma'(\ldots)\right) \boldsymbol{W}_{\boldsymbol{hh}}{}^{\top}$$

Chain Rule

$$\left\| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-1}} \right\| \leq \gamma\lambda_1$$

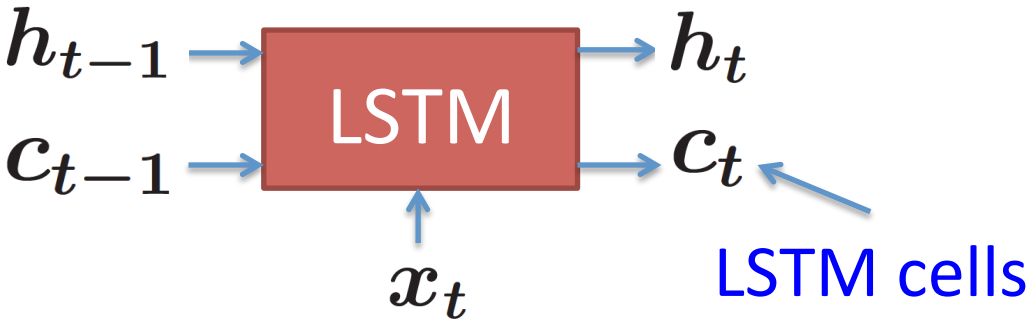
Bound Rules

$$\left\| \frac{\partial \boldsymbol{h}_t}{\partial \boldsymbol{h}_{t-k}} \right\| \leq (\gamma\lambda_1)^k \rightarrow 0 \text{ if } \lambda_1 < \frac{1}{\gamma}$$

Sufficient Cond

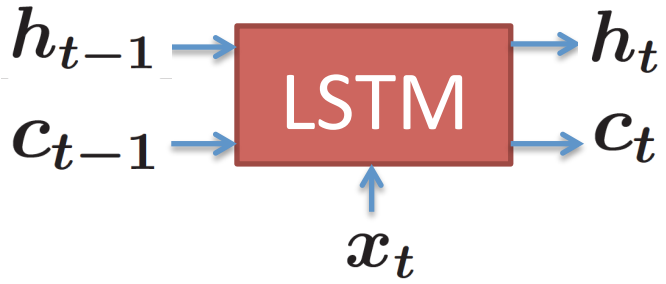
(Pascanu et al., 2013)

Recurrent types – LSTM



- Long-Short Term Memory (LSTM)
  - (Hochreiter & Schmidhuber, 1997)
- LSTM cells are additively updated
  - Make backprop through time easier.

Building LSTM



$$\begin{pmatrix} \\ \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \\ \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = c_{t-1} + \hat{h}_t$$

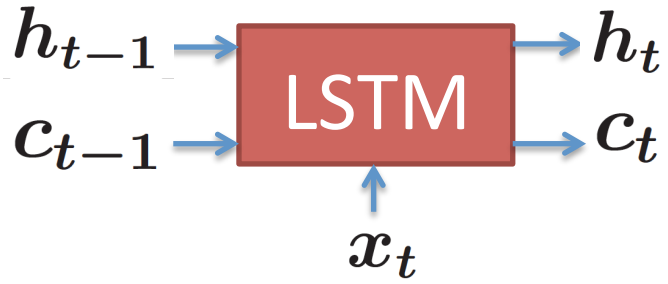
$$h_t = c_t$$

Nice gradients!

$$\frac{\partial c_t}{\partial c_{t-1}} = I$$

- A naïve version.

Building LSTM



Input gates

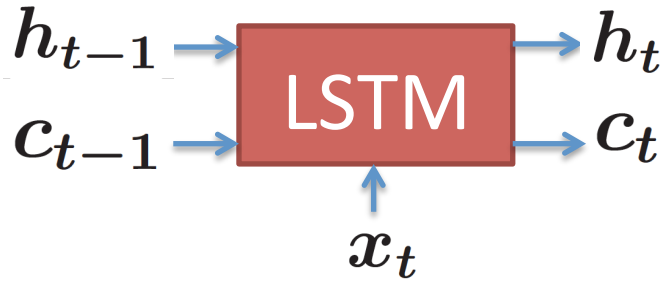
$$\begin{pmatrix} i_t \\ \\ \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \\ \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = c_{t-1} + i_t \circ \hat{h}_t$$

$$h_t = c_t$$

- Add input gates: control input signal.

Building LSTM



Forget gates

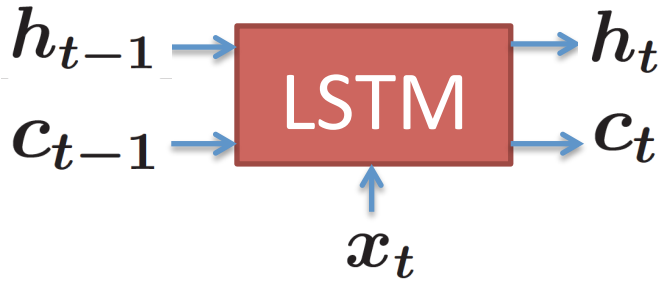
$$\begin{pmatrix} i_t \\ f_t \\ \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$

$$h_t = c_t$$

- Add forget gates: control memory.

Building LSTM



Output gates

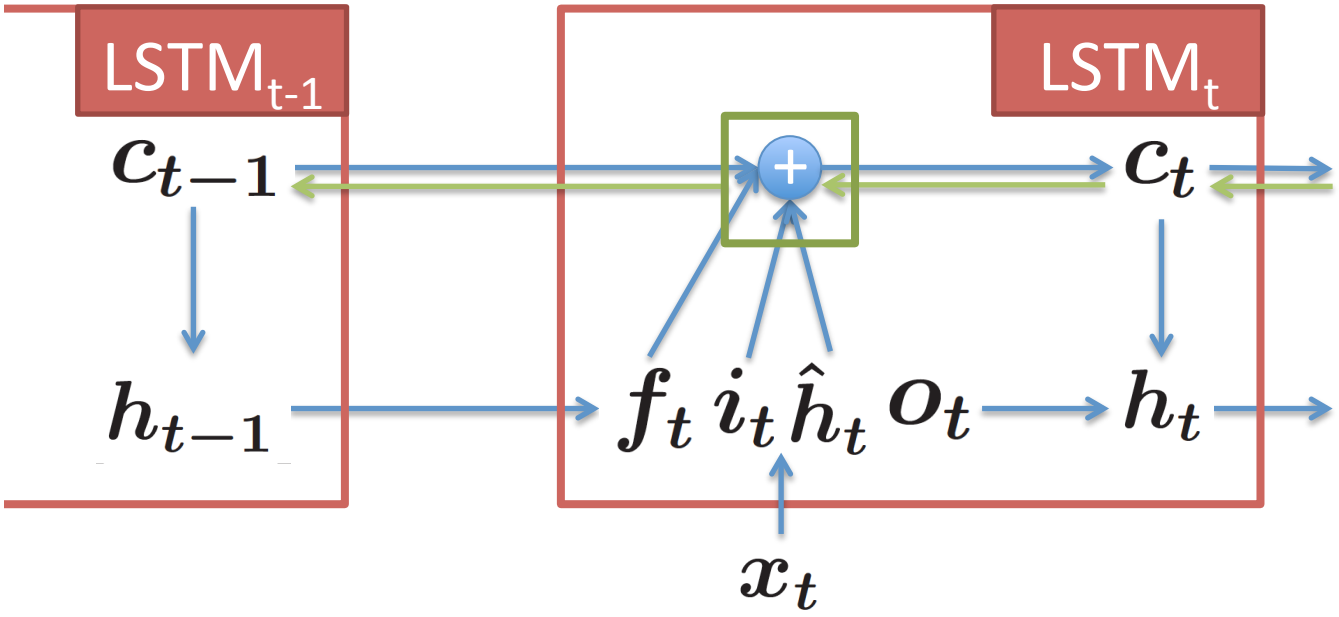
$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$

$$h_t = \boxed{o_t} \circ \tanh(c_t)$$

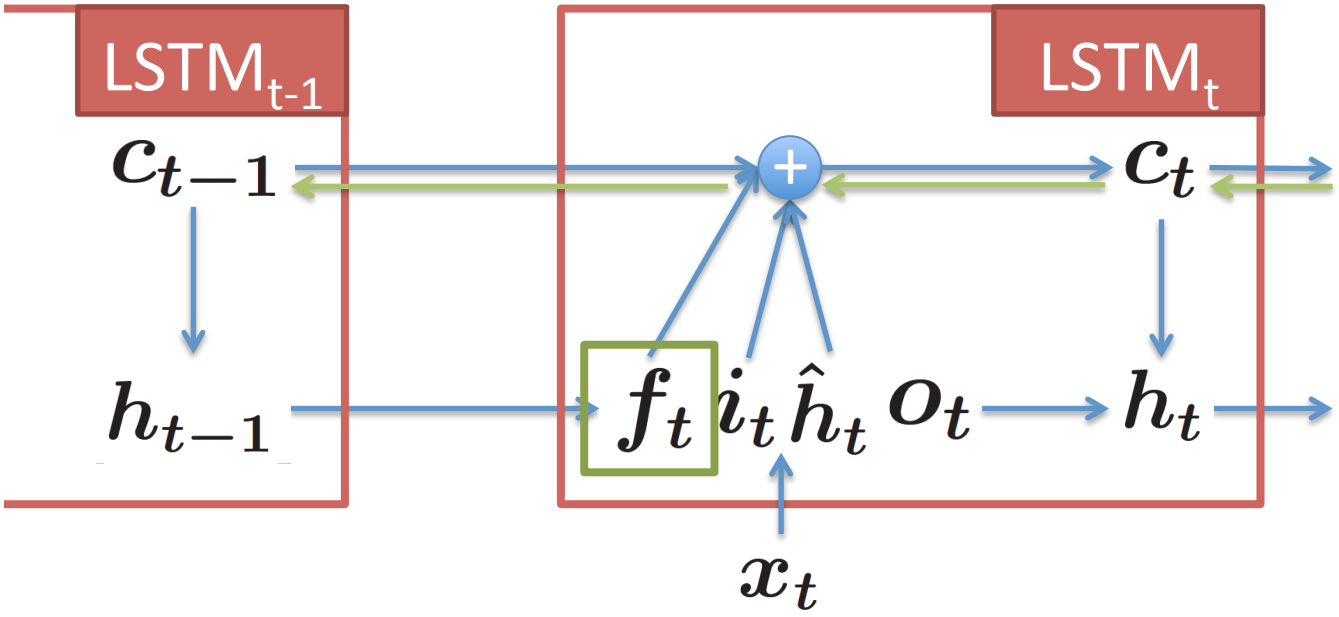
- Add output gates: extract information.
- (Zaremba et al., 2014).

Why LSTM works?



- The additive operation is the key!
- Backpropation path through the cell is effective.

Why LSTM works?



- The additive operation is the key!
- Backpropation path through the cell is effective.

Forget gates are important!

Other RNN units

- (Graves, 2013): revived LSTM.
  - Direct connections between cells and gates.
- Gated Recurrent Unit (GRU) – (Cho et al., 2014a)
  - No cells, same additive idea.
- LSTM vs. GRU: mixed results (Chung et al., 2015).

English – French WMT Results

Systems	BLEU
SOTA in WMT'14 *(Durrani et al., 2014)*	**37.0**
Standard MT + neural components	
Schwenk (2014) – neural language model	33.3
Cho et al. (2014) – phrase table neural features	34.5

