Neural Machine Translation

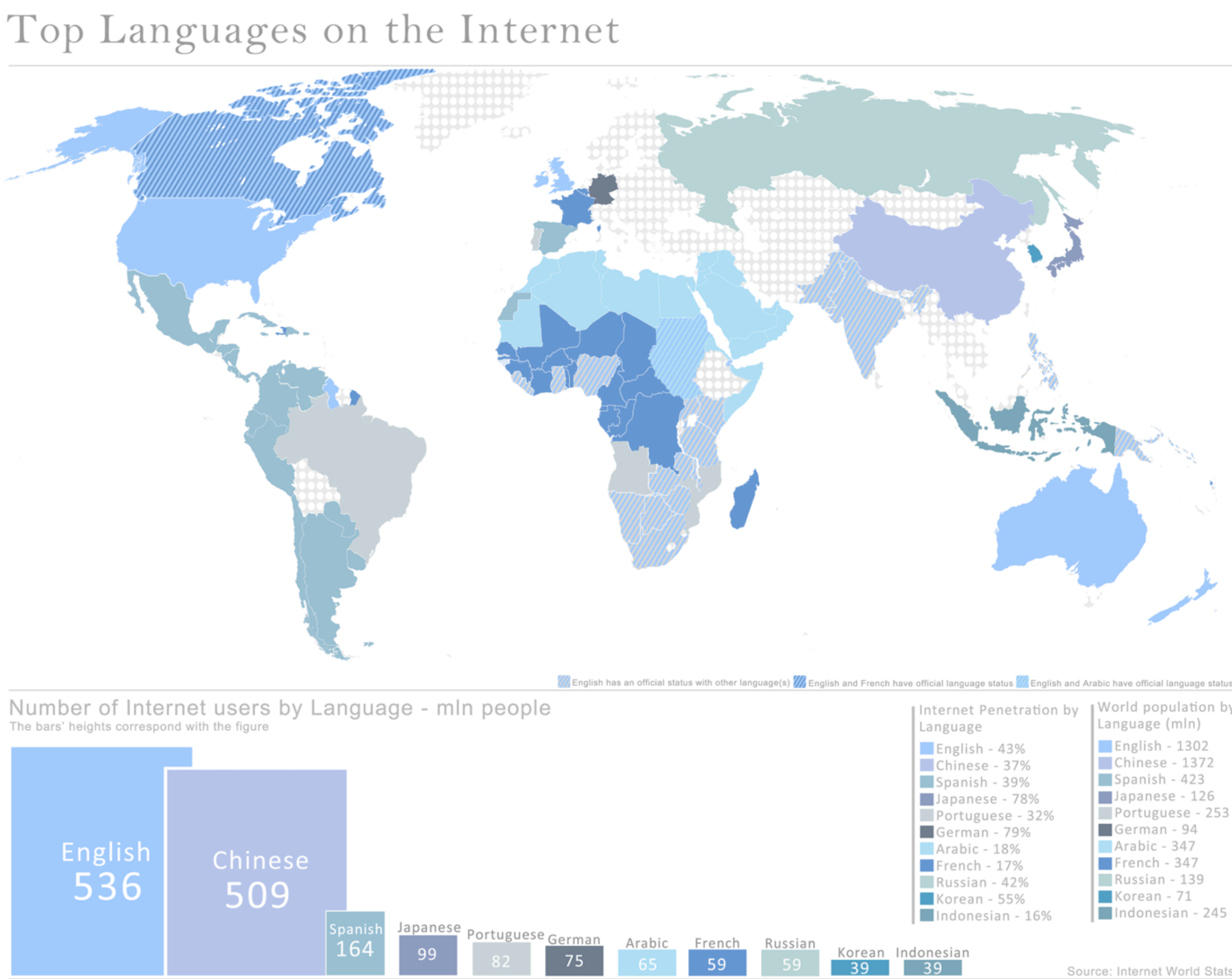
Thang Luong

Lecture @ CS224D

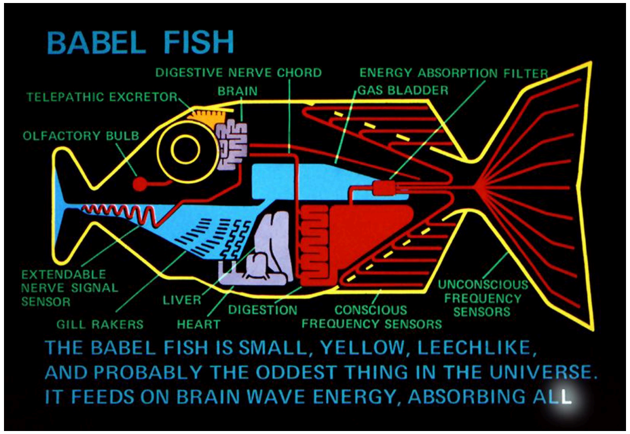
Spring 2016

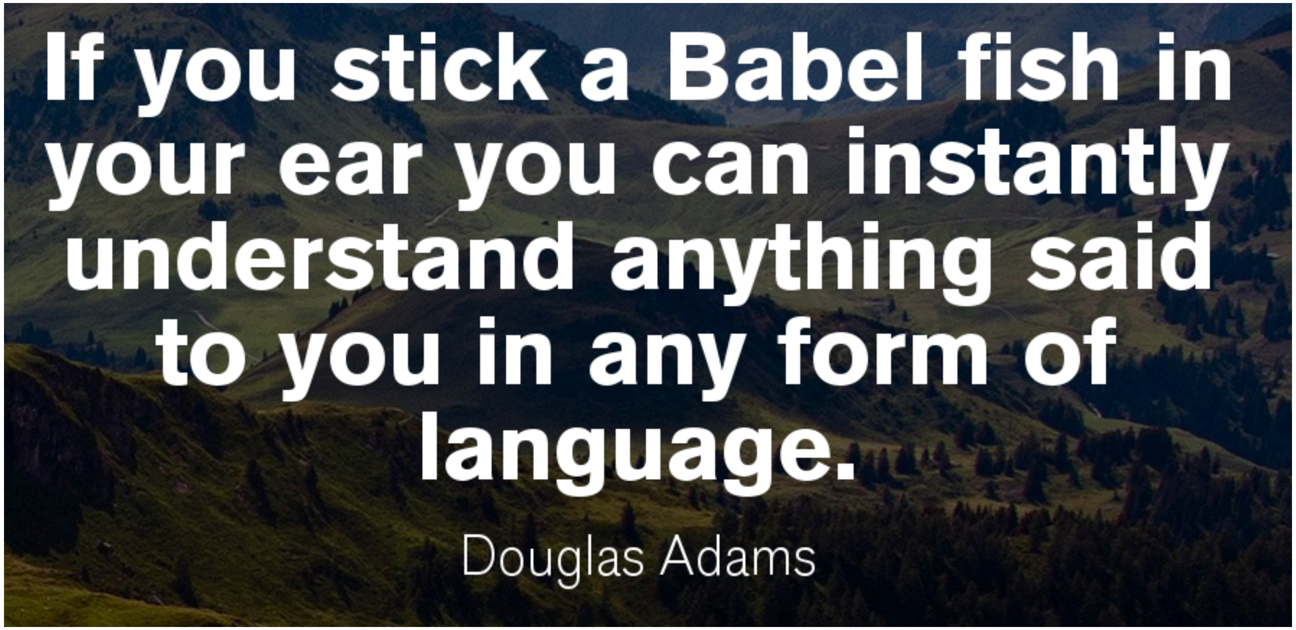
(Special thanks to Chris Manning for feedback!)

7 billion people, 7000 languages



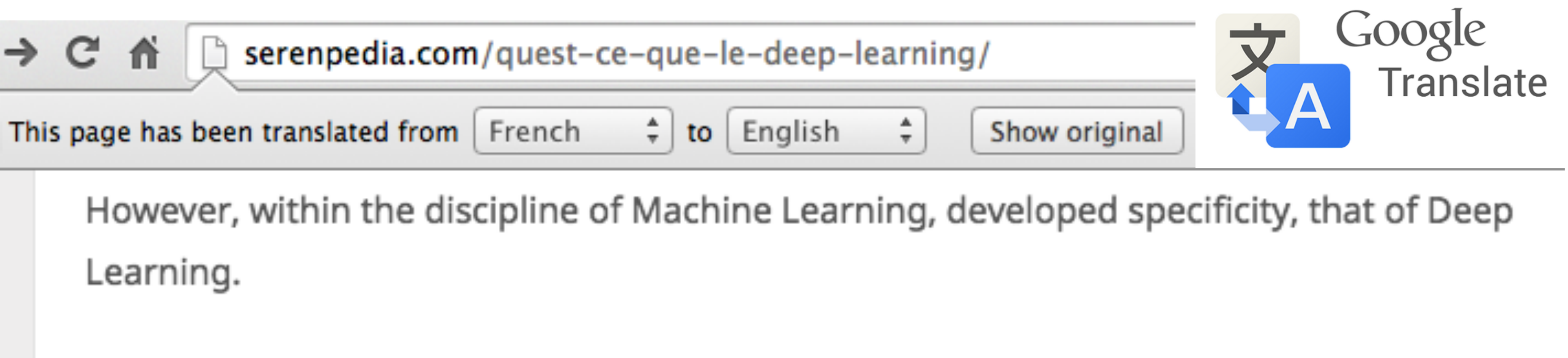
A universal translator





(The **Babel Fish** from "the Hitchhiker's Guide to the Galaxy")

Machine vs. Human Translation

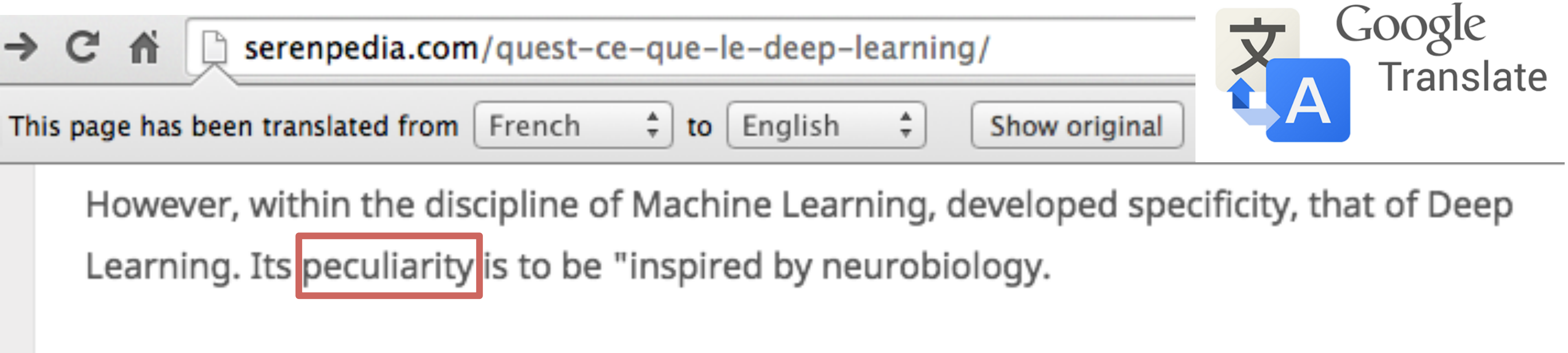


Faithful translation

"Nevertheless, within the discipline of machine learning, a specialization called deep learning has been developed.

- Grammatically incorrect.

Machine vs. Human Translation

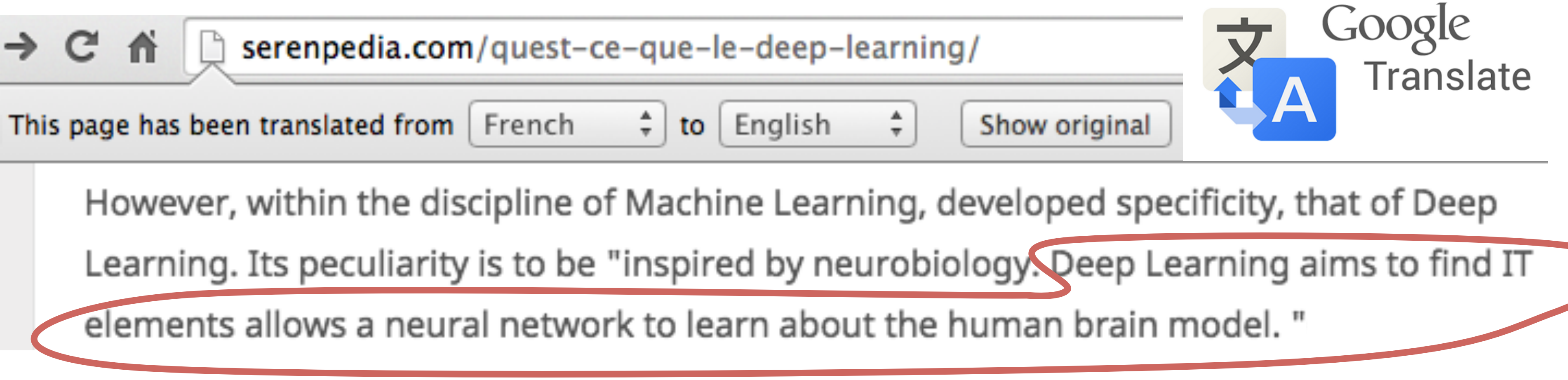


Faithful translation

"Nevertheless, within the discipline of machine learning, a specialization called deep learning has been developed. Its distinguishing feature is that it is inspired by neurobiology.

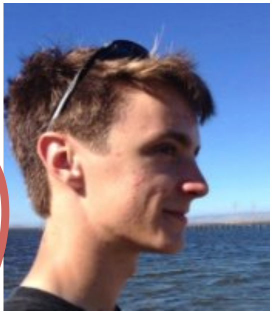
- Bad word choices.

Machine vs. Human Translation



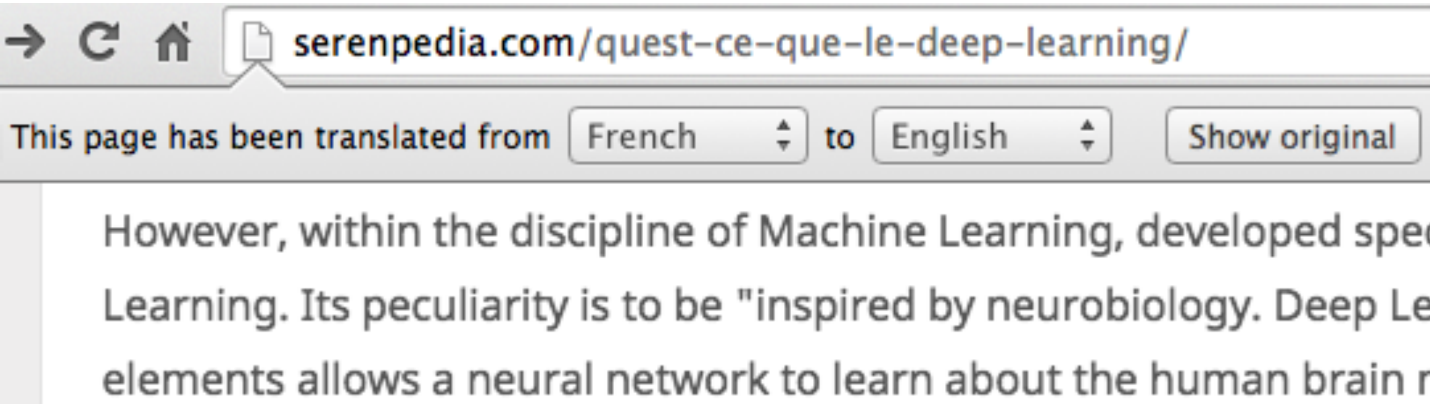
Faithful translation

"Nevertheless, within the discipline of machine learning, a specialization called deep learning has been developed. Its distinguishing feature is that it is inspired by neurobiology. Deep learning deals with computational elements which allow a network of artificial neurons to learn a model of the human brain."



- Bad sentence structures.

Machine vs. Human Translation



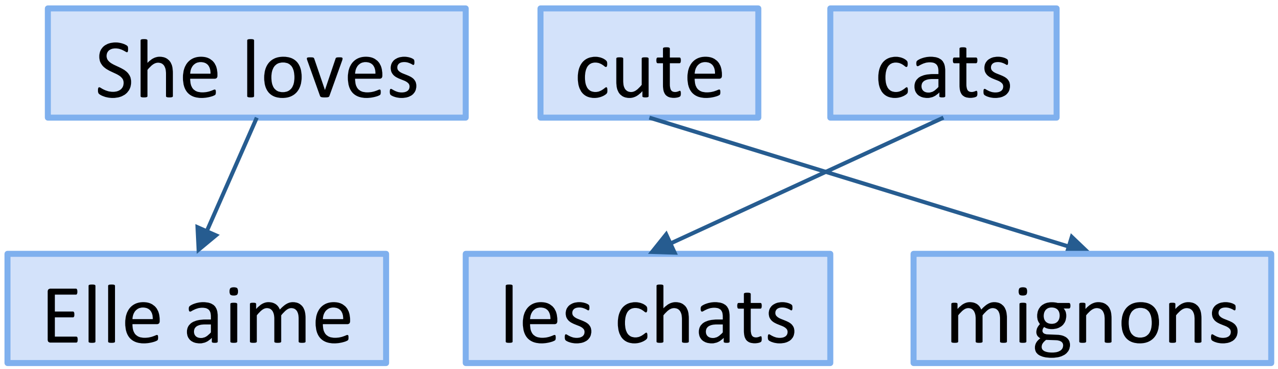
Fluent translation

"However, in machine learning a specialization called deep learning has emerged. It can be recognized by its distinctive neurobiological influence. Deep learning is centered around networks of artificial neurons which can learn models of the human brain."

A big gap!

How has MT evolved?

Phrase-based MT



(Brown et al., 1993; Koehn et al., 2003; Och & Ney, 2004)

- Break sentences into chunks.
- *Translation model*: look up phrase translations.
- *Language model*: tie phrases together.

Translate locally

LM uses only target words

Joint Neural Language Model

We went for a stroll along the South Bank

… allés pour une promenade le long de la rive

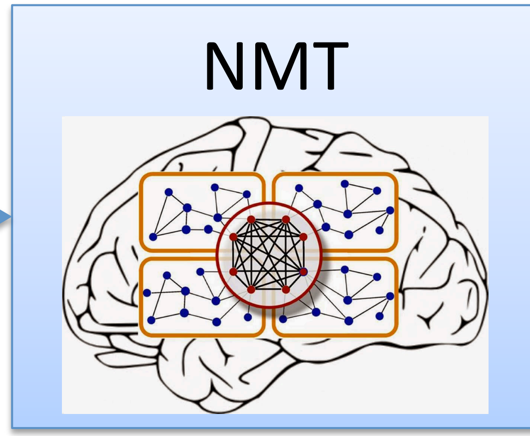
walk *along* *bank (river)*

- Conditioned on source words (Devlin et al., 2014)
- Still translate locally.

MT systems become more complex!

Neural Machine Translation to the rescue!

I am a student →



→ Je suis étudiant

(Sutskever et al., 2014; Cho et al., 2014)

- *Sequence-to-sequence*: translate globally.
- *End-to-end*: simple & generalizable.

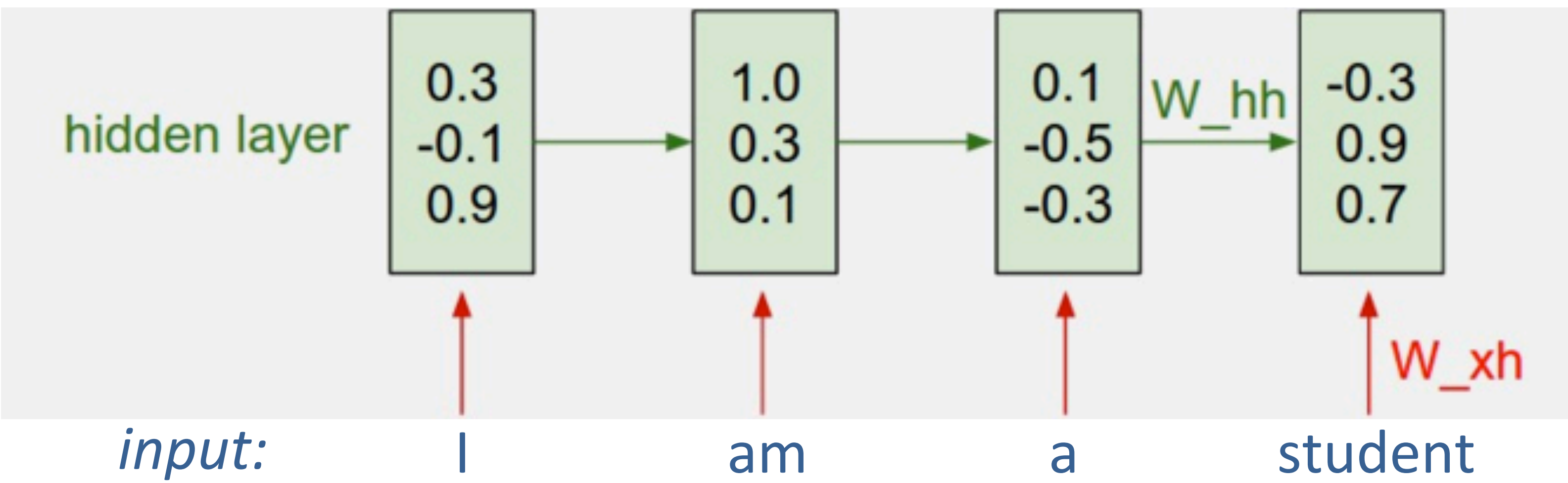
Let's find out!

Outline

- Basic NMT
  - RNN Recap.
  - Encoder-Decoder.
  - Training.
  - Testing.

- Advanced NMT

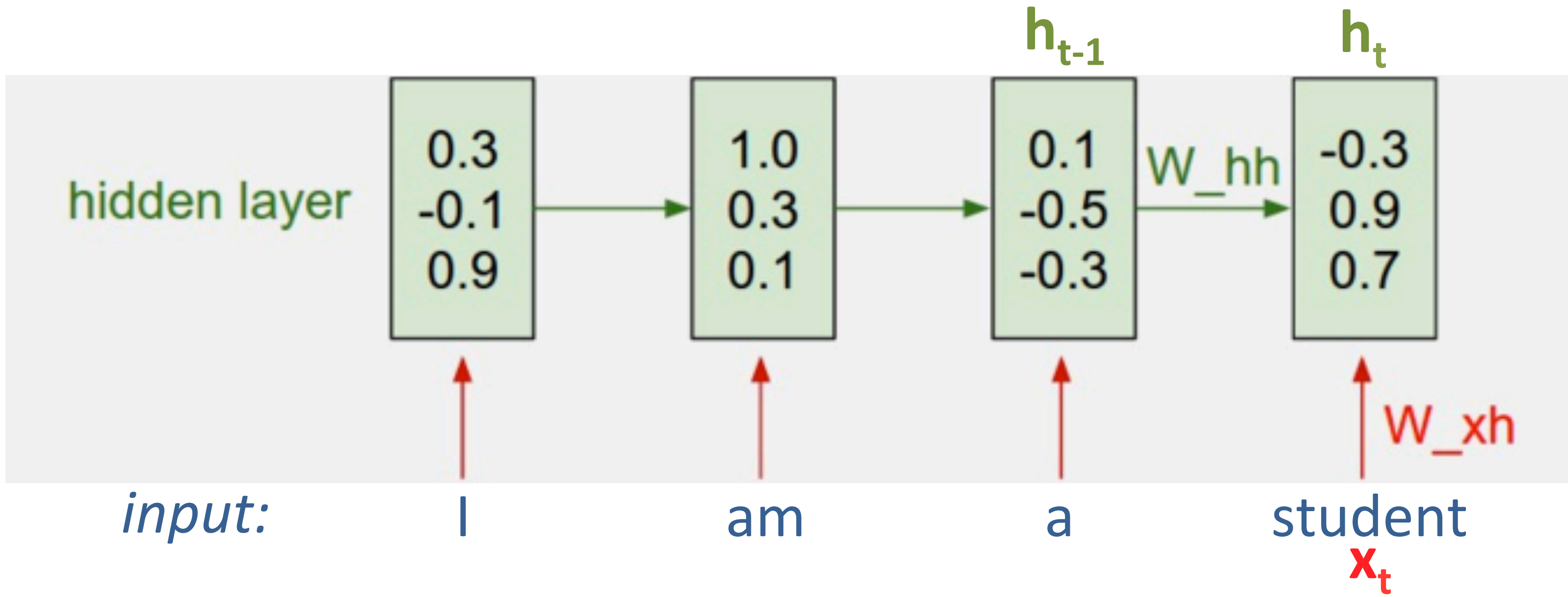
Recurrent Neural Networks (RNNs)



(Picture adapted from Andrej Karparthy)

Recurrent Neural Networks (RNNs)

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



RNNs to represent sequences!

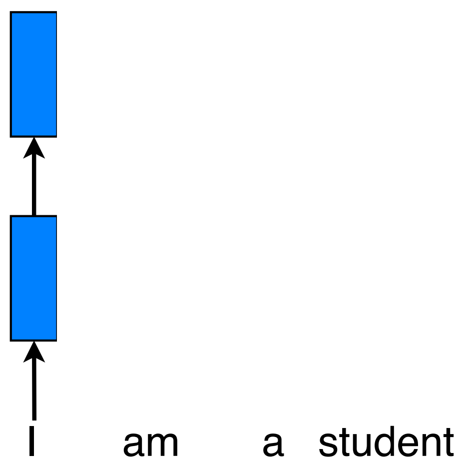
(Picture adapted from Andrej Karparthy)

Neural Machine Translation (NMT)

I am a student Je suis étudiant

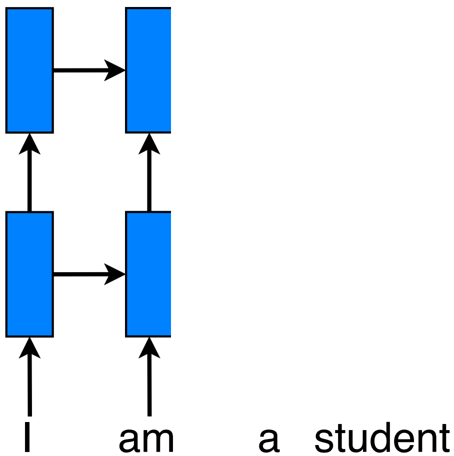
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



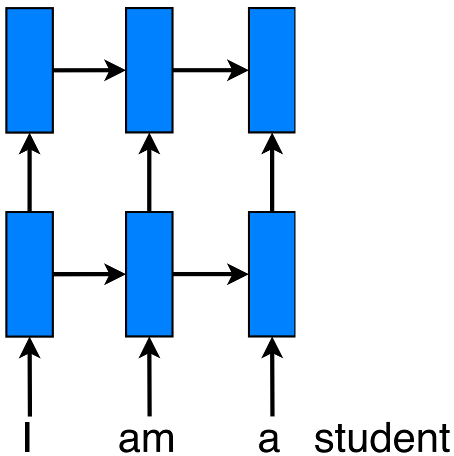
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



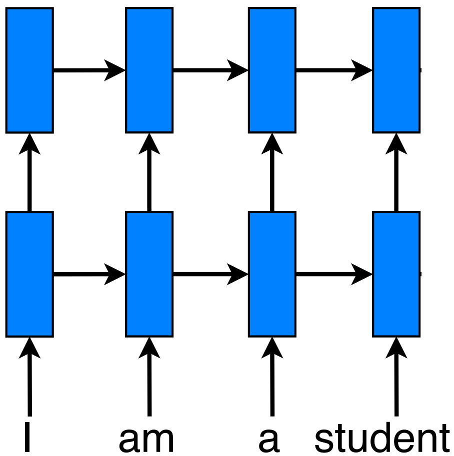
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



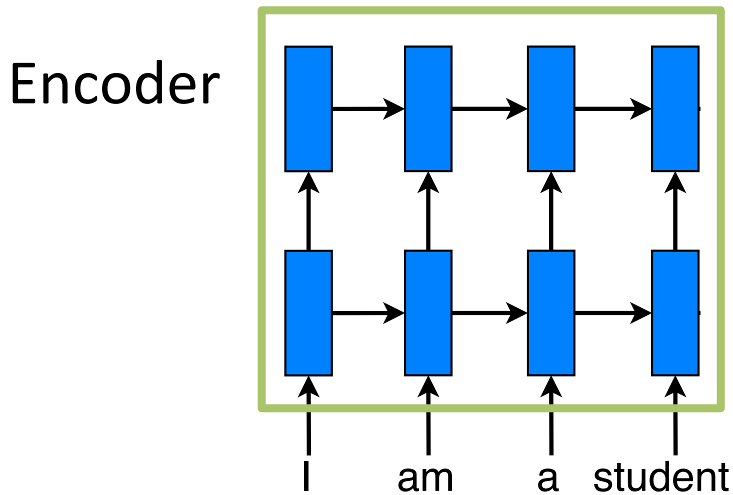
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