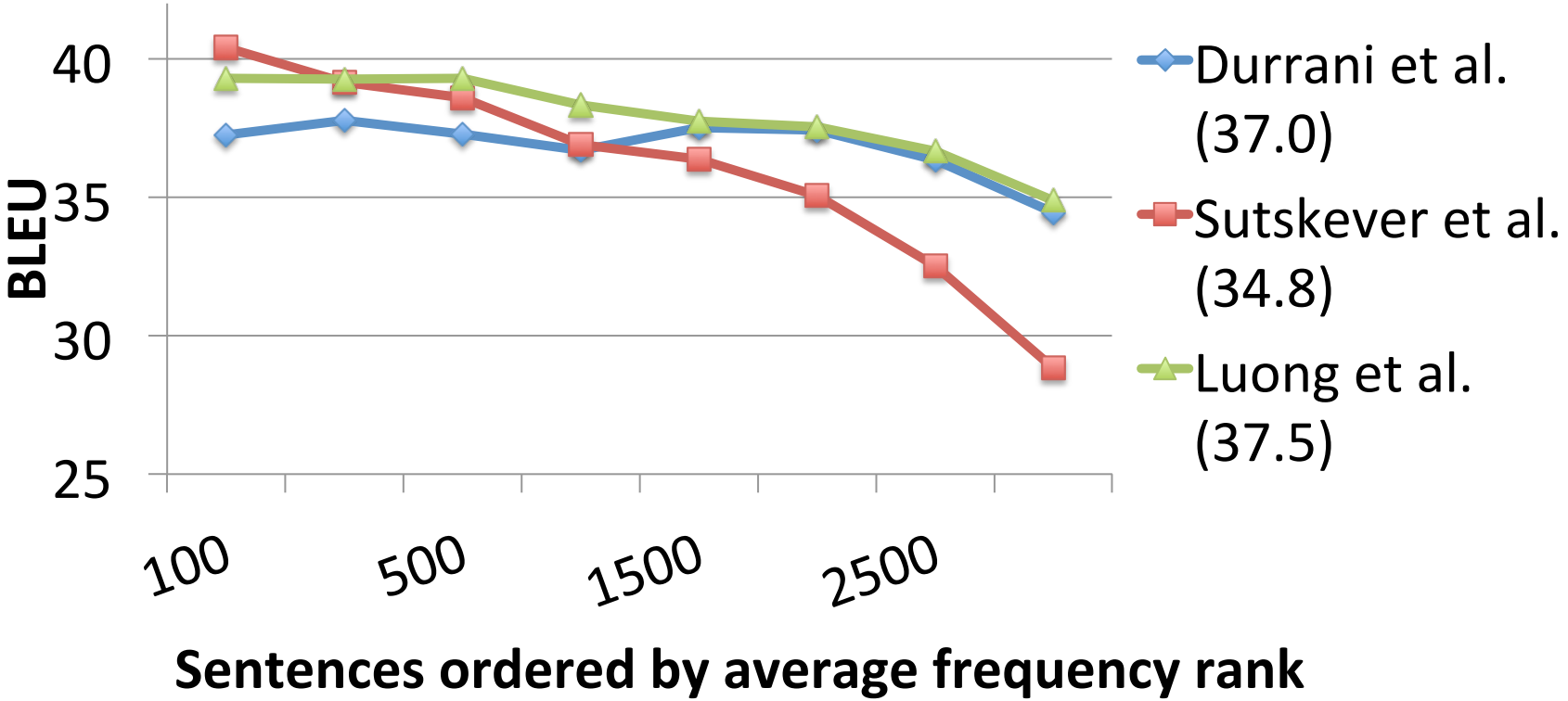
English – French WMT Results

Systems	BLEU
SOTA in WMT'14 *(Durrani et al., 2014)*	**37.0**
Standard MT + neural components	
Schwenk (2014) – neural language model	33.3
Cho et al. (2014) – phrase table neural features	34.5
NMT	
Sutskever et al. (2014) – ensemble LSTMs	34.8

English – French WMT Results

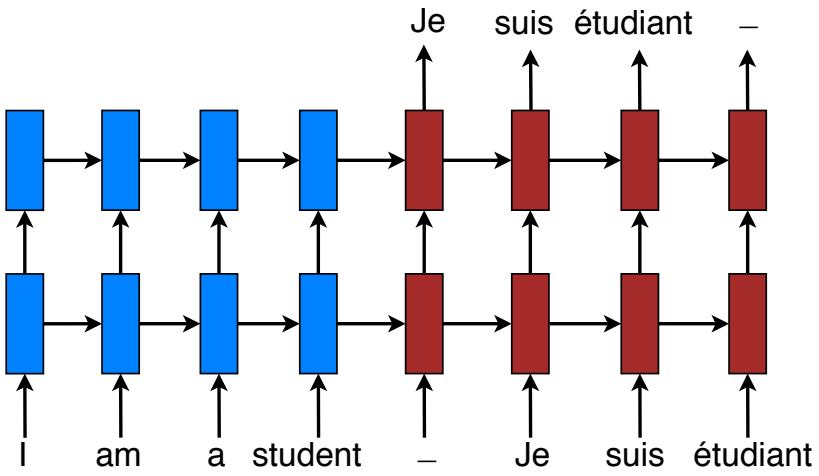
Systems	BLEU
SOTA in WMT'14 *(Durrani et al., 2014)*	**37.0**
Standard MT + neural components	
Schwenk (2014) – neural language model	33.3
Cho et al. (2014) – phrase table neural features	34.5
NMT	
Sutskever et al. (2014) – ensemble LSTMs	34.8
Luong et al. (2015a) – ensemble LSTMs + rare word	**37.5**

Effects of Translating Rare Words



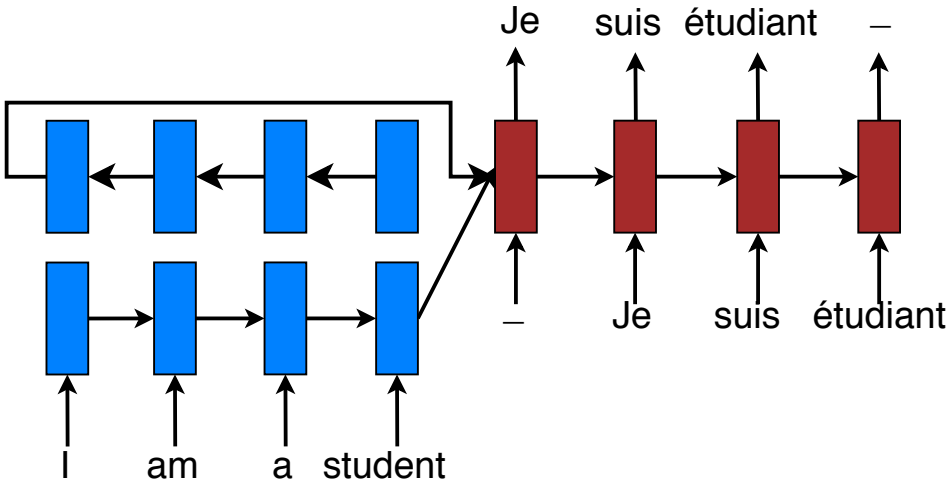
Deep RNNs

(Sutskever et al., 2014)



Bidirectional RNNs

(Bahdanau et al., 2015)



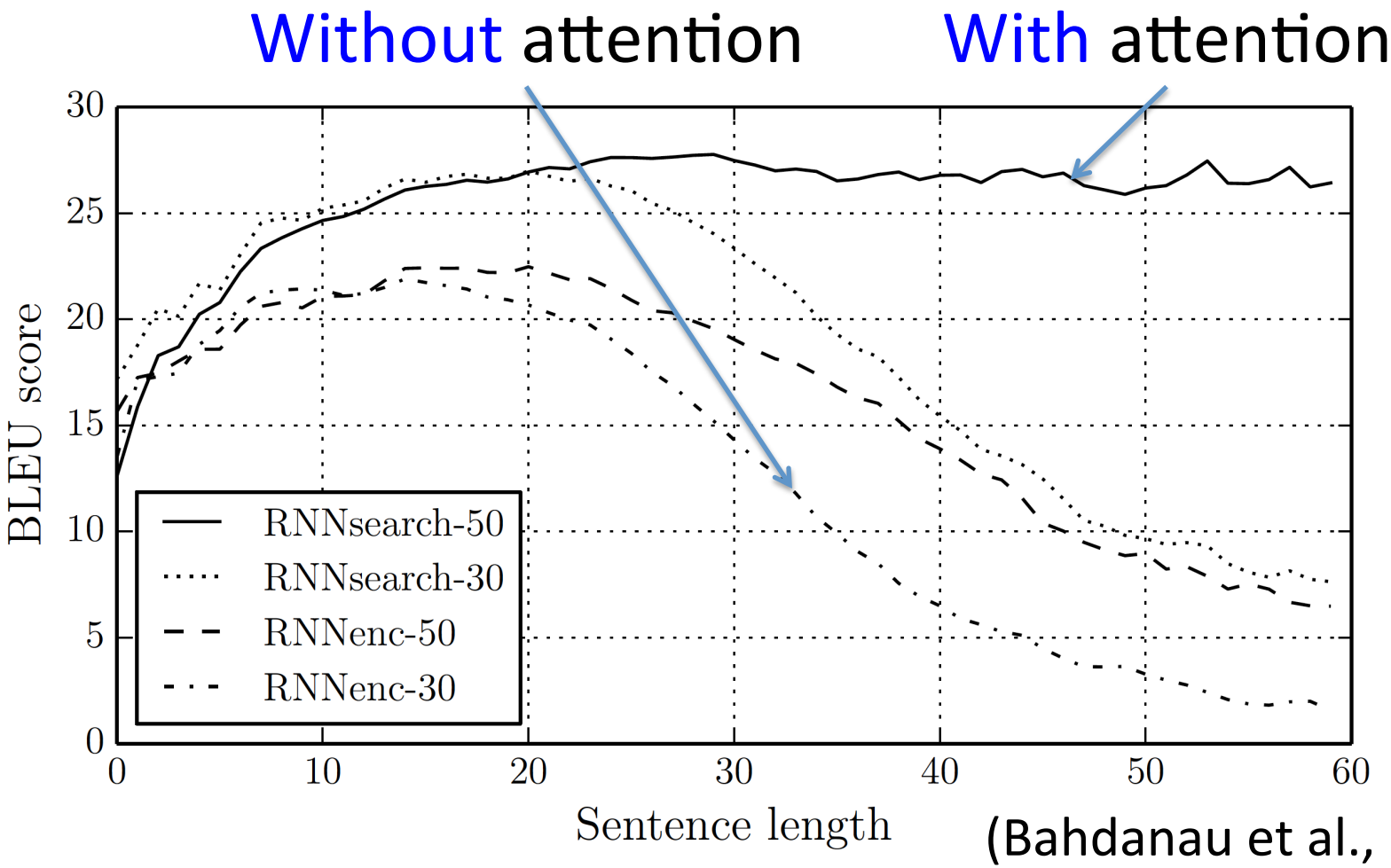
Summary

- Generalize well.
- Small memory.
- Simple decoder.

Outline

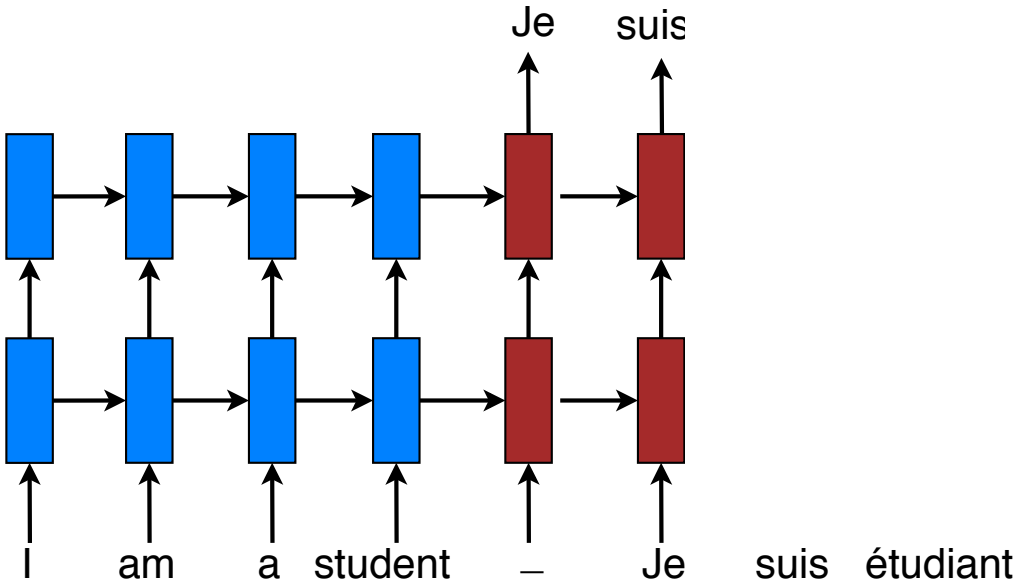
- Recurrent Neural Networks (RNNs)
- NMT basics (Sutskever et al., 2014)
- Attention mechanism (Bahdanau et al., 2015)

Sentence Length Problem



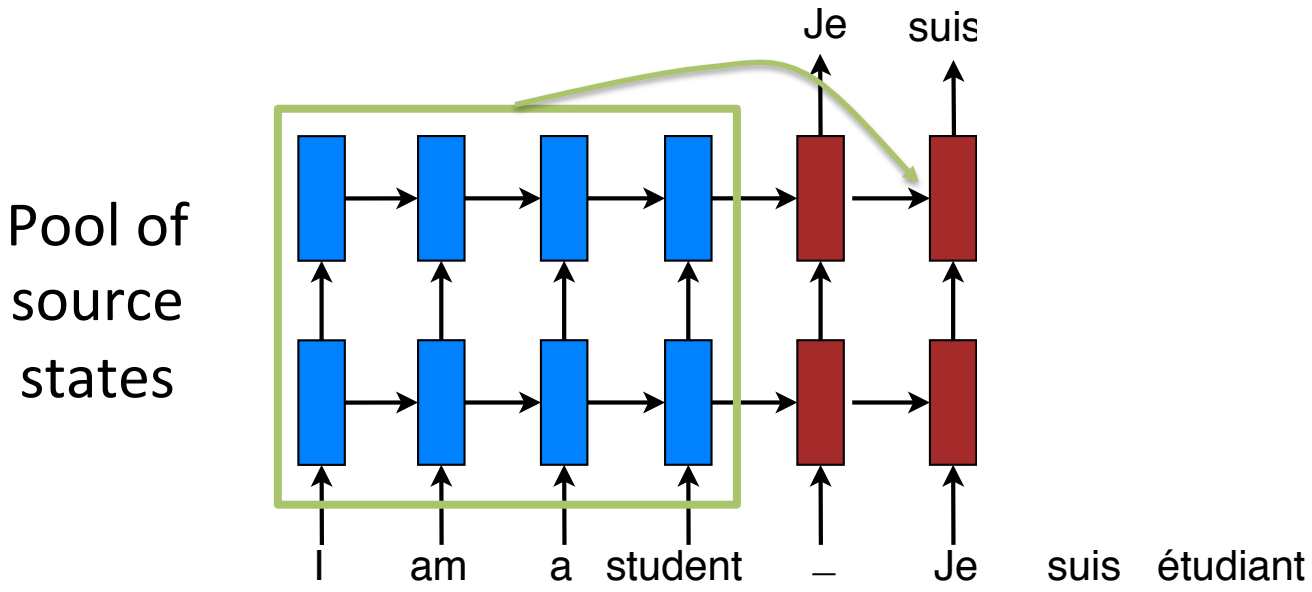
(Bahdanau et al., 2015)

Why?