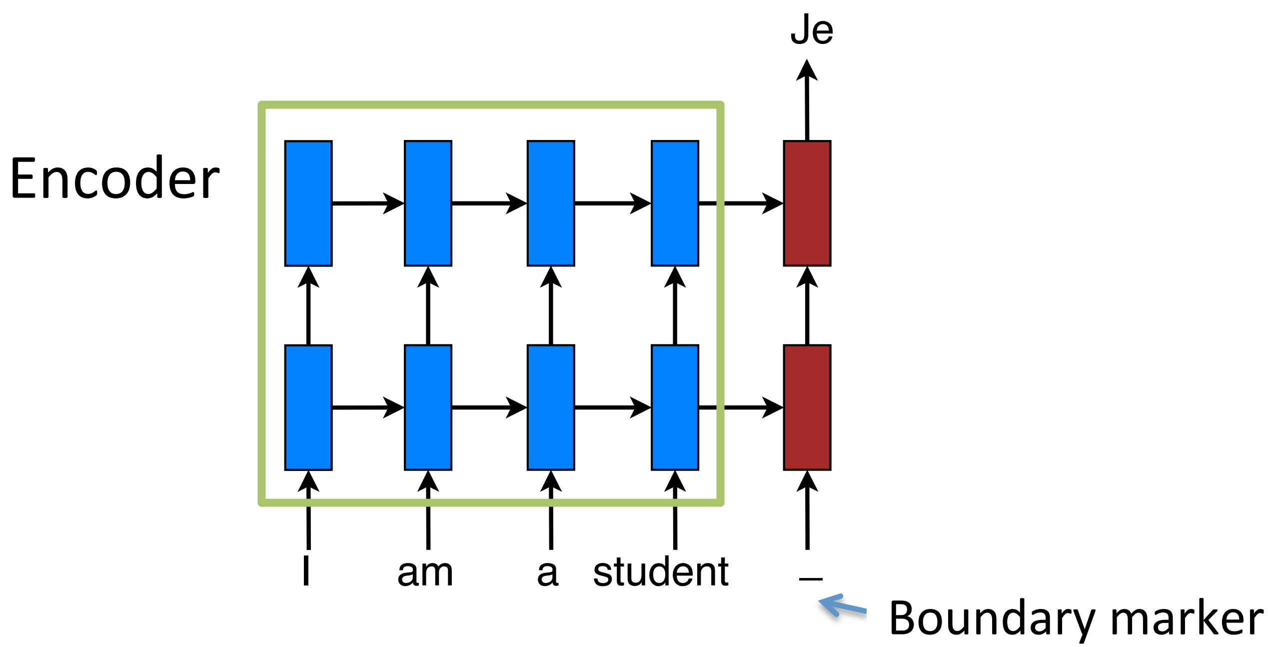
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



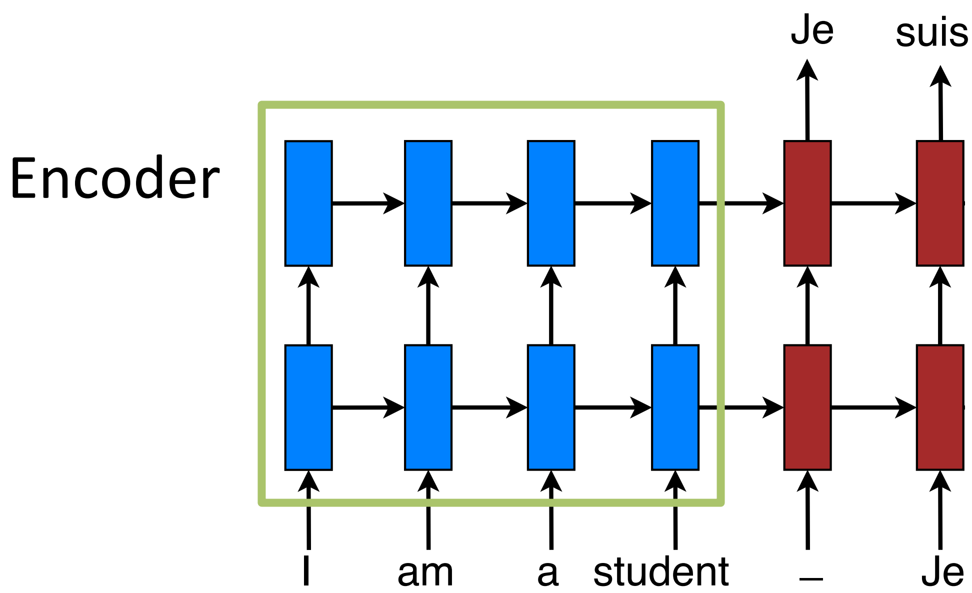
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



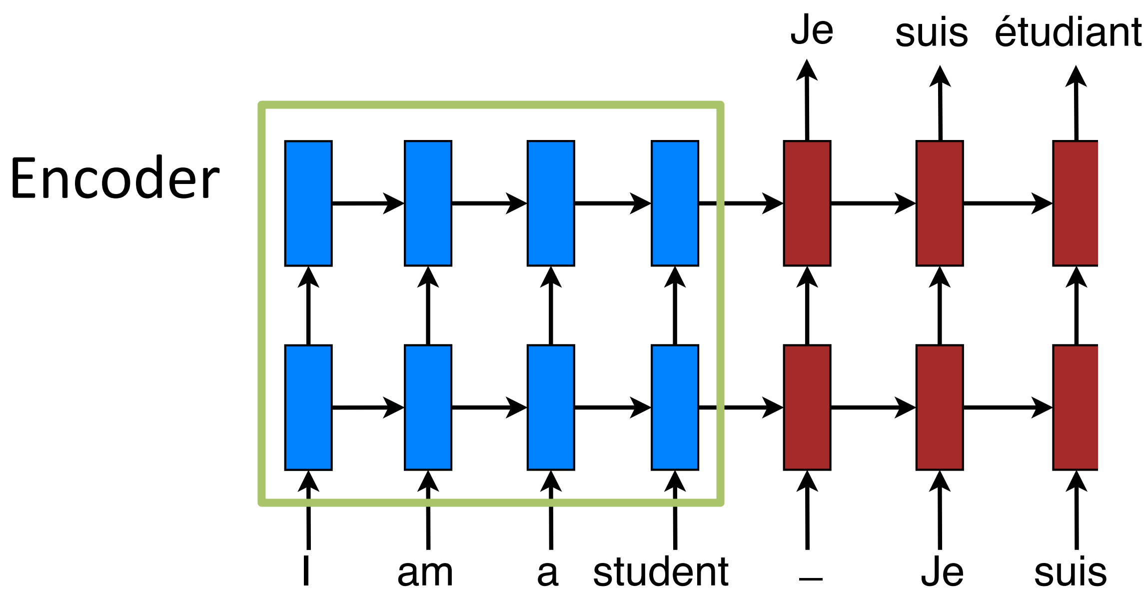
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



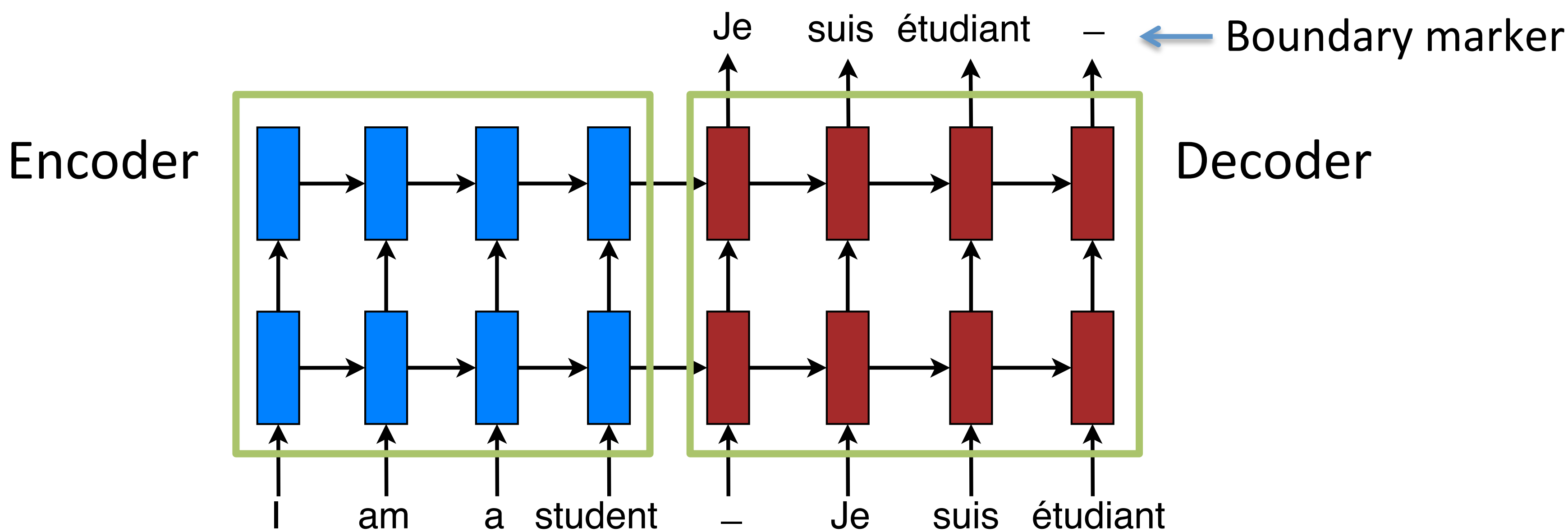
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



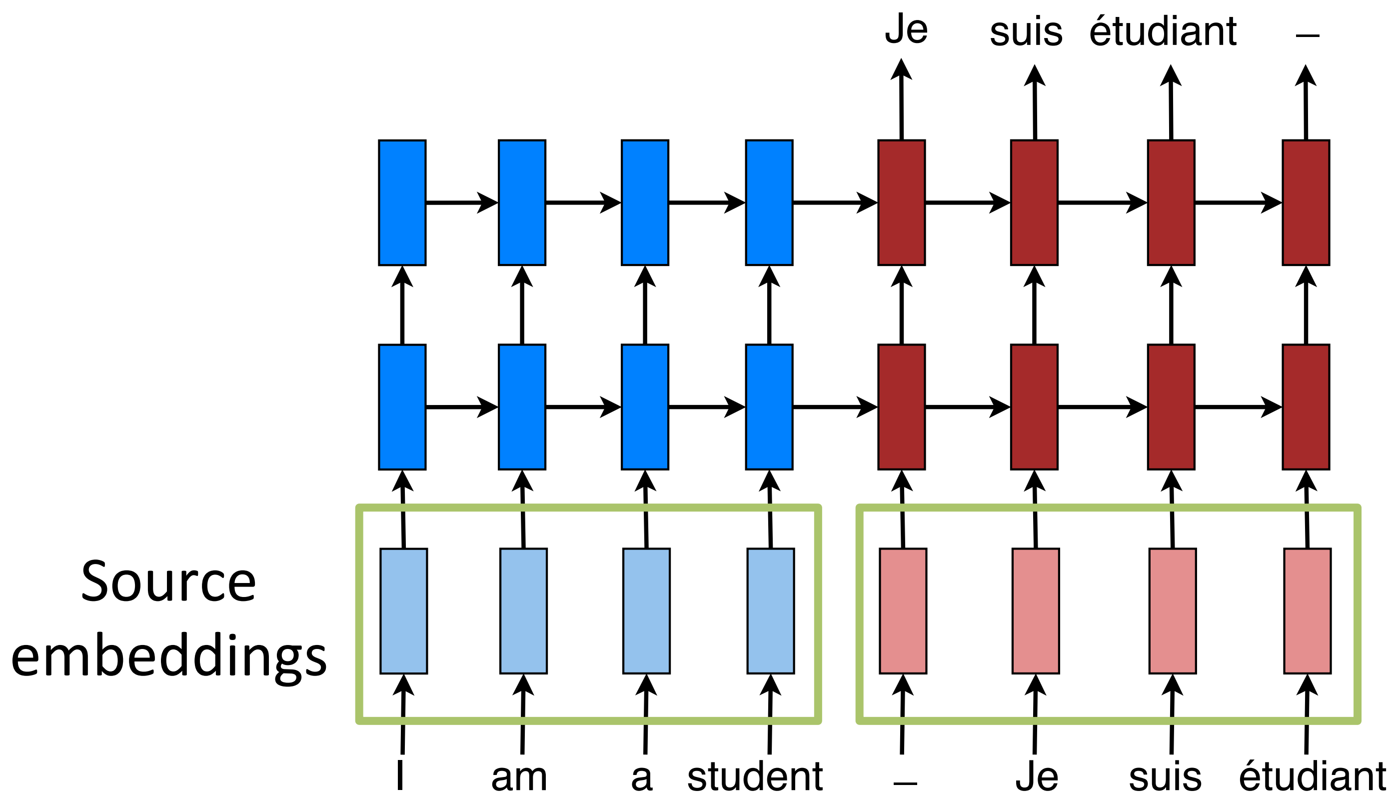
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Neural Machine Translation (NMT)



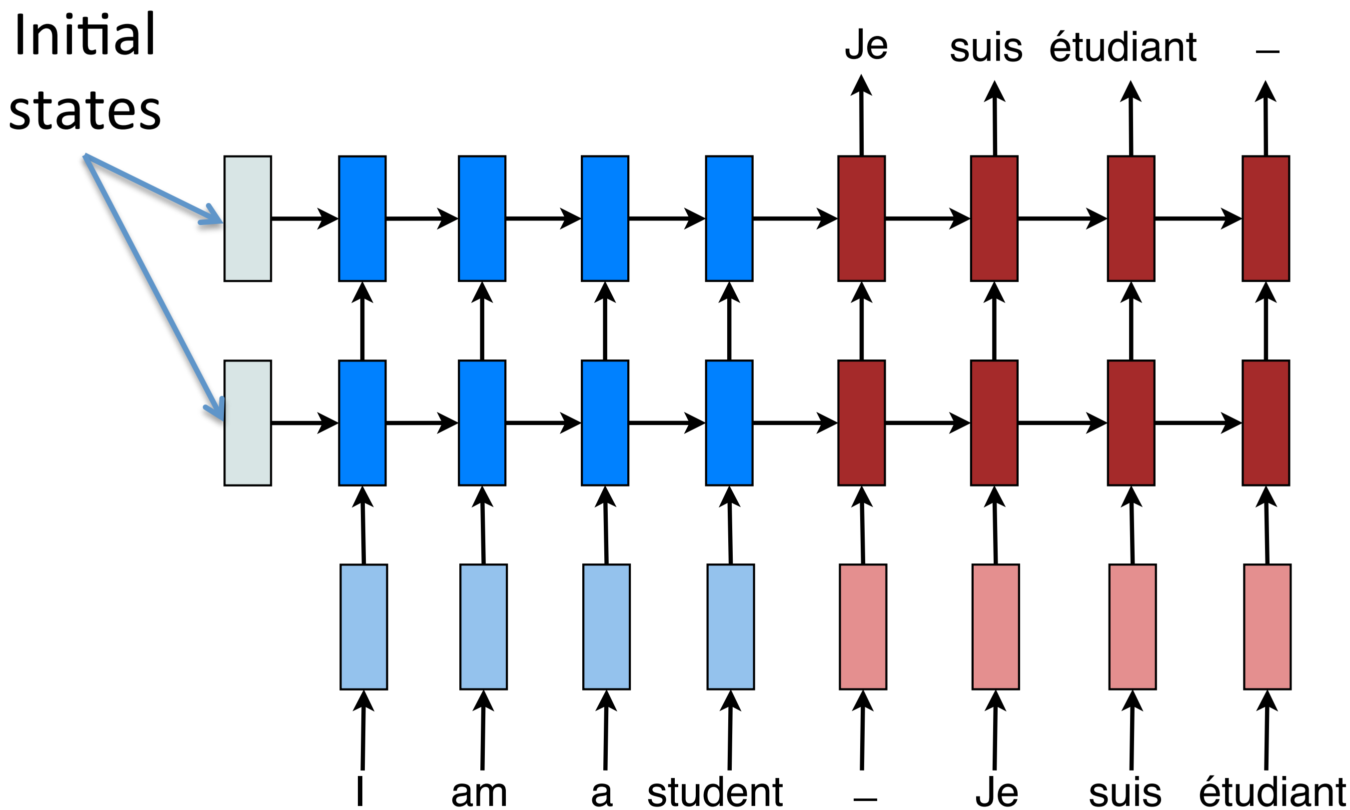
- Recurrent Neural Networks:
  - Model P(target | source) directly.
  - Can be trained end-to-end.

Word Embeddings



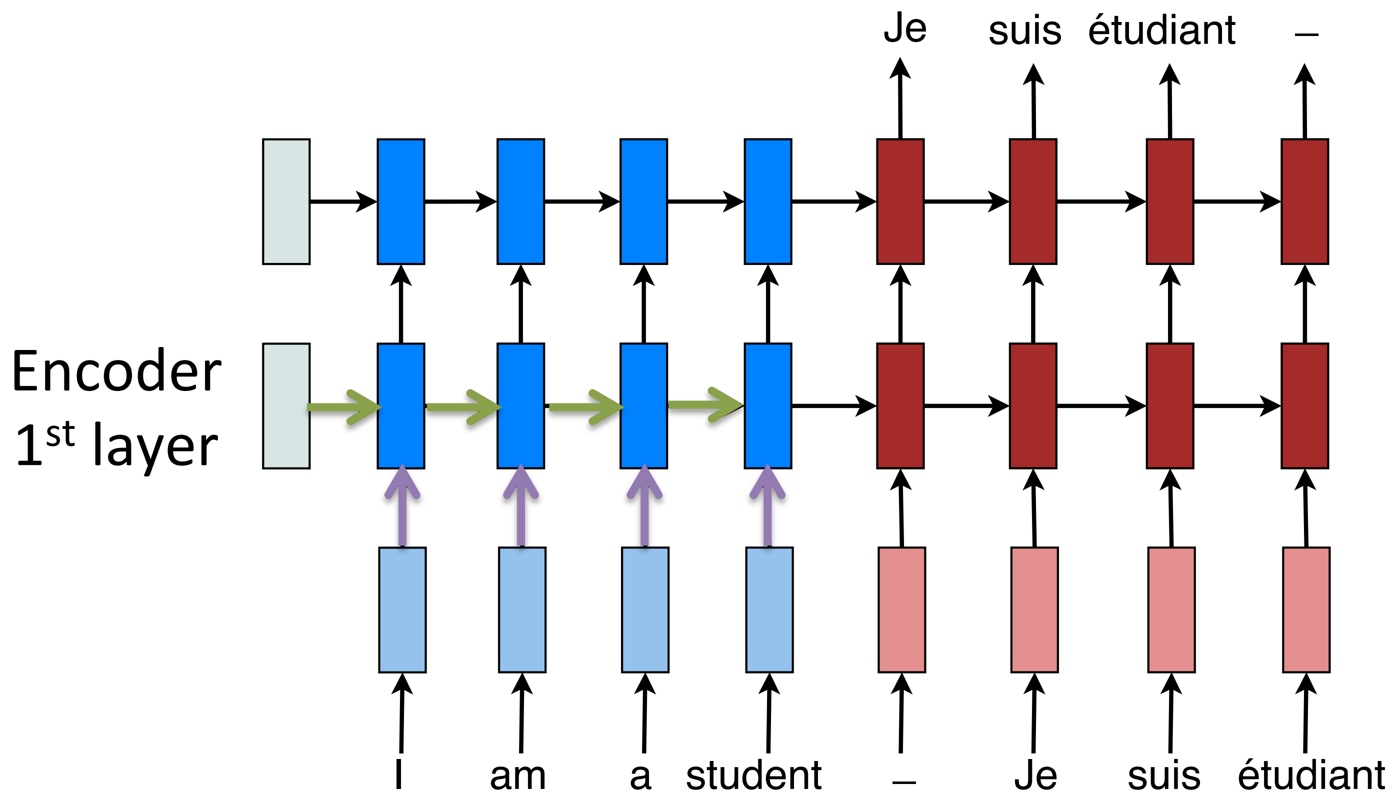
- One for each language: can learn from scratch.

Recurrent Connections



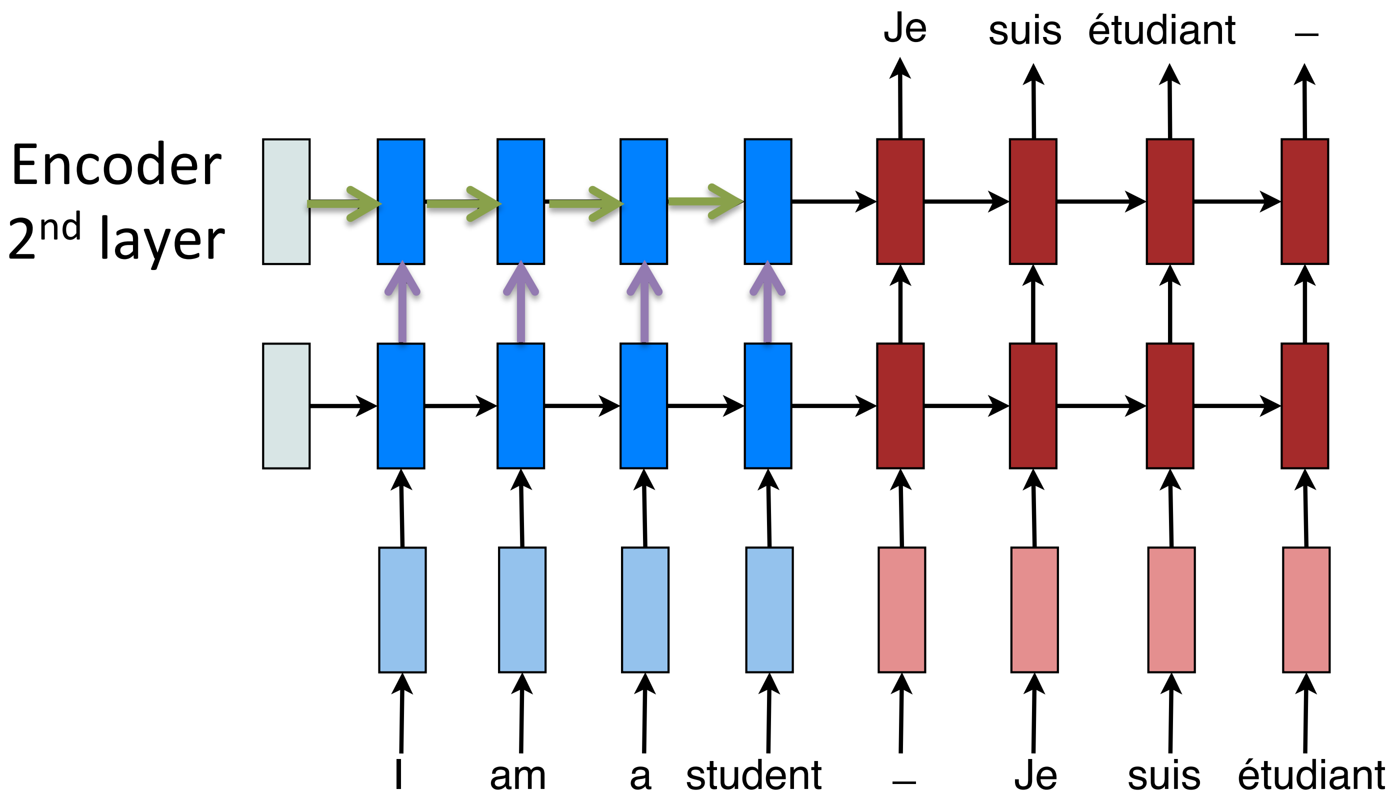
- Often set to 0.

Recurrent Connections



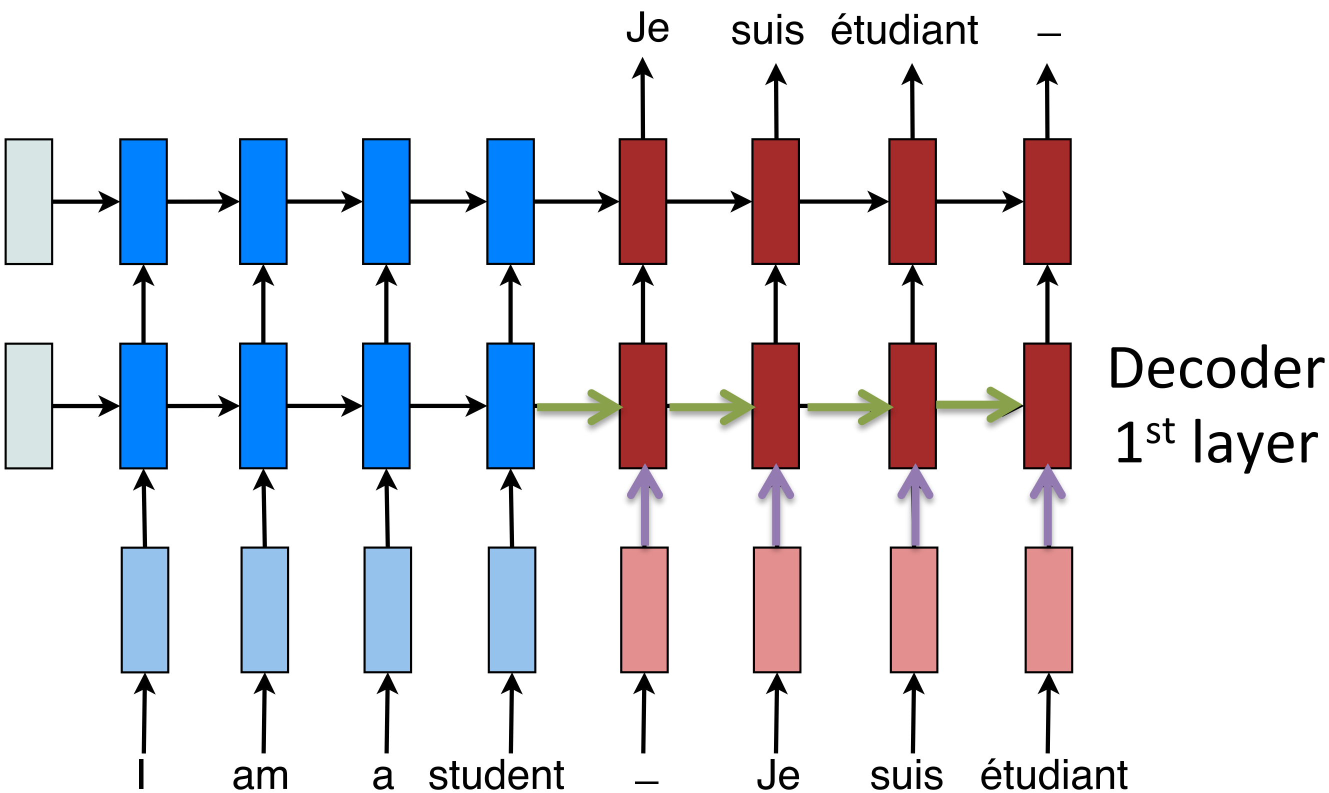
- Different: {1st layer, 2nd layer} x {encoder, decoder}.