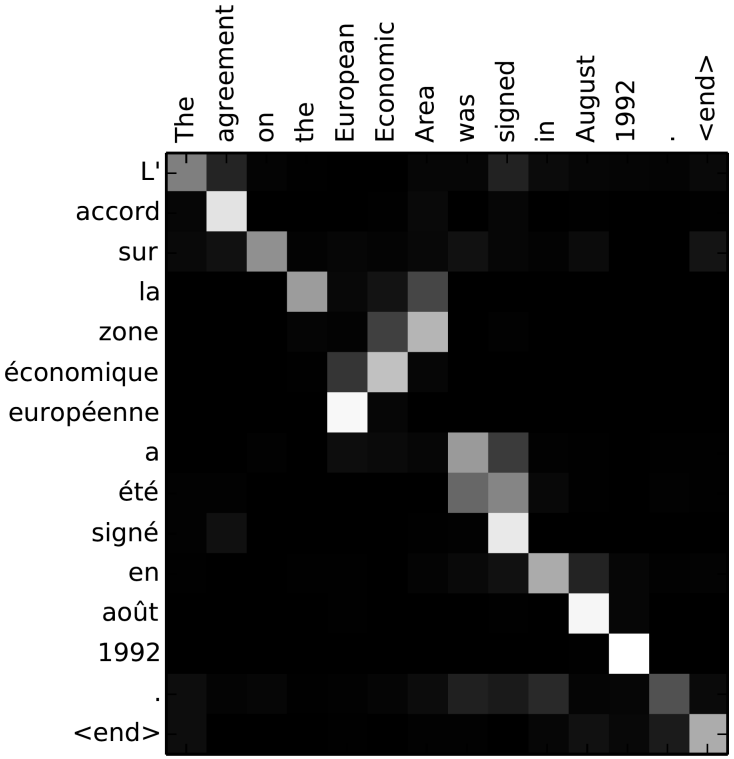
- A fixed-dimensional source vector.
- Problem: Markovian process.

Attention Mechanism



- Solution: random access memory
  - Retrieve as needed.
  - cf. Neural Turing Machine (Graves et al., 2014).

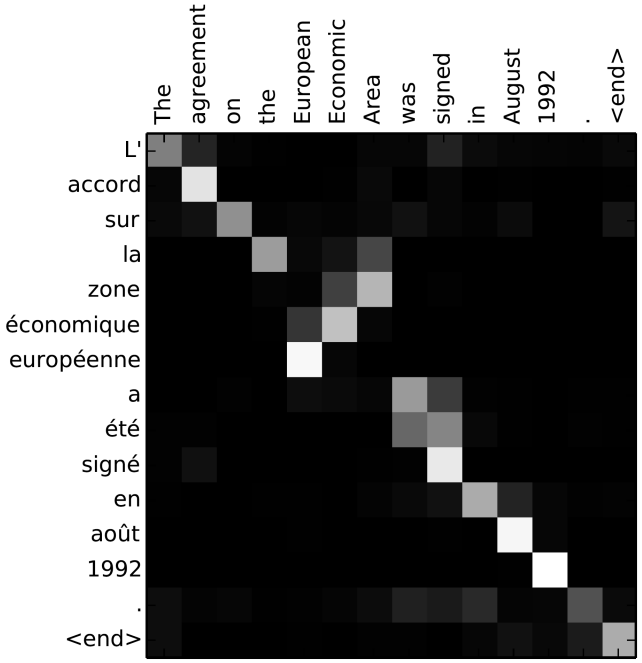
Alignments as a by-product



(Bahdanau et al., 2015)

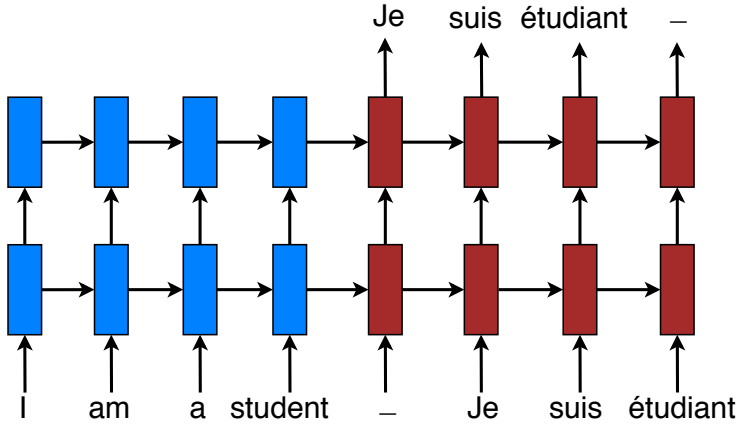
- Recent innovation in deep learning:
  - Control problem (Mnih et al., 14)
  - Speech recognition (Chorowski et al., 15)
  - Image caption generation (Xu et al., 15)

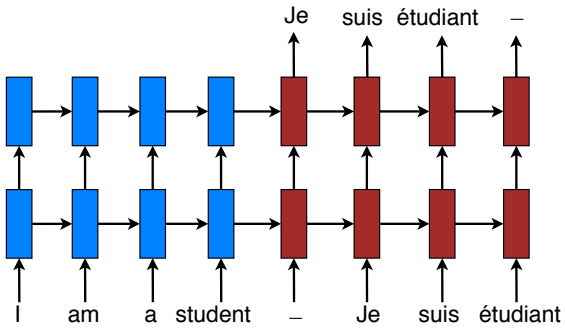
Simplified Attention
(Bahdanau et al., 2015)

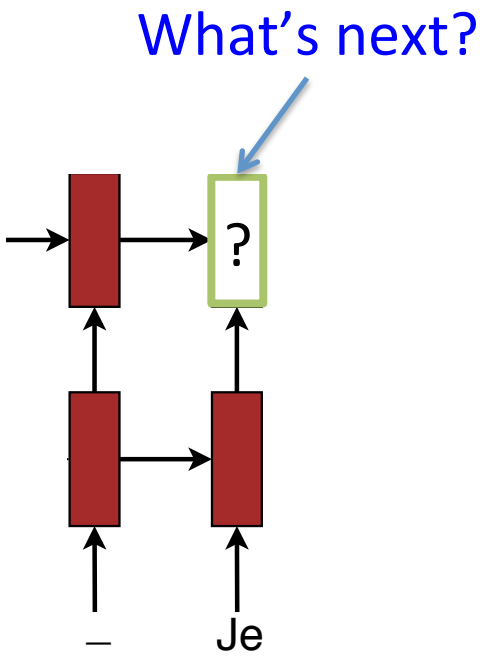


+

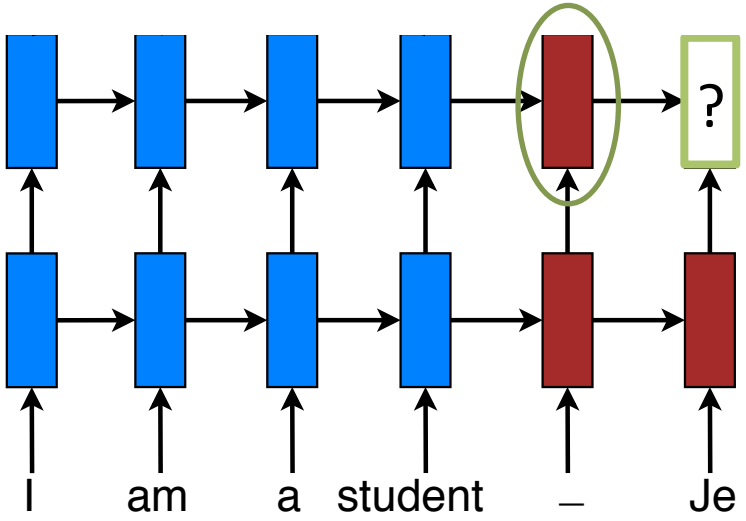
Deep LSTM
(Sutskever et al., 2014)




Je
suis
étudiant
–
I
am
a
student
–
Je
suis
étudiant


What's next?
?
–
Je

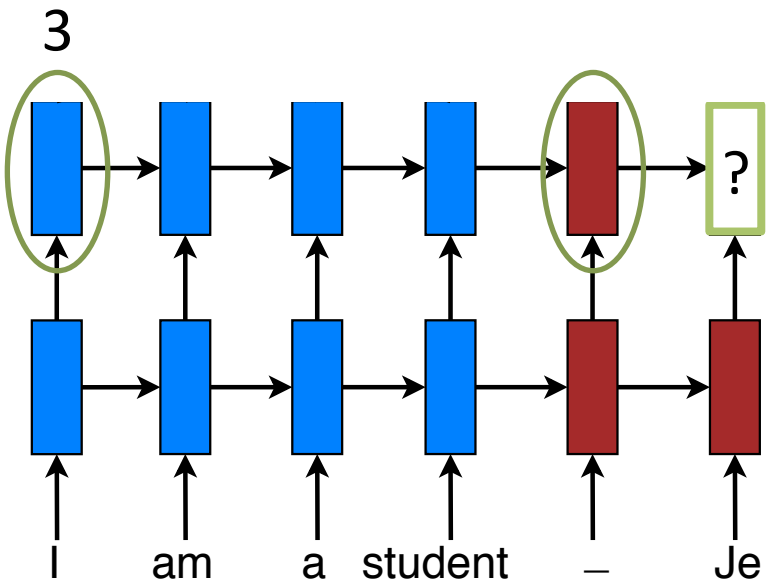
Attention Mechanism – *Scoring*



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

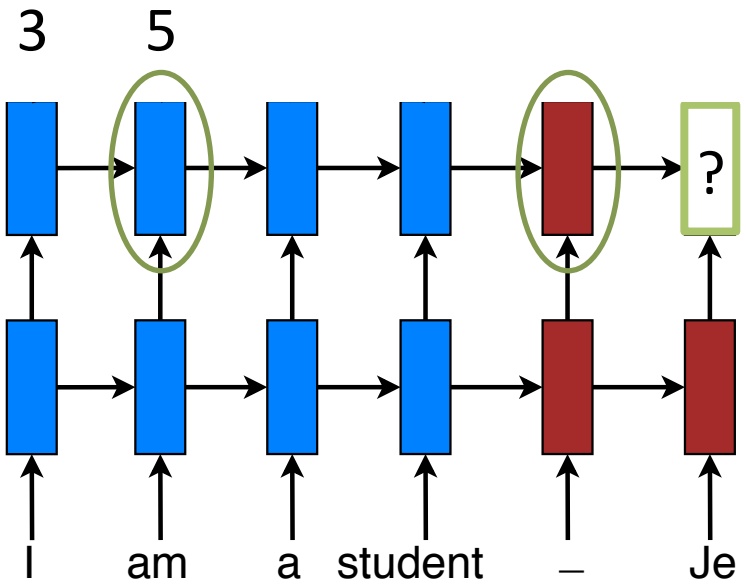
$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

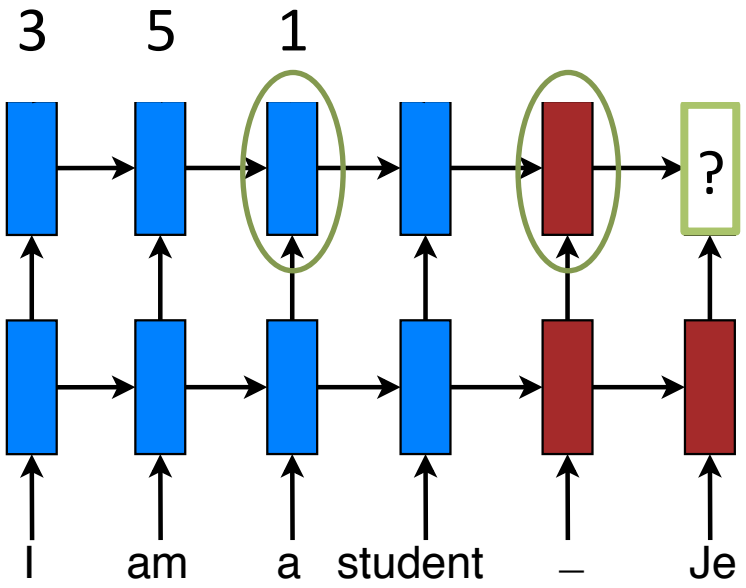
$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