Recurrent Connections



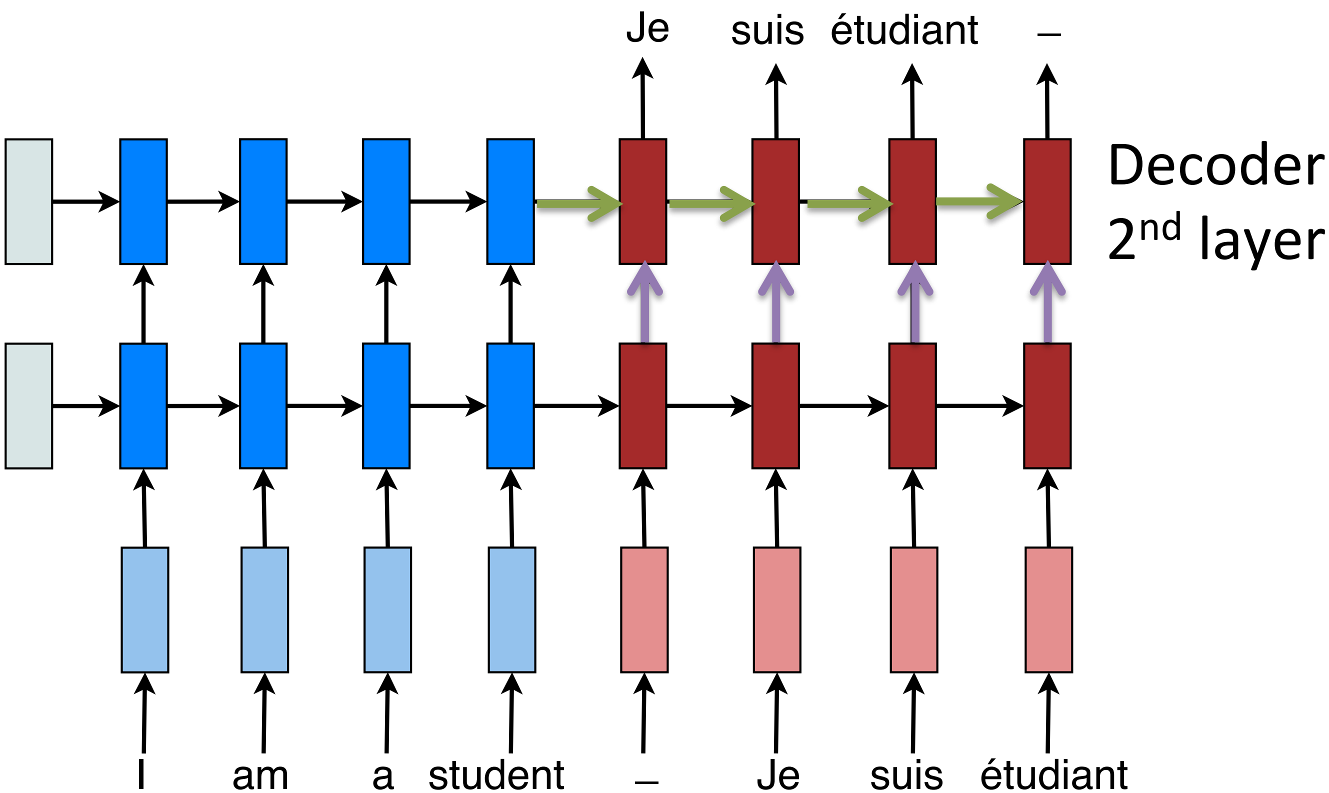
- Different: {1st layer, 2nd layer} x {encoder, decoder}.

Recurrent Connections



- Different: {1st layer, 2nd layer} x {encoder, decoder}.

Recurrent Connections



- Different: {1st layer, 2nd layer} x {encoder, decoder}.

Recurrent Units

- Vanilla:

$$h_{t-1} \rightarrow \boxed{\text{RNN}} \rightarrow h_t, \quad x_t \uparrow$$

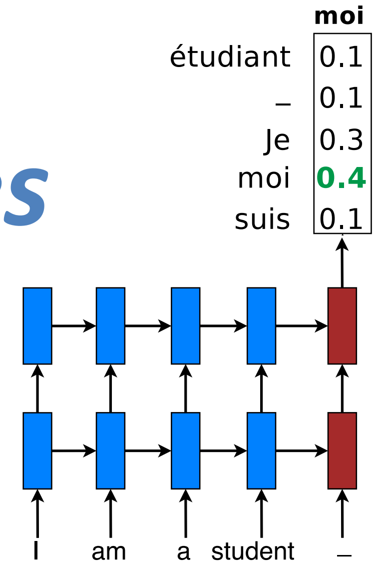
Vanishing gradient problem!

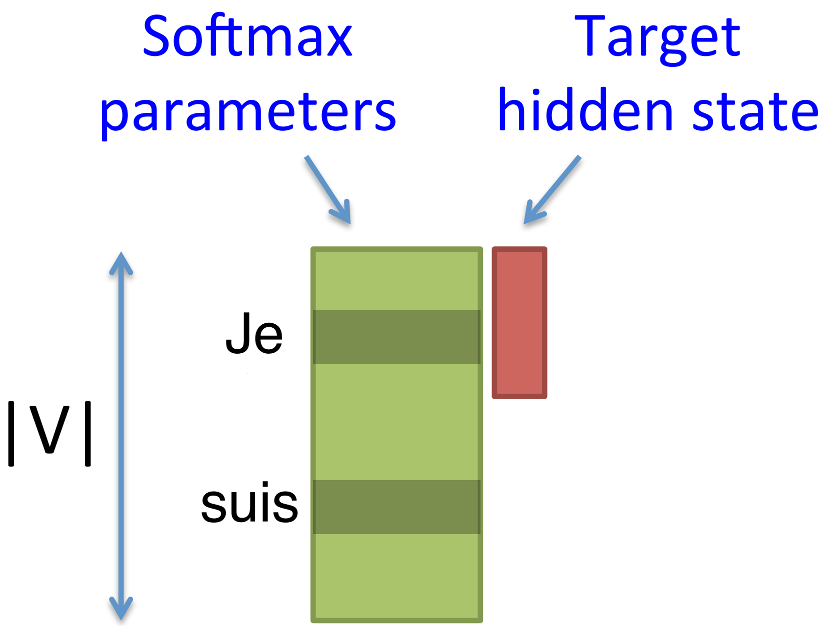
- LSTM:

$$\begin{matrix} h_{t-1} \rightarrow \\ c_{t-1} \rightarrow \end{matrix} \boxed{\text{LSTM}} \begin{matrix} \rightarrow h_t \\ \rightarrow c_t \end{matrix}, \quad x_t \uparrow$$

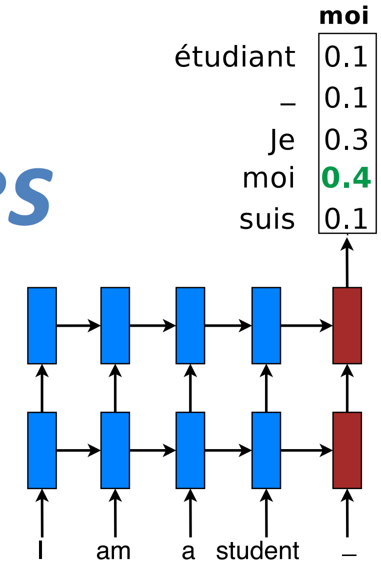
C'mon, it's been around for 20 years!

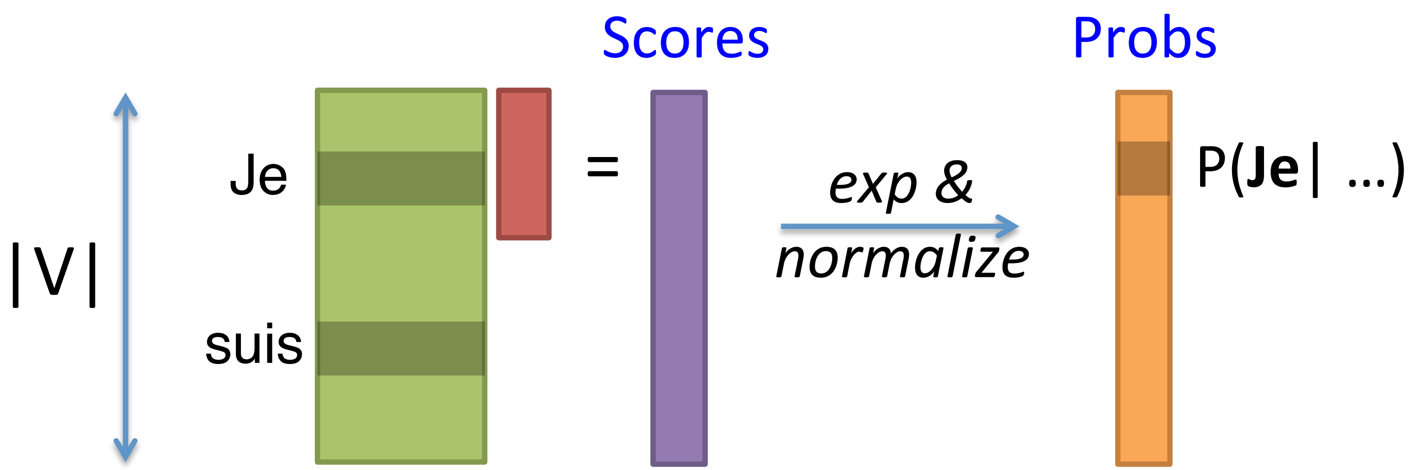
Softmax: *vectors* ↦ *categories*





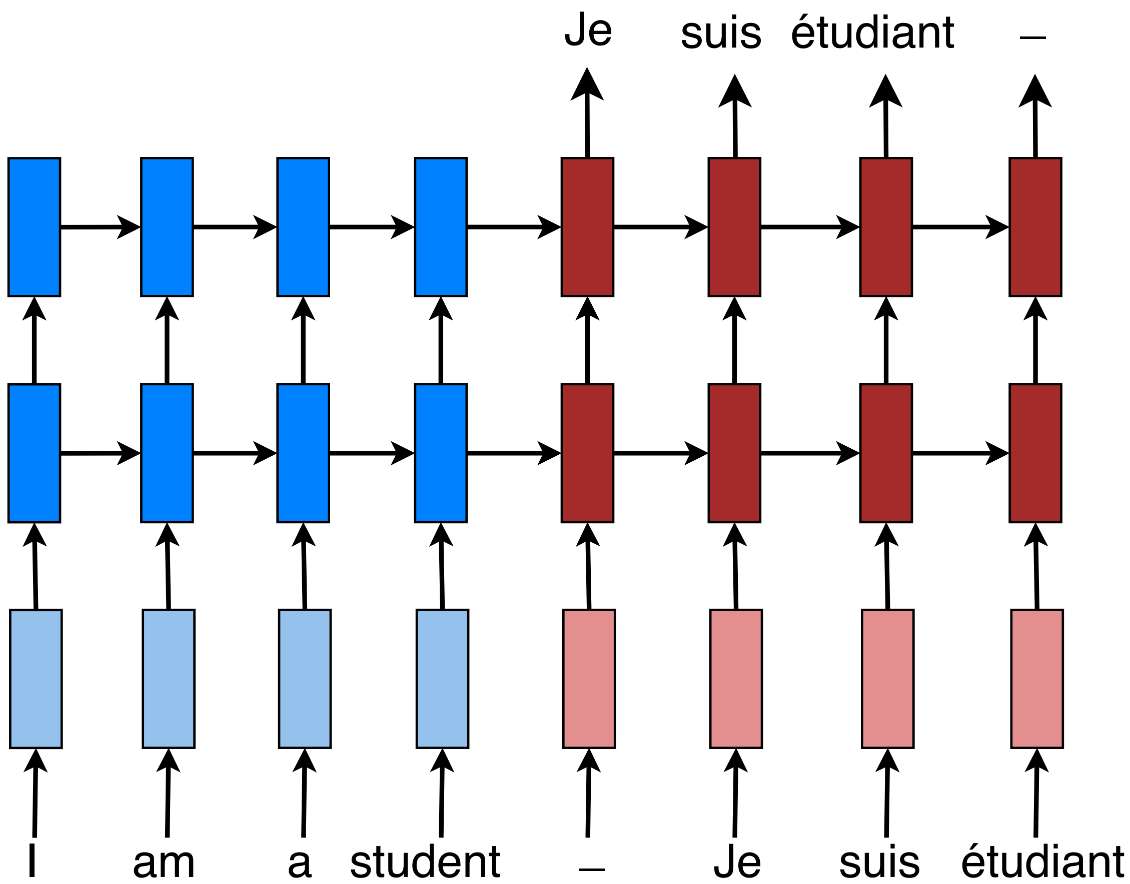
Softmax: *vectors* ↦ *categories*





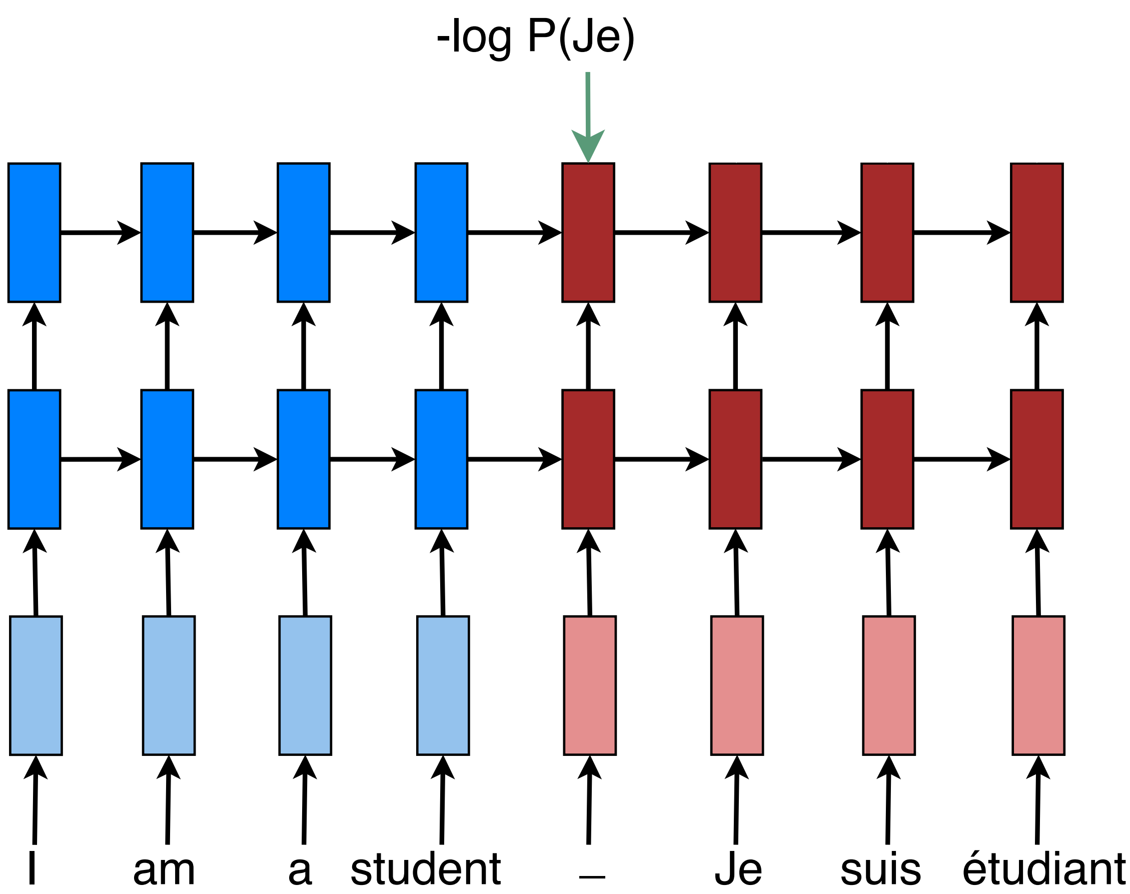
- Hidden states ↦ scores ↦ probabilities.