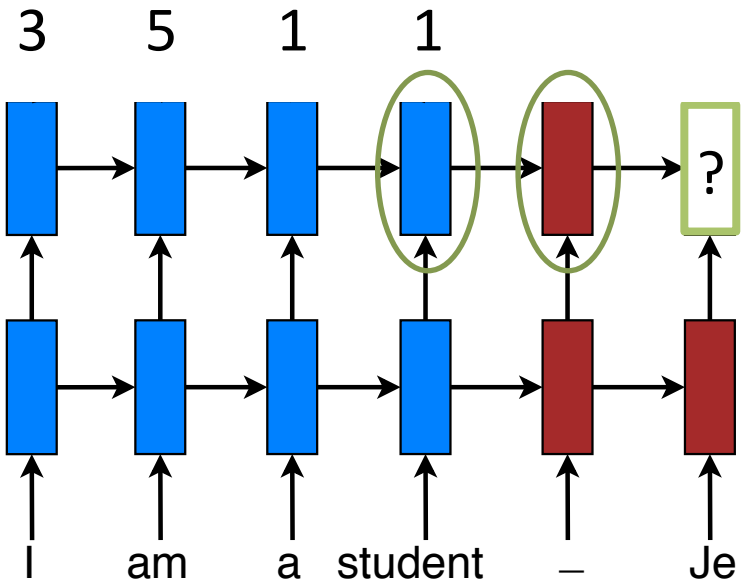
$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

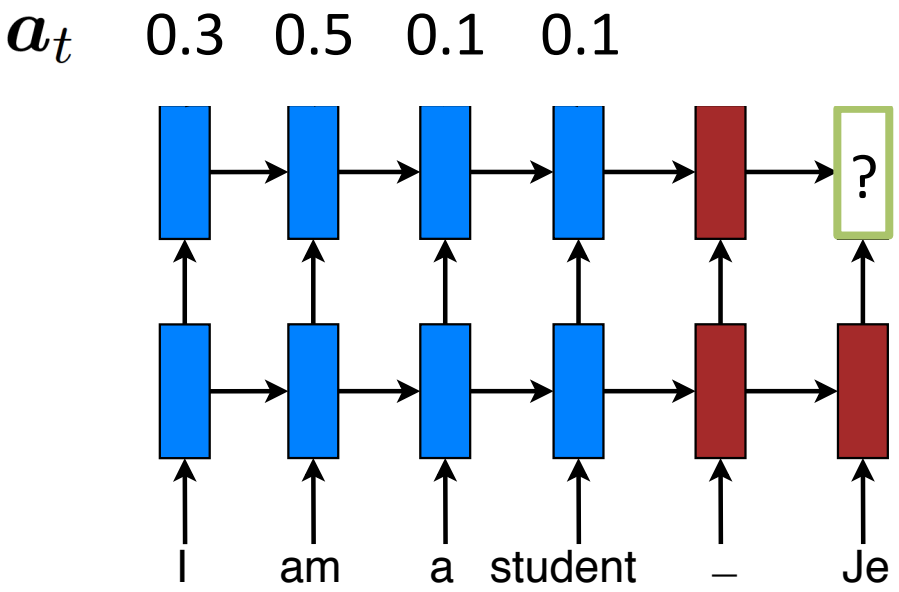
Attention Mechanism – *Scoring*

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



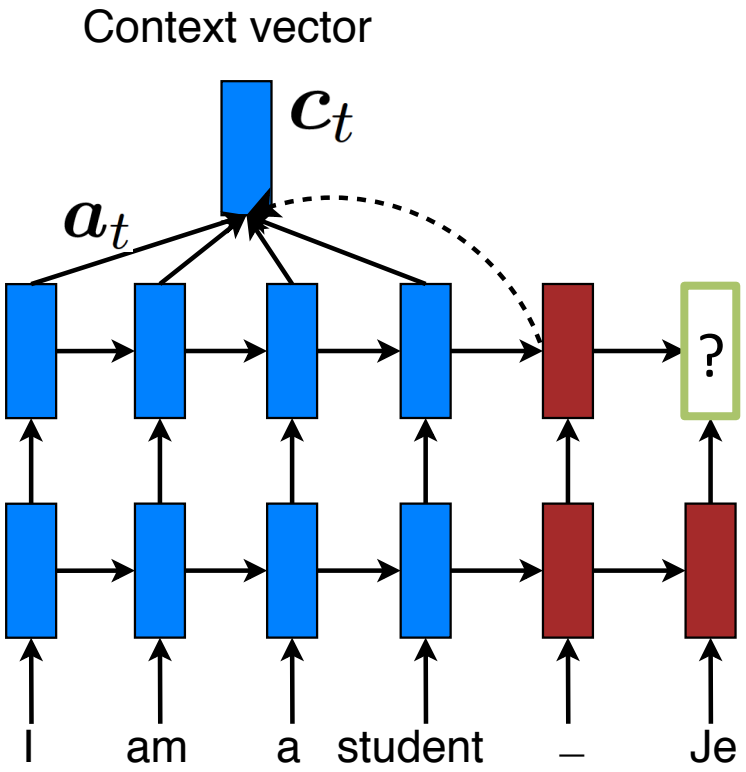
- Compare target and source hidden states.

Attention Mechanism – *Normalization*



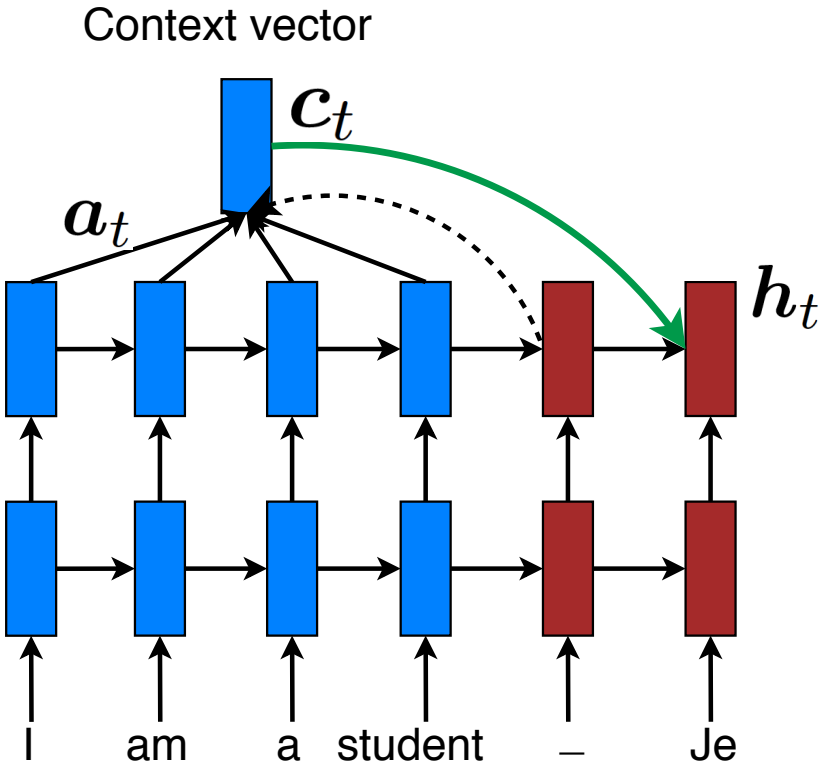
- Convert into alignment weights.

Attention Mechanism – *Context vector*



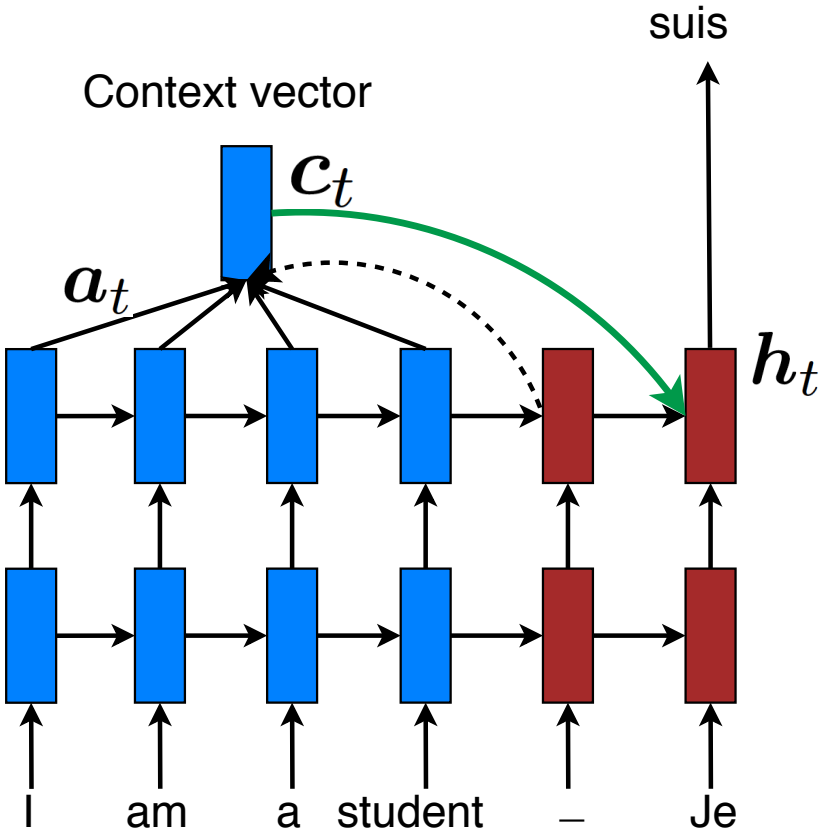
- Build context vector: weighted average.

Attention Mechanism – *Hidden state*



- Compute the next hidden state.

Attention Mechanism – *Predict*



- Predict the next word.

Attention Mechanism – *Score Functions*

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W}_{\boldsymbol{a}}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) \end{cases}$$ (Bahdanau et al., 2015)

Attention Mechanism – *Score Functions*

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s \\ \boldsymbol{h}_t^\top \boldsymbol{W_a} \bar{\boldsymbol{h}}_s \\ \boldsymbol{v}_a^\top \tanh\left(\boldsymbol{W_a}[\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s]\right) \end{cases}$$

(Luong et al., 2015b)

(Bahdanau et al., 2015)

- More focused attention (Luong et al., 2015b)
  - Focus on a subset of words each time.

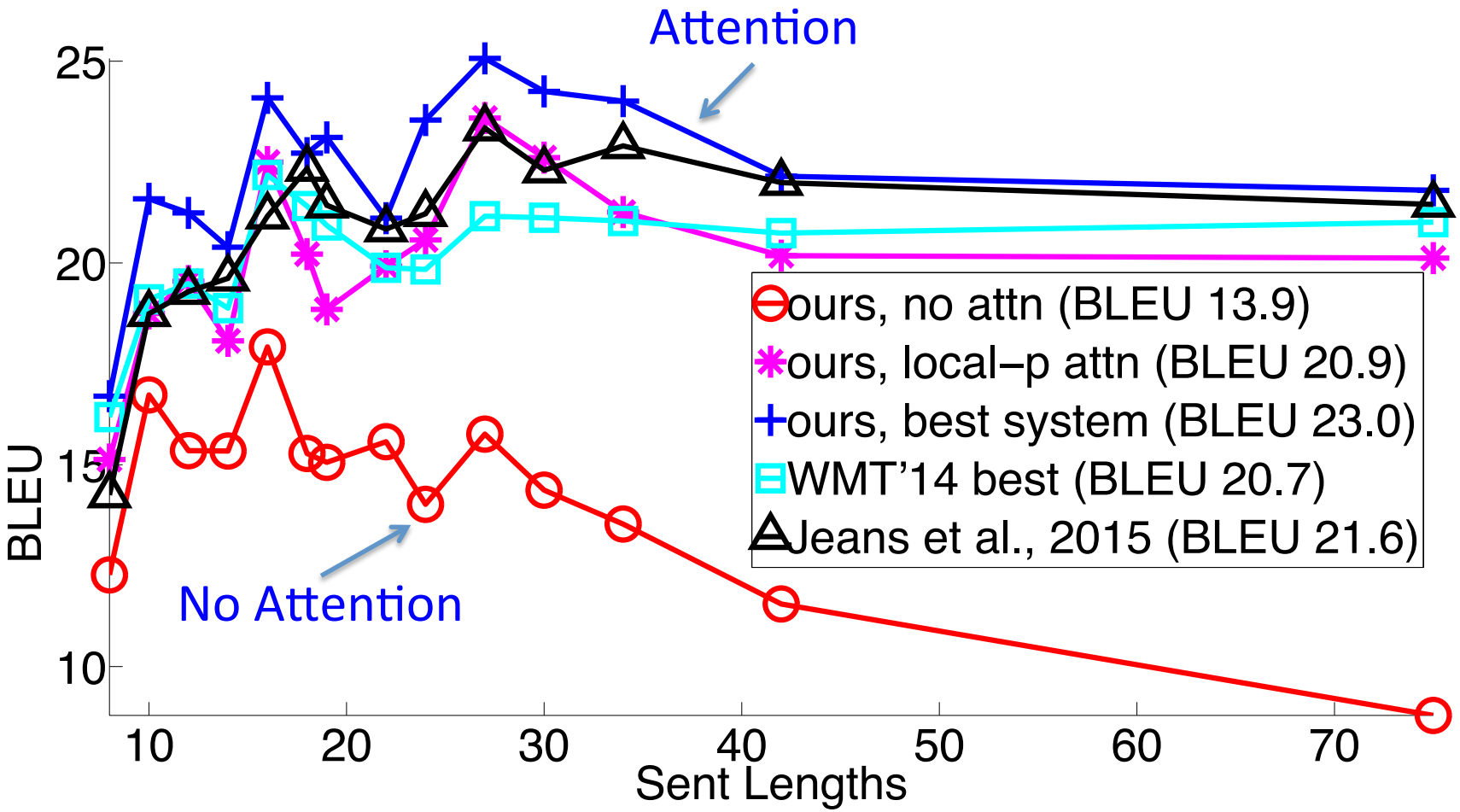
English-German WMT Results

Systems	BLEU
SOTA in WMT'14 (Buck et al., 2014)	20.7
NMT	
Jean et al., (2015) – GRUs + attention	21.6

English-German WMT Results

Systems	BLEU
SOTA in WMT'14 (Buck et al., 2014)	20.7
NMT	
Jean et al., (2015) – GRUs + attention	21.6
Luong et al. (2015b) – LSTMs + improved attention	**23.0 (+2.3)**

Translate Long Sentences



(Luong et al., 2015b)

Sample English-German translations

src	Orlando Bloom and *Miranda Kerr* still love each other
ref	Orlando Bloom und **Miranda Kerr** lieben sich noch immer
best	Orlando Bloom und **Miranda Kerr** lieben einander noch immer .
base	Orlando Bloom und **Lucas Miranda** lieben einander noch immer .

- Translate names correctly.

(Luong et al., 2015b)

Sample English-German translations

src	’’ We ’ re pleased the FAA recognizes that an enjoyable passenger experience is ***not incompatible*** with safety and security , ’’ said Roger Dow , CEO of the U.S. Travel Association .
ref	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur Sicherheit steht** ” , sagte Roger Dow , CEO der U.S. Travel Association .
best	’’ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit **unvereinbar** ist ’’ , sagte Roger Dow , CEO der US - die .
base	’’ Wir freuen uns u ̈ber die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit ’’ , sagte Roger Cameron , CEO der US - <unk> .

- Translate a doubly-negated phrase correctly

(Luong et al., 2015b)

Sample English-German translations

src	’’ We ’ re pleased the FAA recognizes that an enjoyable passenger experience is ***not incompatible*** with safety and security , ’’ said Roger Dow , CEO of the U.S. Travel Association .
ref	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur Sicherheit steht** ” , sagte Roger Dow , CEO der U.S. Travel Association .
best	’’ Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit **unvereinbar** ist ’’ , sagte Roger Dow , CEO der US - die .
base	’’ Wir freuen uns u¨ber die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit ’’ , sagte Roger Cameron , CEO der US - <unk> .

- Translate a doubly-negated phrase correctly

(Luong et al., 2015b)

Sample English-German translations

src	'' We ' re pleased the FAA recognizes that an enjoyable passenger experience is ***not incompatible*** with safety and security , '' said Roger Dow , CEO of the U.S. Travel Association .
ref	" Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur Sicherheit steht** " , sagte Roger Dow , CEO der U.S. Travel Association .
best	'' Wir freuen uns , dass die FAA anerkennt , dass ein angenehmes ist nicht mit Sicherheit und Sicherheit **unvereinbar** ist '' , sagte Roger Dow , CEO der US - die .
base	'' Wir freuen uns u¨ber die <unk> , dass ein <unk> <unk> mit Sicherheit nicht **vereinbar** ist mit Sicherheit und Sicherheit '' , sagte Roger Cameron , CEO der US - <unk> .

- Fail to translate "passenger experience".

(Luong et al., 2015b)

Sample German-English translations

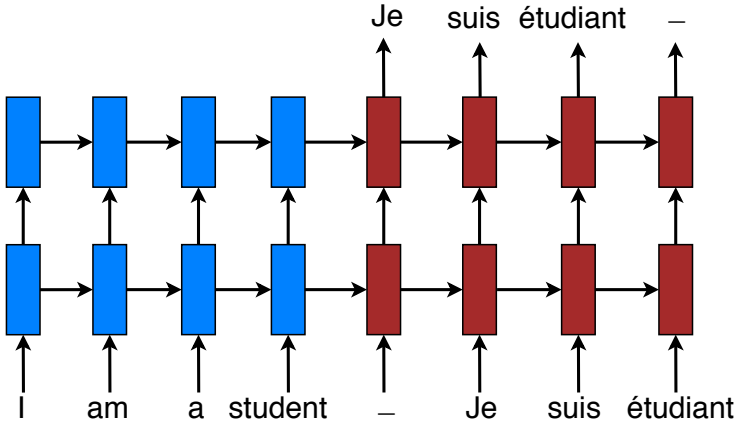
src	Wegen der von Berlin und der Europa¨ischen Zentralbank verha¨ngten strengen Sparpolitik in Verbindung mit der Zwangsjacke , in die die jeweilige nationale Wirtschaft durch das Festhal- ten an der gemeinsamen Wa¨hrung geno¨tigt wird , sind viele Menschen der Ansicht , das Projekt Europa sei zu weit gegangen
ref	The **austerity imposed by Berlin and the European Central Bank** , **coupled with the straitjacket** imposed on national economies through adherence to the common currency , has led many people to think Project Europe has gone too far .
best	Because of the strict **austerity measures imposed by Berlin and the European Central Bank** **in connection with the straitjacket** in which the respective national economy is forced to adhere to the common currency , many people believe that the European project has gone too far .
base	Because of the pressure **imposed by the European Central Bank and the Federal Central Bank with the strict austerity** imposed on the national economy in the face of the single currency , many people believe that the European project has gone too far .

- Translate well long sentences. (Luong et al., 2015b)

Thank you!

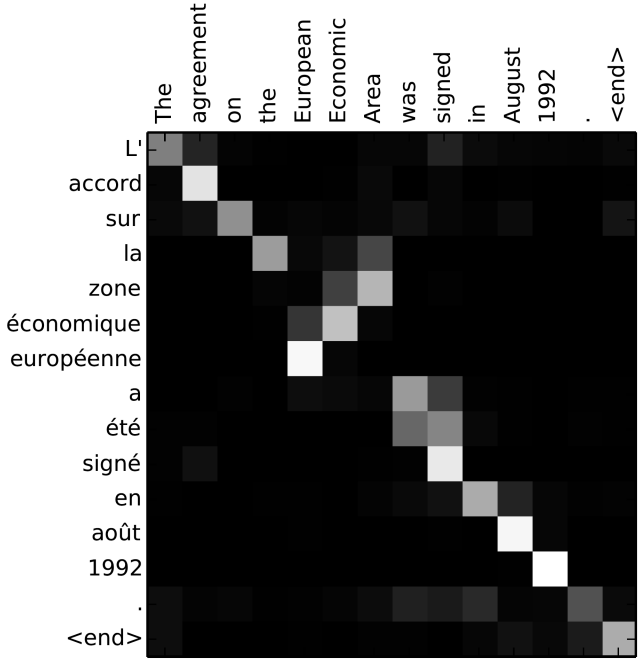
Summary

Deep LSTM
(Sutskever et al., 2014)



+

Simplified Attention
(Bahdanau et al., 2015)



References (1)

- [Bahdanau et al., 2015] Neural Translation by Jointly Learning to Align and Translate. http://arxiv.org/pdf/1409.0473.pdf
- [Cho et al., 2014a] Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. http://aclweb.org/anthology/D/D14/D14-1179.pdf
- [Cho et al., 2014b] On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. http://www.aclweb.org/anthology/W14-4012
- [Chorowski et al., 2015] Attention-Based Models for Speech Recognition. http://arxiv.org/pdf/1506.07503v1.pdf
- [Chung et al., 2015] Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. http://arxiv.org/pdf/1412.3555.pdf
- [Graves, 2013] Generating Sequences With Recurrent Neural Networks. http://arxiv.org/pdf/1308.0850v5.pdf
- [Graves, 2014] Neural Turing Machine. http://arxiv.org/pdf/1410.5401v2.pdf.
- [Hochreiter & Schmidhuber, 1997] Long Short-term Memory. http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf

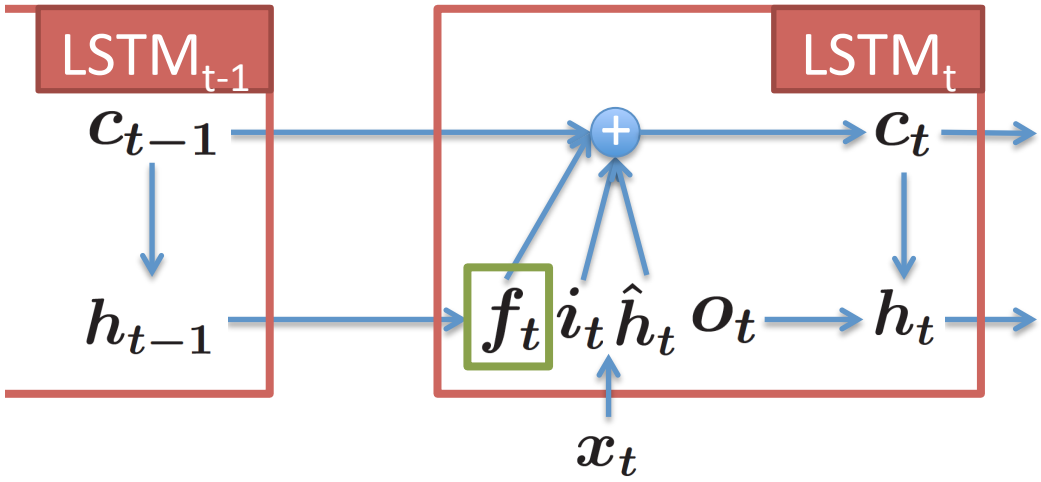
References (2)

- [Kalchbrenner & Blunsom, 2013] Recurrent Continuous Translation Models. http://nal.co/papers/KalchbrennerBlunsom_EMNLP13
- [Luong et al., 2015a] Addressing the Rare Word Problem in Neural Machine Translation. http://www.aclweb.org/anthology/P15-1002
- [Luong et al., 2015b] Effective Approaches to Attention-based Neural Machine Translation. https://aclweb.org/anthology/D/D15/D15-1166.pdf
- [Mnih et al., 2014] Recurrent Models of Visual Attention. http://papers.nips.cc/paper/5542-recurrent-models-of-visual-attention.pdf
- [Pascanu et al., 2013] On the difficulty of training Recurrent Neural Networks. http://arxiv.org/pdf/1211.5063v2.pdf
- [Xu et al., 2015] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. http://jmlr.org/proceedings/papers/v37/xuc15.pdf
- [Sutskever et al., 2014] Sequence to Sequence Learning with Neural Networks. http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf
- [Zaremba et al., 2015] Recurrent Neural Network Regularization. http://arxiv.org/pdf/1409.2329.pdf

Encoder-decoder Summary

	Encoder	Decoder
(Sutskever et al., 2014) (Luong et al., 2015a) (Luong et al., 2015b)	Deep LSTM	Deep LSTM
(Cho et al., 2014a) (Bahdanau et al., 2015) (Jean et al., 2015)	(Bidirectional) GRU	GRU
(Kalchbrenner & Blunsom, 2013)	CNN	(Inverse CNN) RNN
(Cho et al., 2014b)	Gated Recursive CNN	GRU

Important LSTM components?



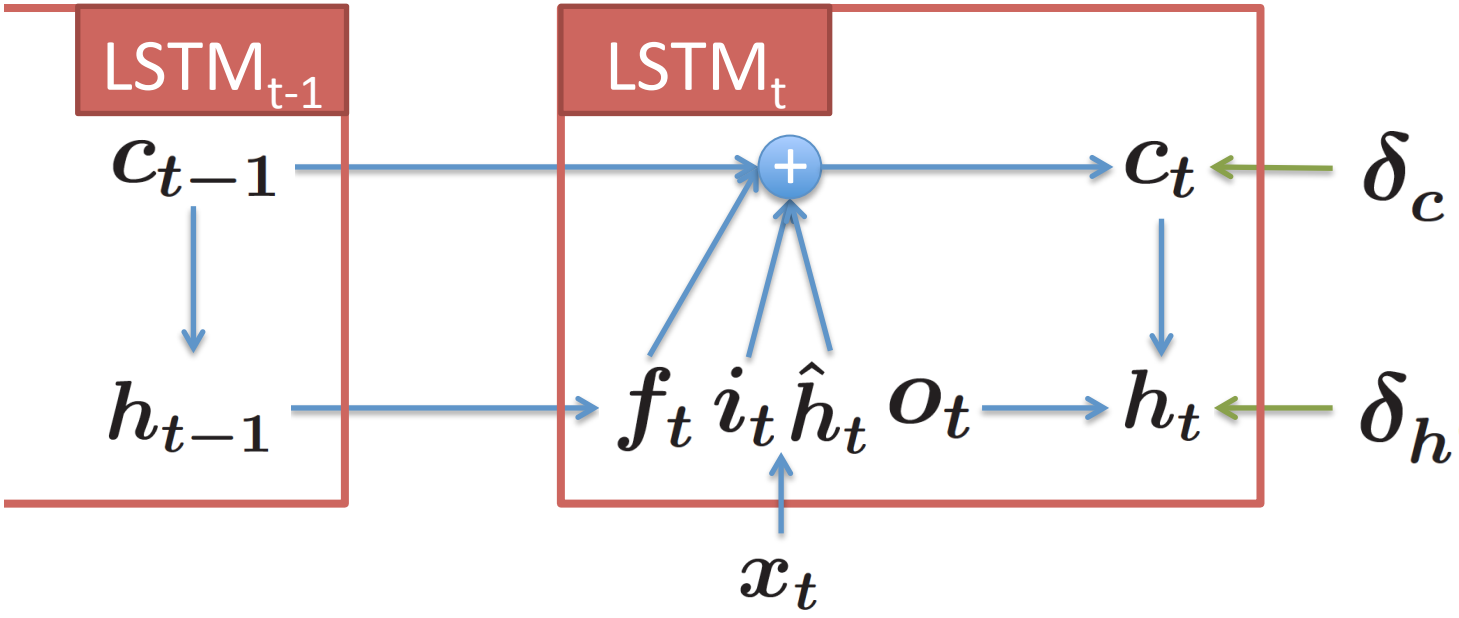
- (Jozefowicz et al., 2015): forget gate bias of 1.