Training Loss



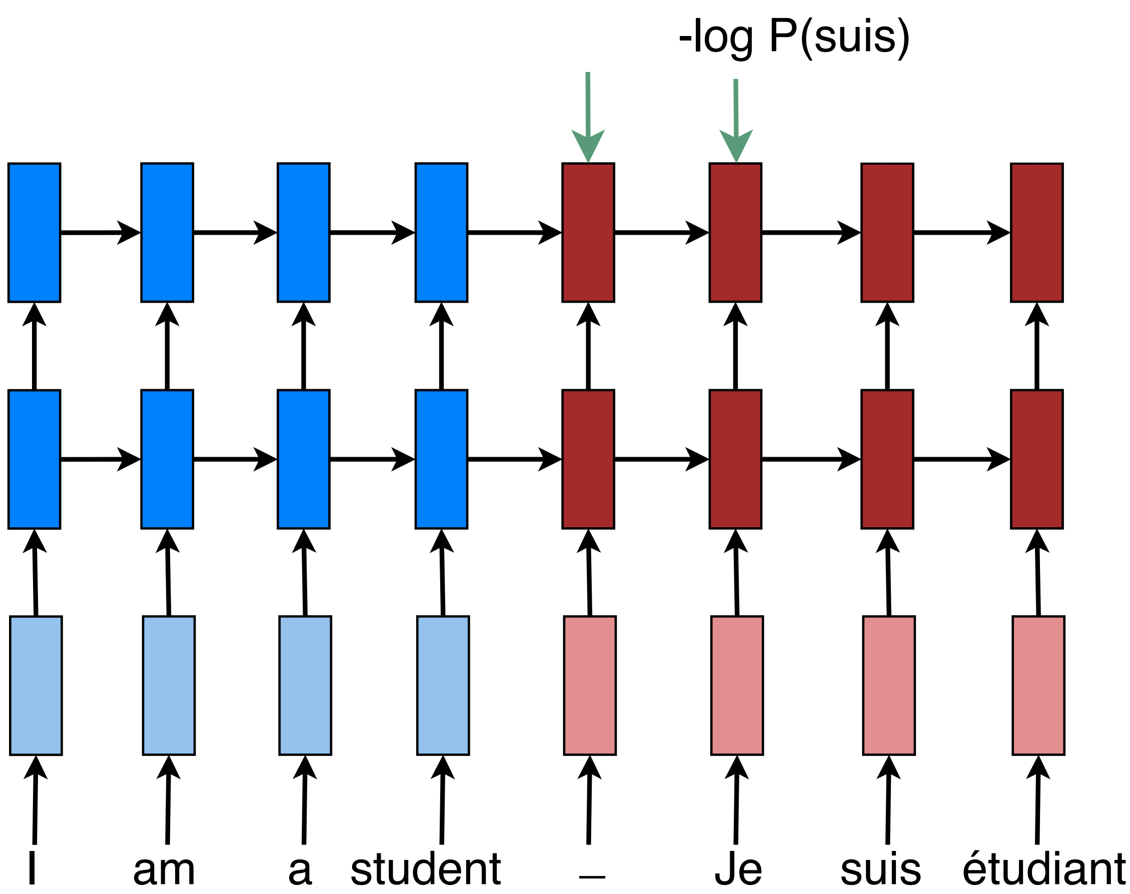
- Maximize P(target | source)