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$

- (Greff et al., 2015): forget gates & output acts.

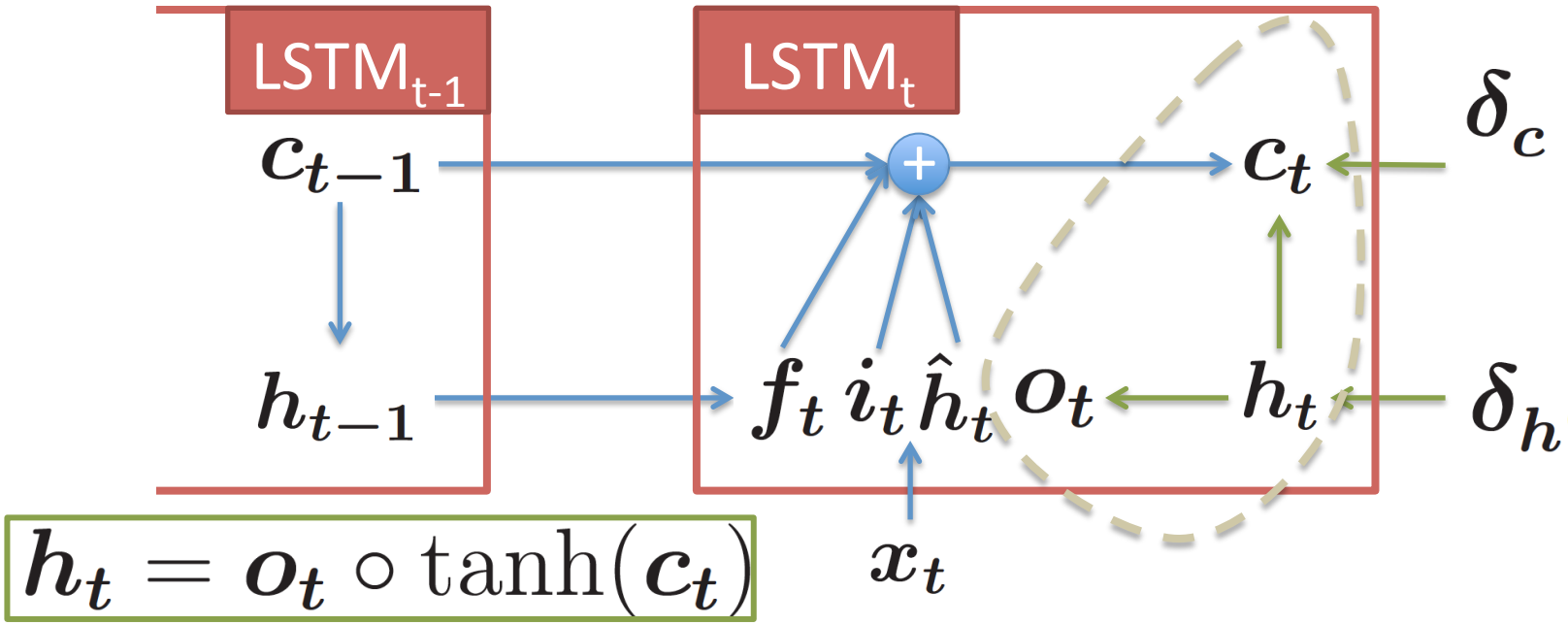
$$h_t = o_t \circ \tanh(c_t)$$

LSTM Backpropagation

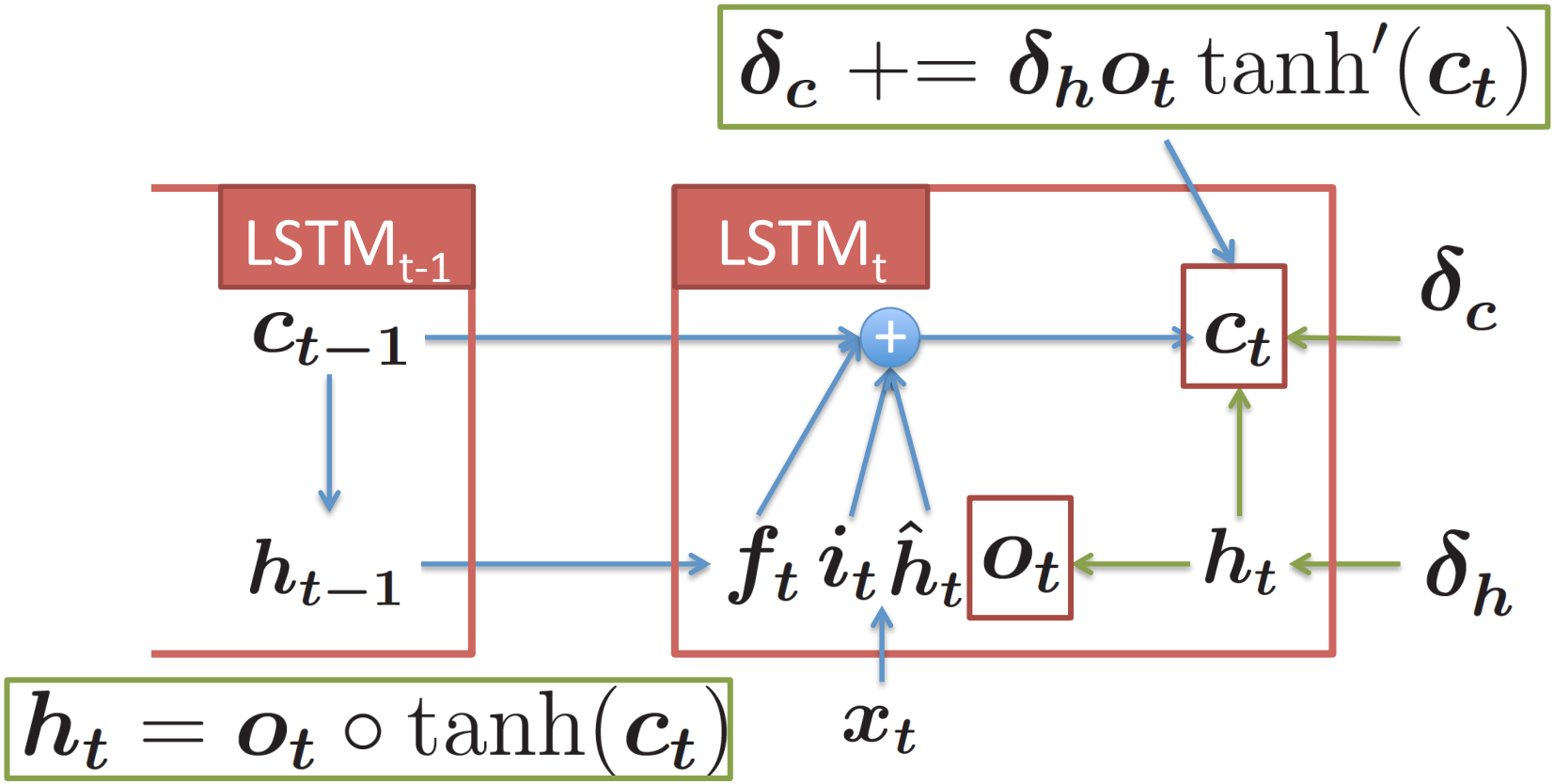


- Deltas sent back from the top layers.

LSTM Backpropagation – *Context*



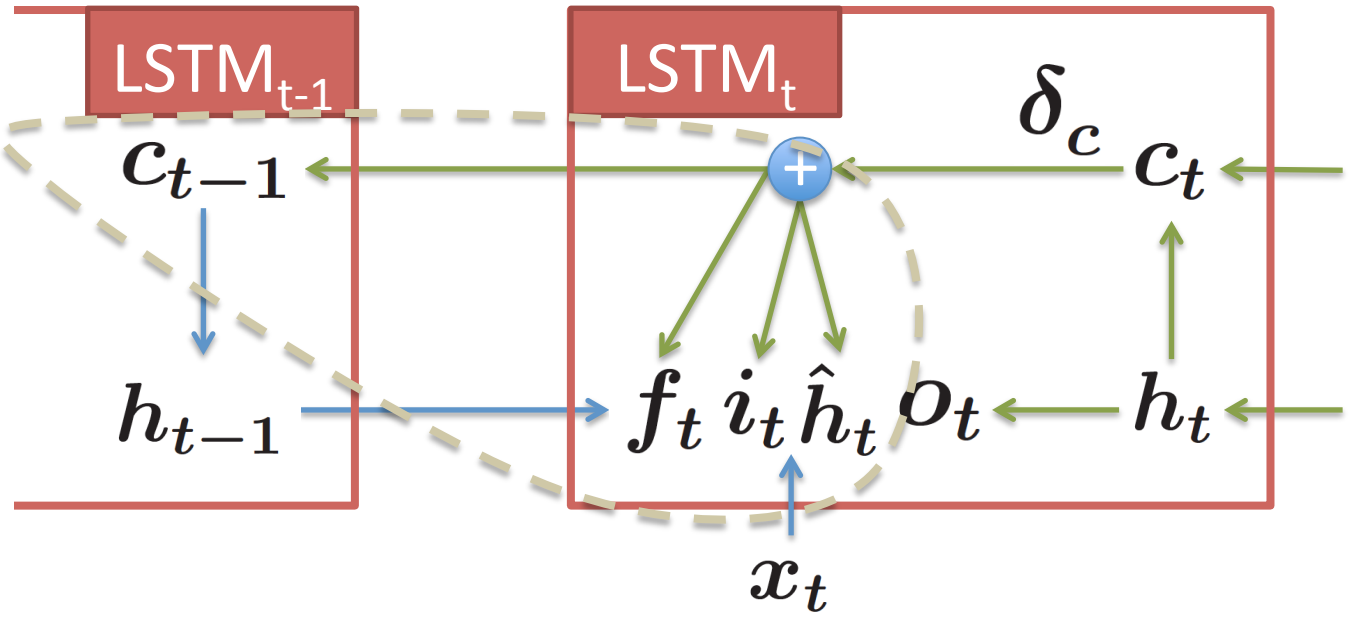
LSTM Backpropagation – *Context*



- Complete context vector gradient.

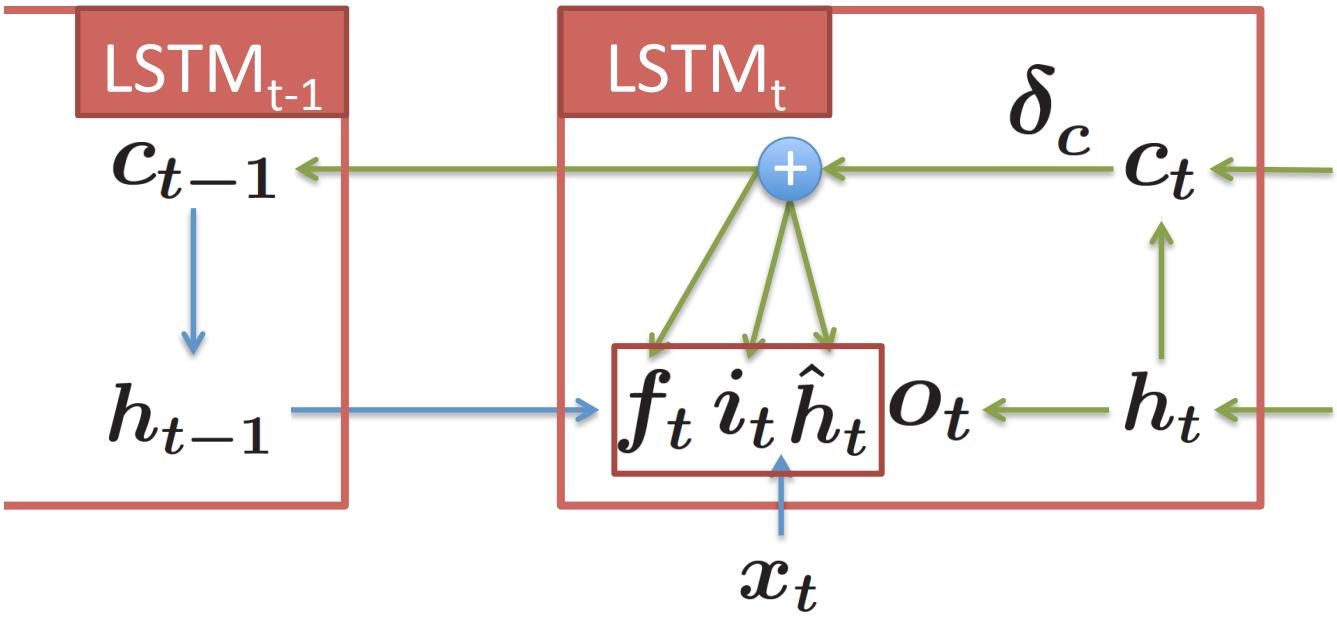
LSTM Backpropagation – *Context*

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$



LSTM Backpropagation – *Context*

$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$

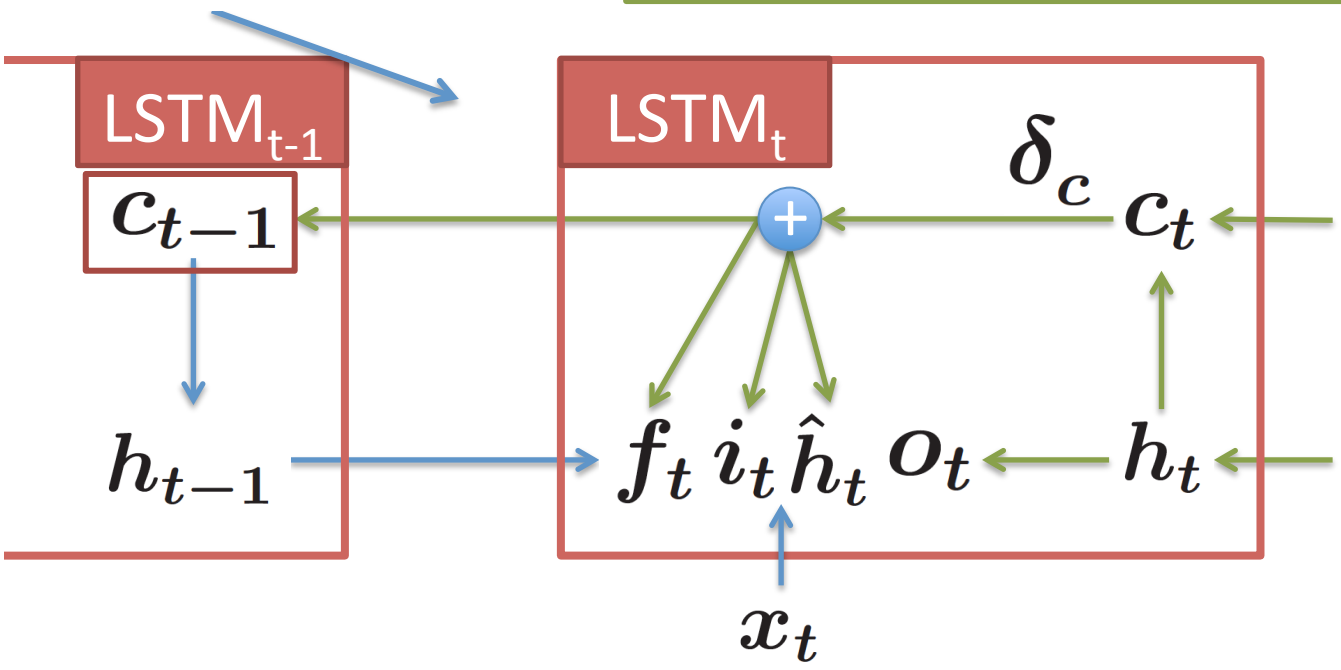


- First, use δ_c to compute gradients for $f_t\ i_t\ \hat{h}_t$.

LSTM Backpropagation – *Context*

$$\delta_c = \delta_c \circ f_t$$

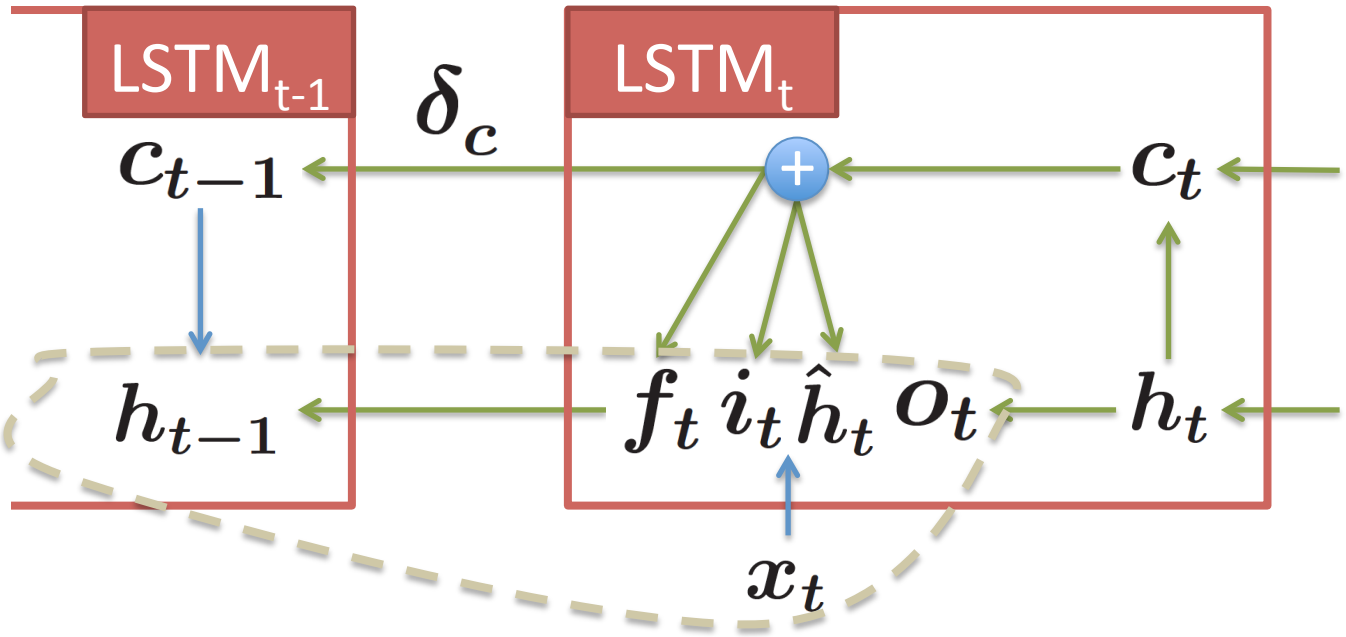
$$c_t = f_t \circ c_{t-1} + i_t \circ \hat{h}_t$$



- Then, update δ_c.

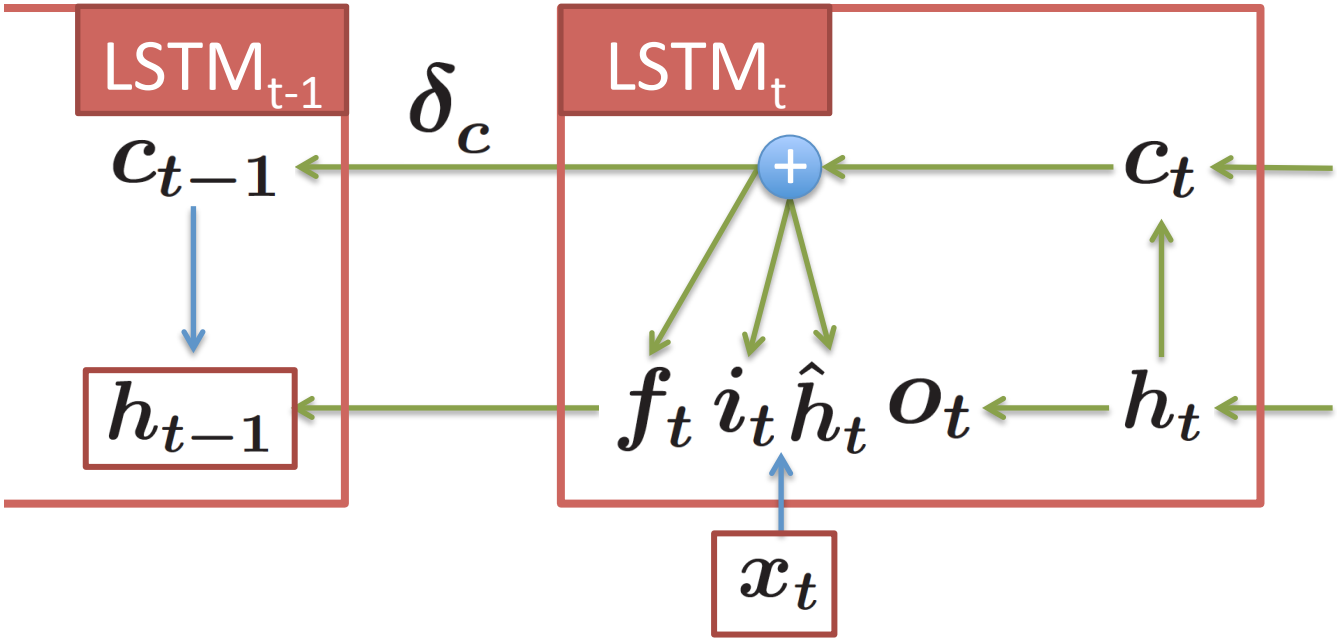
LSTM Backprop

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$



LSTM Backprop

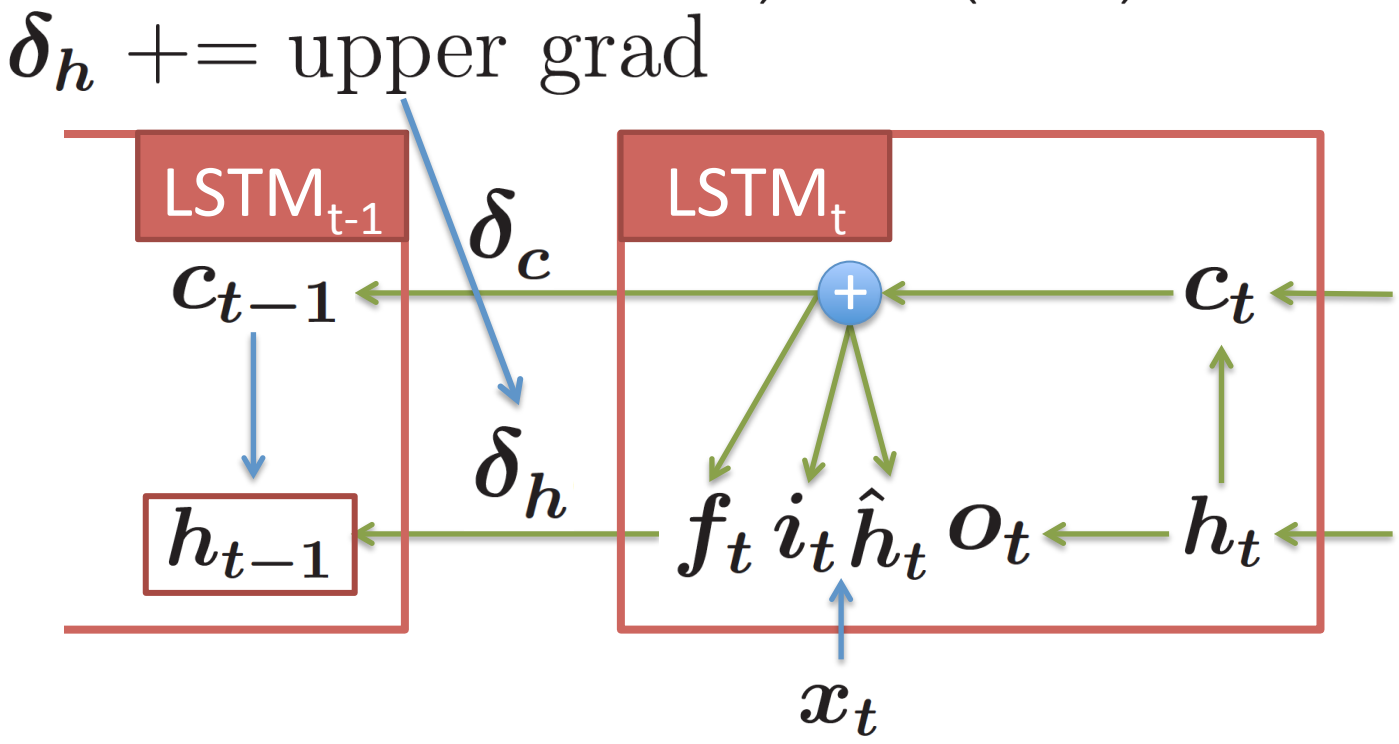
$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$



- Compute gradients for $T_{4n \times 2n}$, x_t, h_{t-1}.

LSTM Backprop

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ \hat{h}_t \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{4n \times 2n} \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$$



- Add gradients from the loss / upper layers.

Summary

- LSTM backpropagation is nasty.