Training Loss



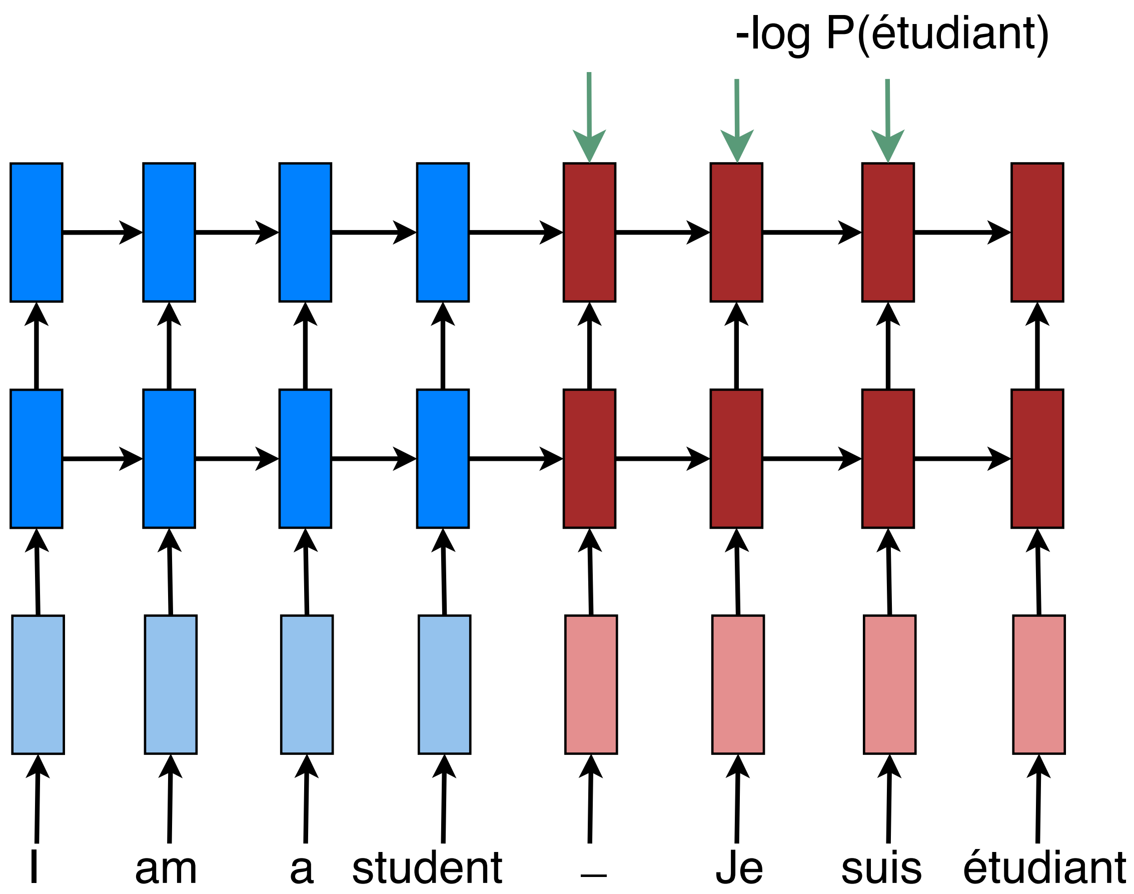
- Sum of all individual losses

Training Loss



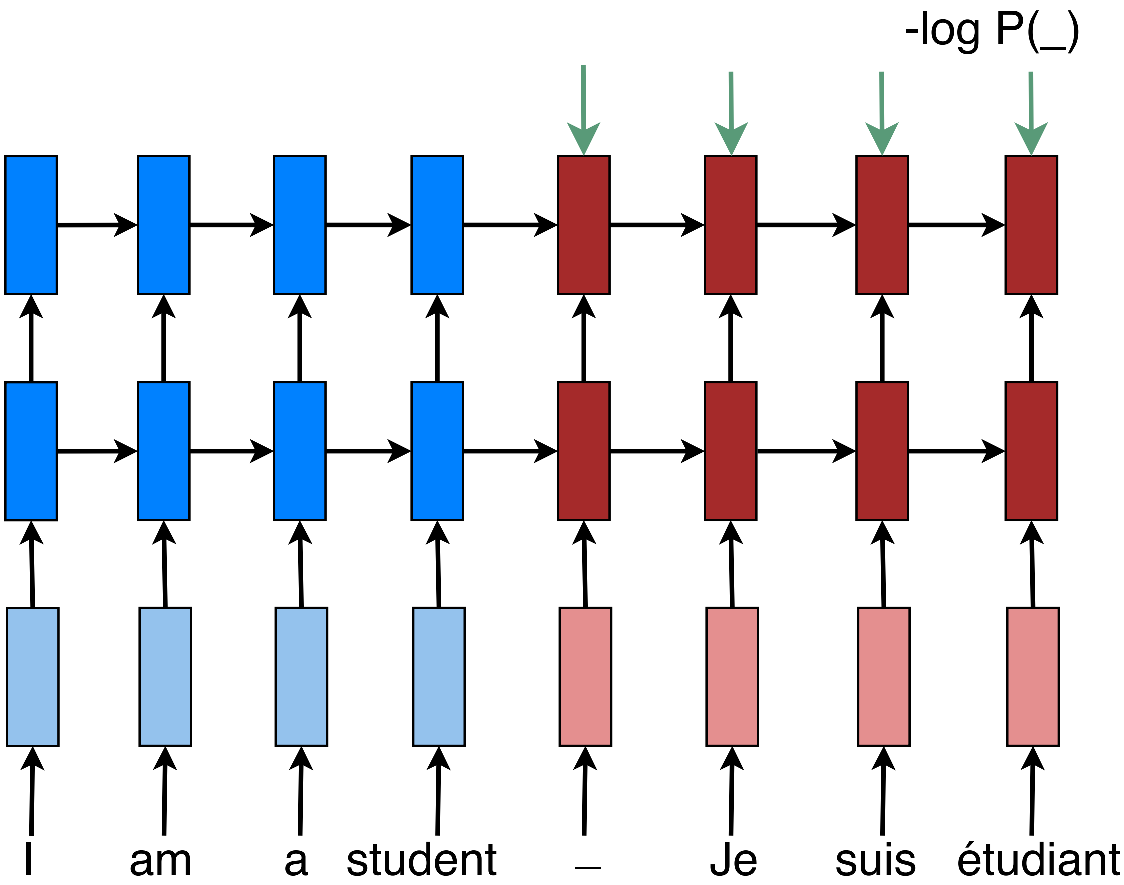
- Sum of all individual losses

Training Loss



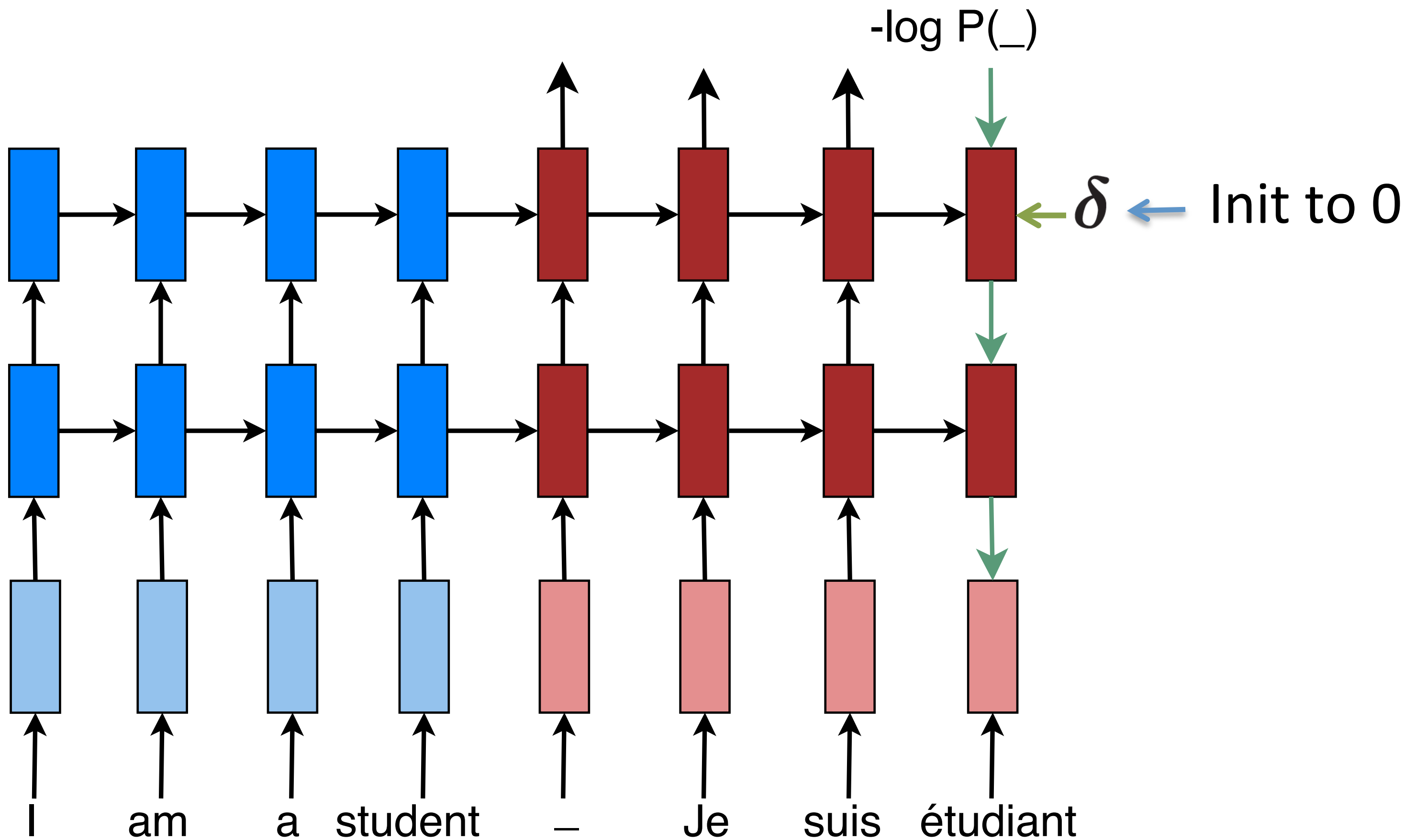
- Sum of all individual losses

Training Loss

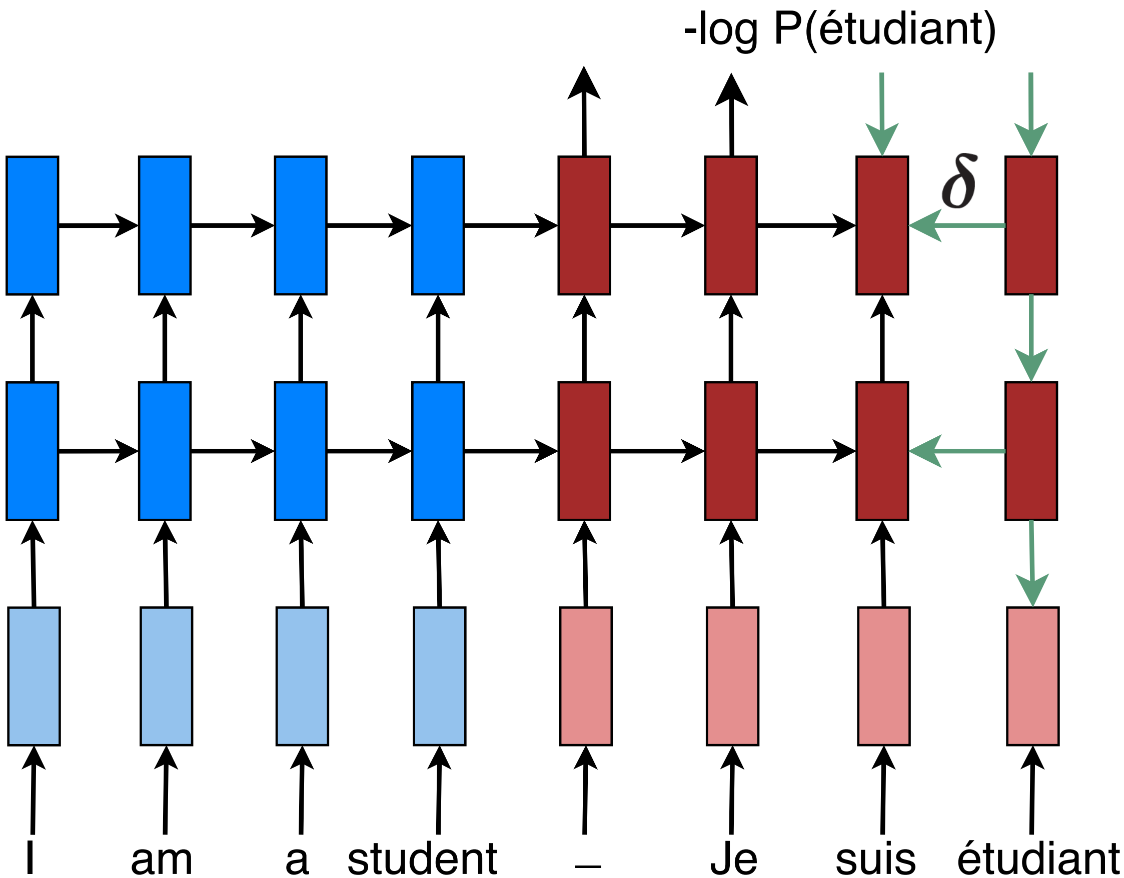


- Sum of all individual losses

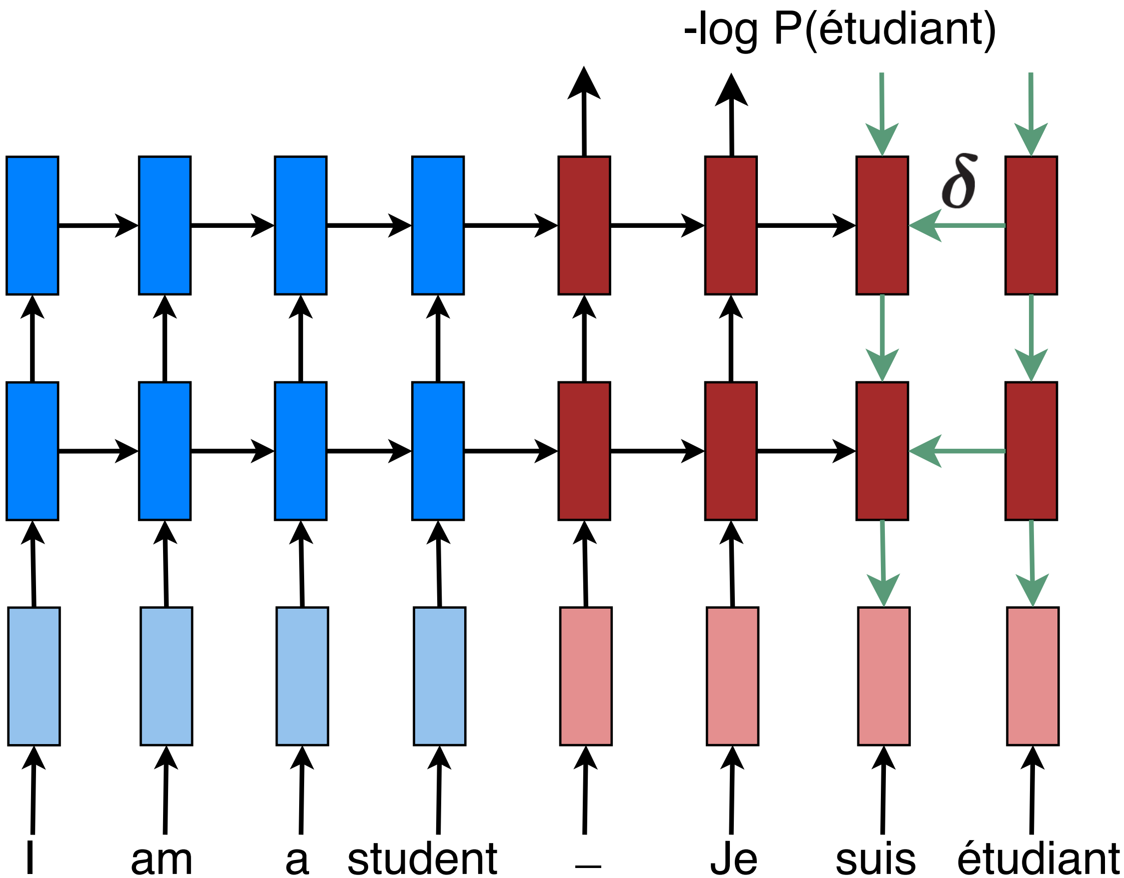
Backpropagation Through Time



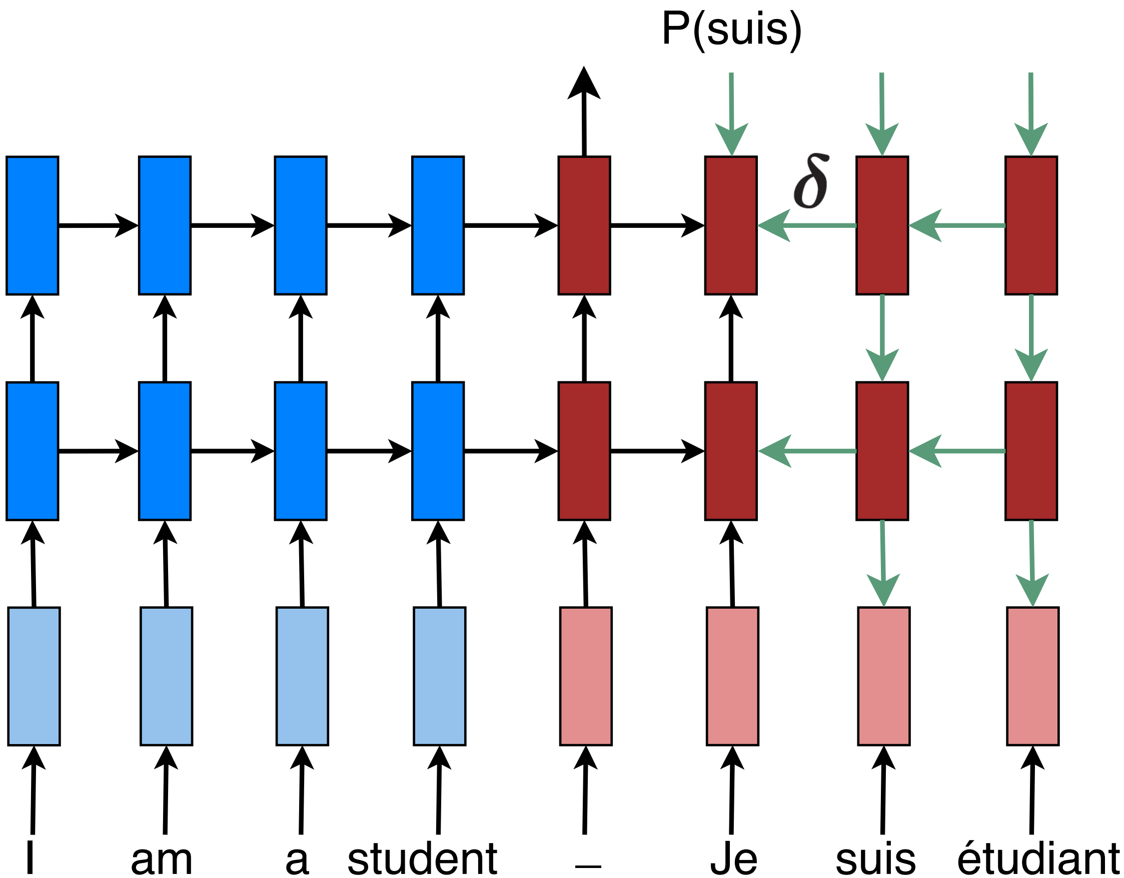
Backpropagation Through Time



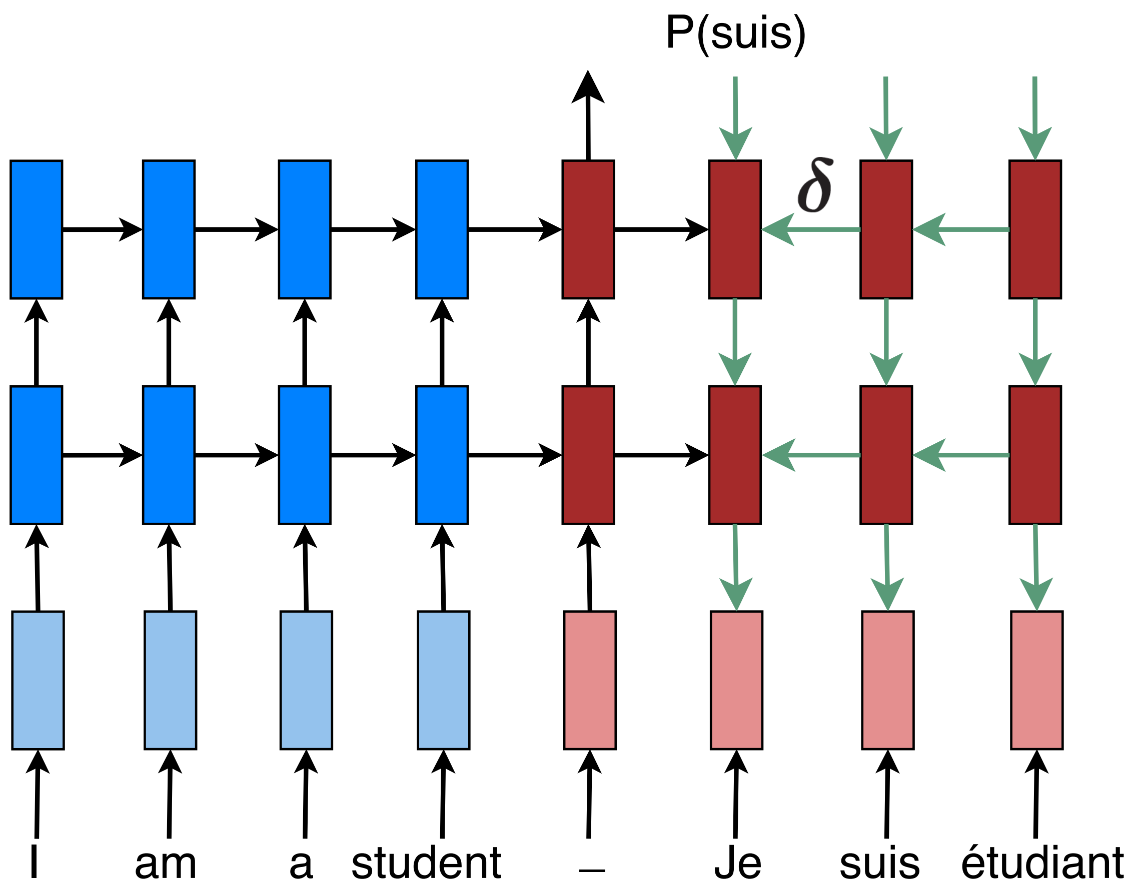
Backpropagation Through Time



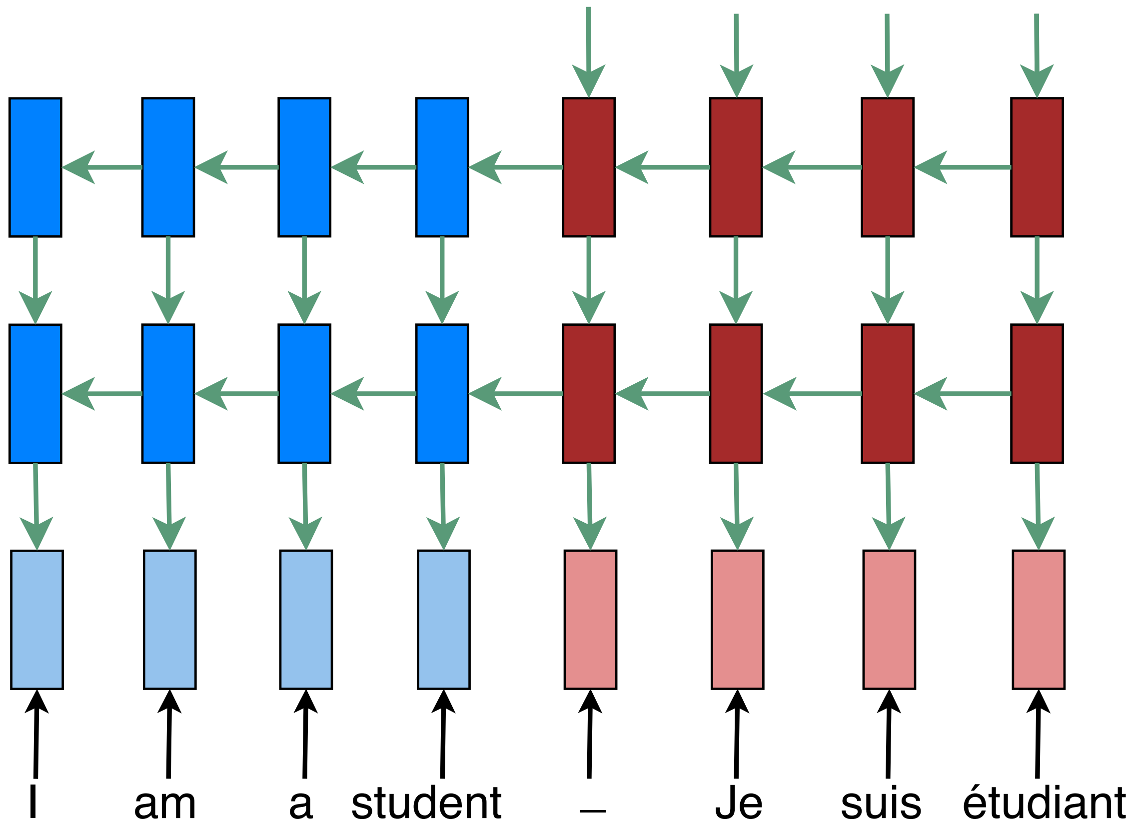