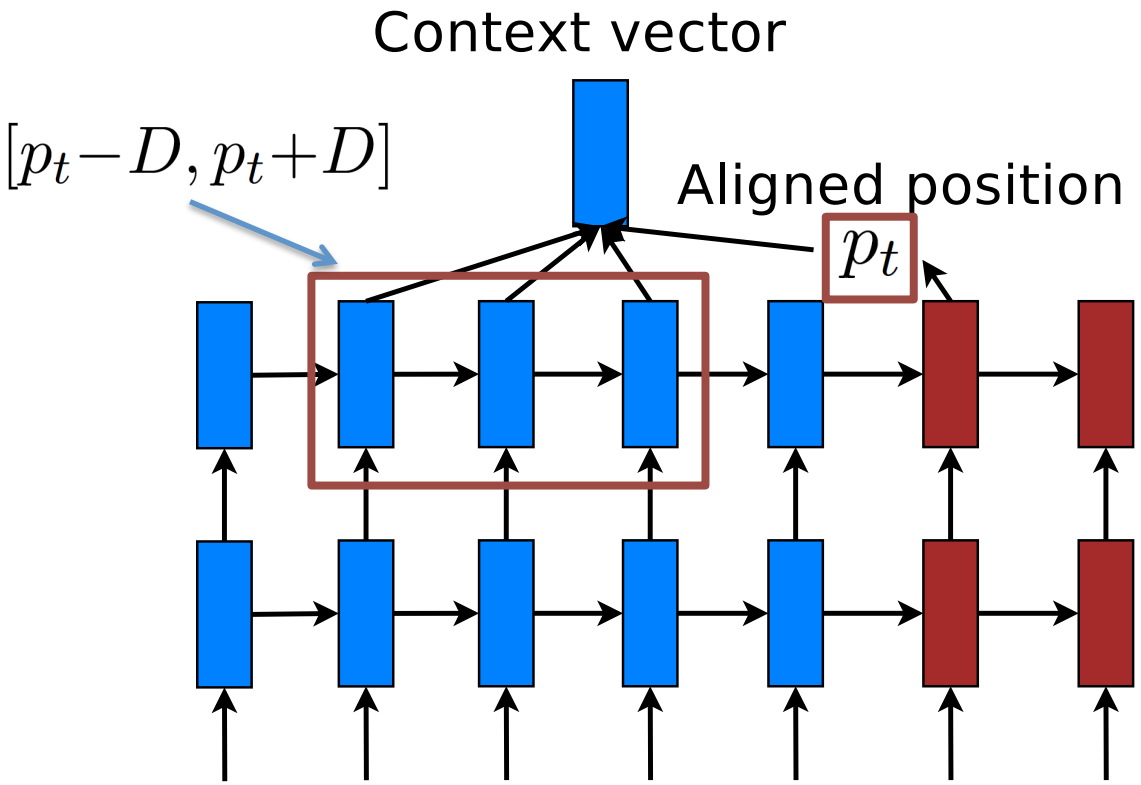
- But it will be much easier if:
  - Know your matrix calculus!
  - Pay attention to $\boldsymbol{\delta}_{\boldsymbol{c}}$ and $\boldsymbol{\delta}_{\boldsymbol{h}}$.

Other Attention Functions

- Content-based: $\boldsymbol{a}_t = Attend(\boldsymbol{h}_{t-1}, \bar{\boldsymbol{h}}_{1...S})$
- Location-based: $\boldsymbol{a}_t = Attend(\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1})$
  - (Graves, 2013): hand-writing synthesis model.
- Hybrid: $\boldsymbol{a}_t = Attend(\boldsymbol{h}_{t-1}, \boldsymbol{a}_{t-1}, \bar{\boldsymbol{h}}_{1...S})$
  - (Chorowski et al., 2015) for speech recognition.

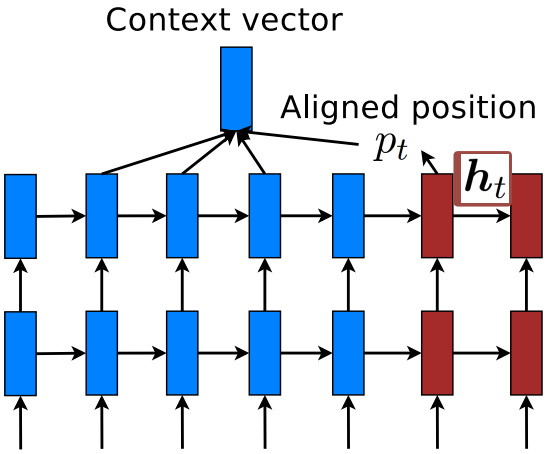
Local Attention (Luong et al., 2015b)

- More focused attention.
  - Potentially useful for longer text sequences.



aligned positions?

Predict aligned positions



[0,1]

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W}_p \boldsymbol{h}_t))$$

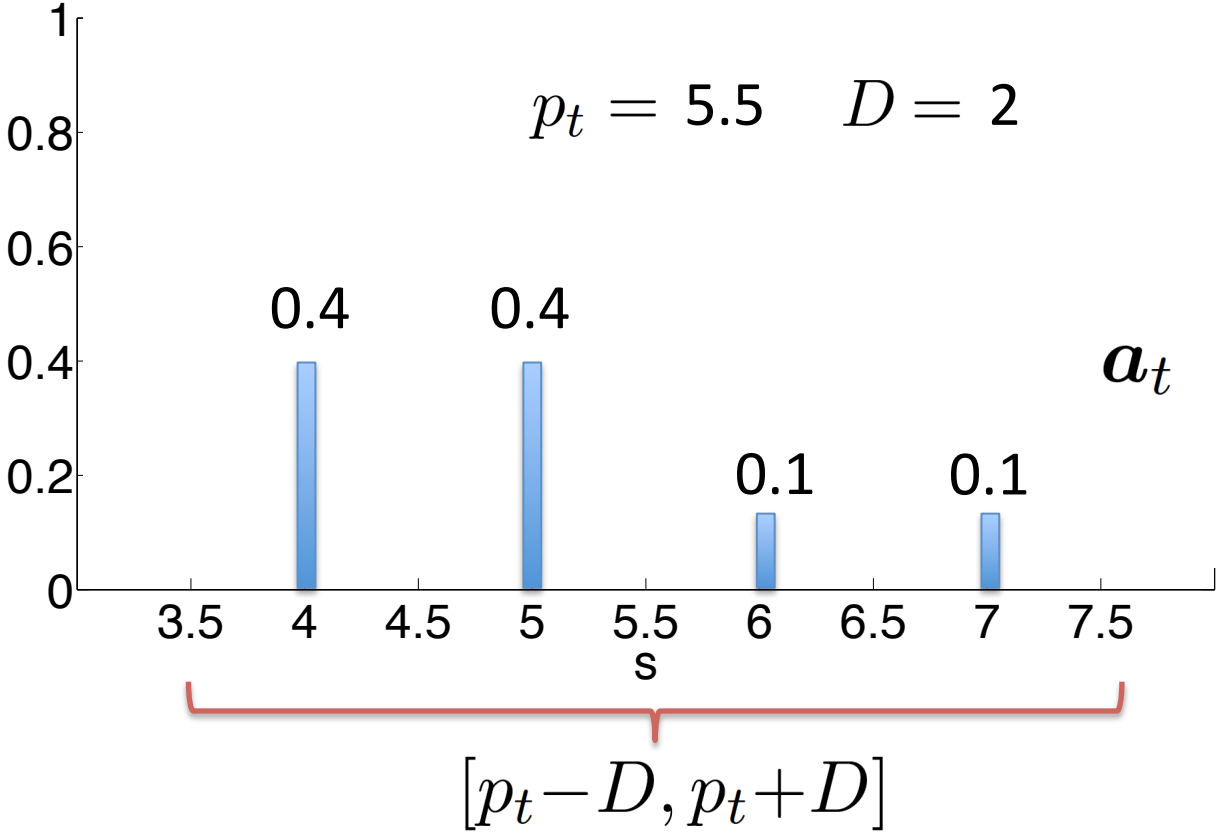
Real value in [0, S]

Source sentence

How do we learn to the position parameters?

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W_p h}_t))$$

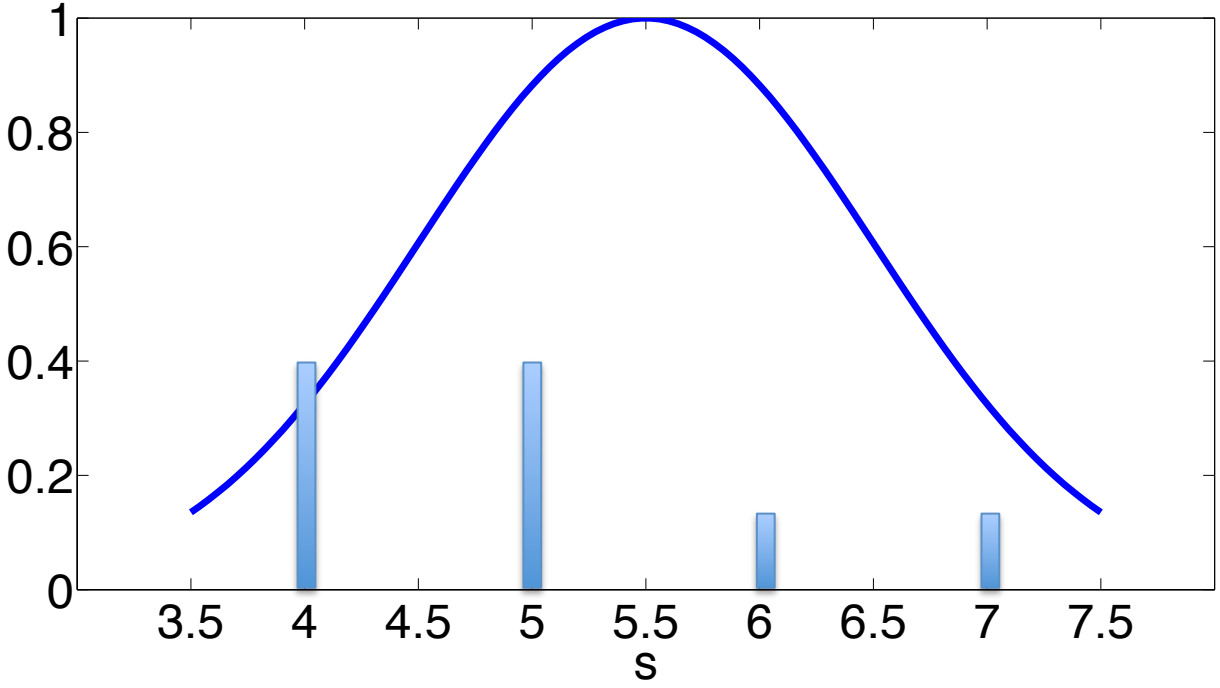
Alignment Weights



$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W_p h}_t))$$

Truncated Gaussian

$$\exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right)$$

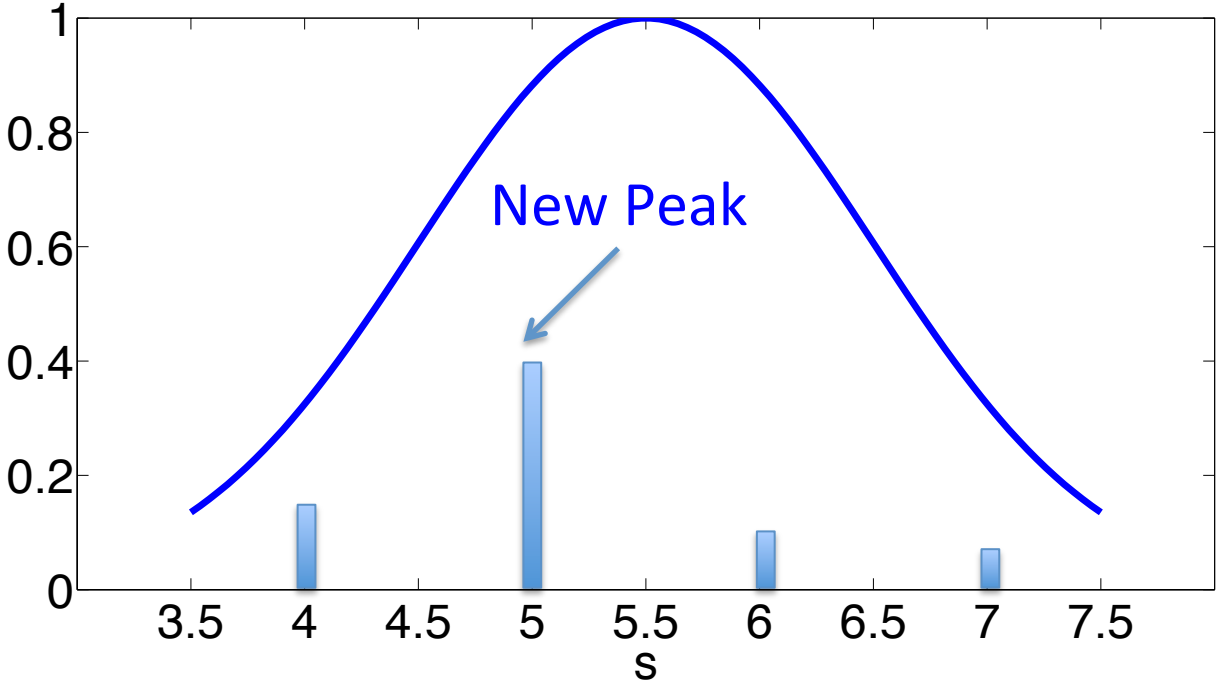


- Favor points close to the center.

$$p_t = S \cdot \text{sigmoid}(\boldsymbol{v}_p^\top \tanh(\boldsymbol{W}_{\boldsymbol{p}} \boldsymbol{h}_t))$$

Scaled Alignment Weights

$$\boldsymbol{a}_t(s) \;\; \exp\left(-\frac{(s - p_t)^2}{2\sigma^2}\right)$$



Differentiable almost everywhere!