Backpropagation Through Time



Backpropagation Through Time



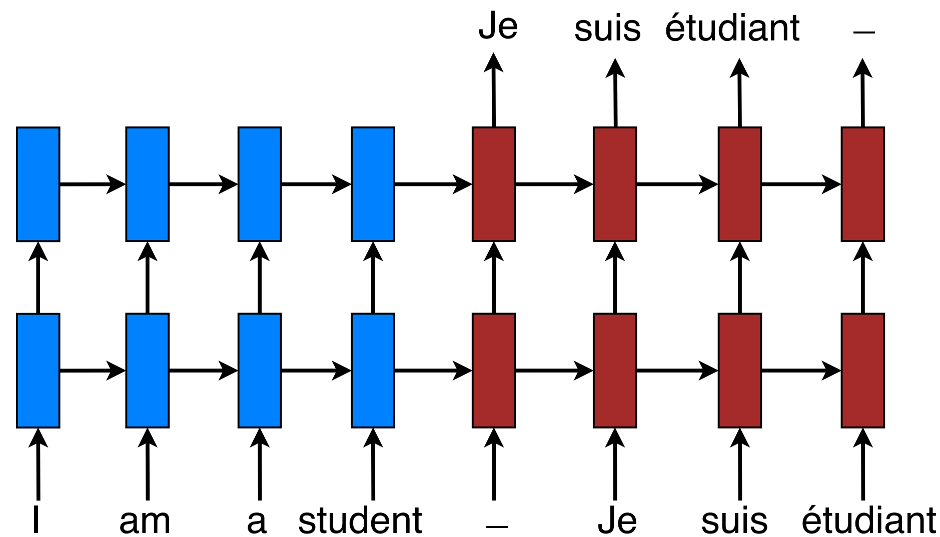
Backpropagation Through Time



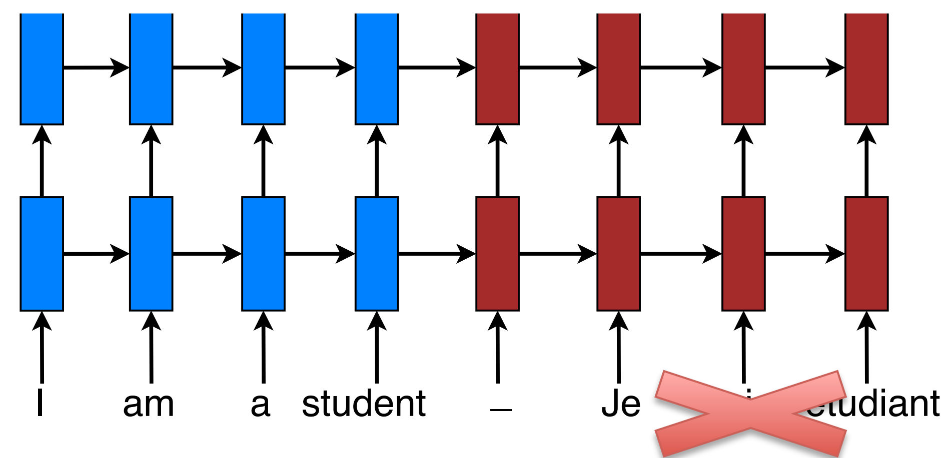
RNN gradients are accumulated.

Training vs. Testing

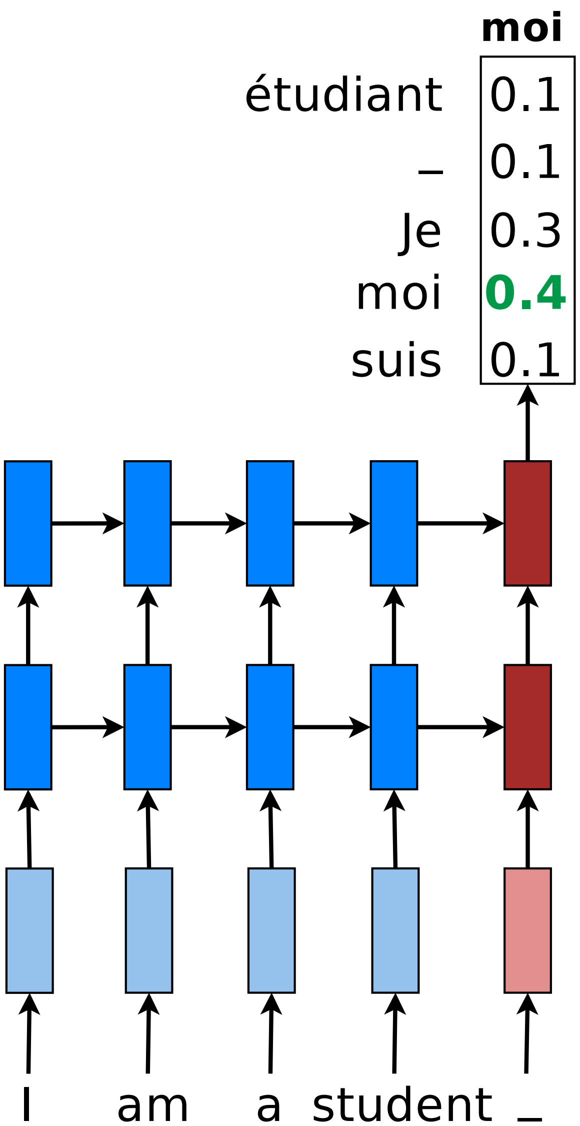
- *Training*
  - Correct translations are available.



- *Testing*
  - Only source sentences are given.

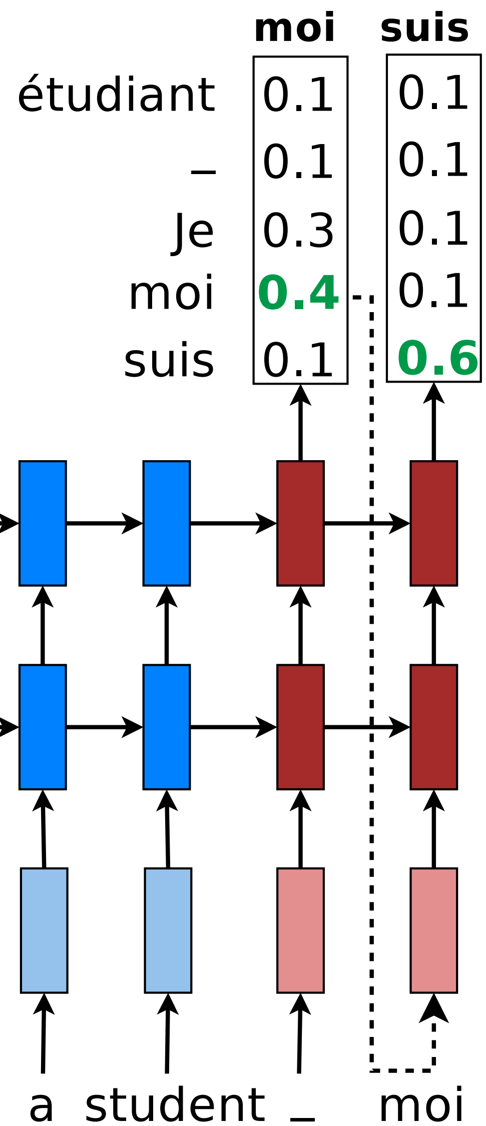


Testing



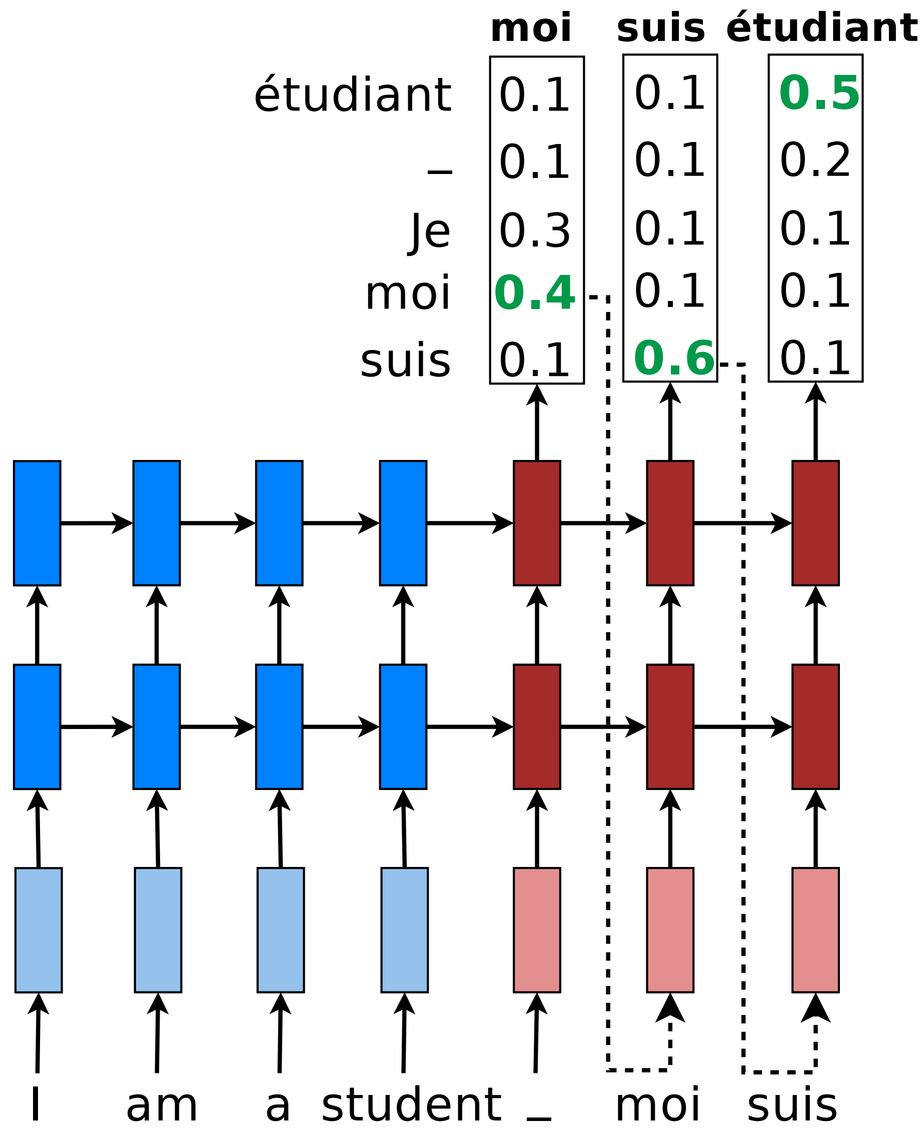
- Feed the most likely word

Testing



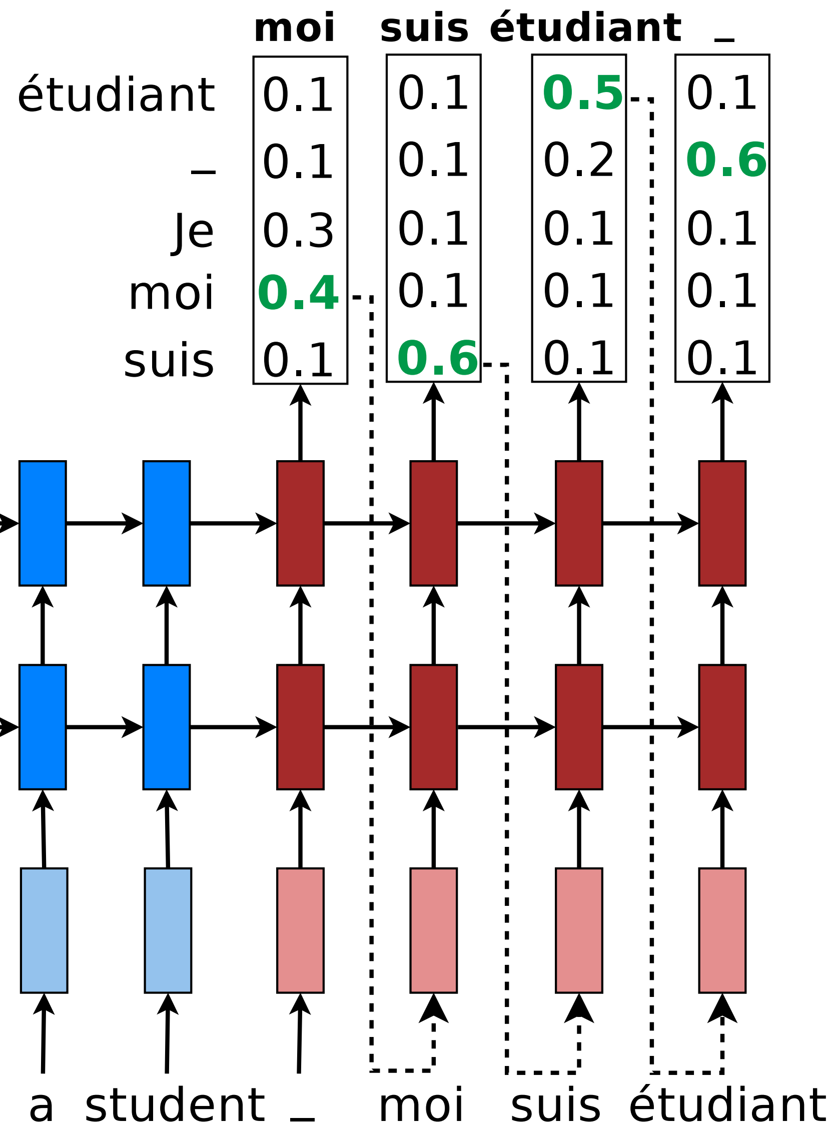
- Feed the most likely word

Testing



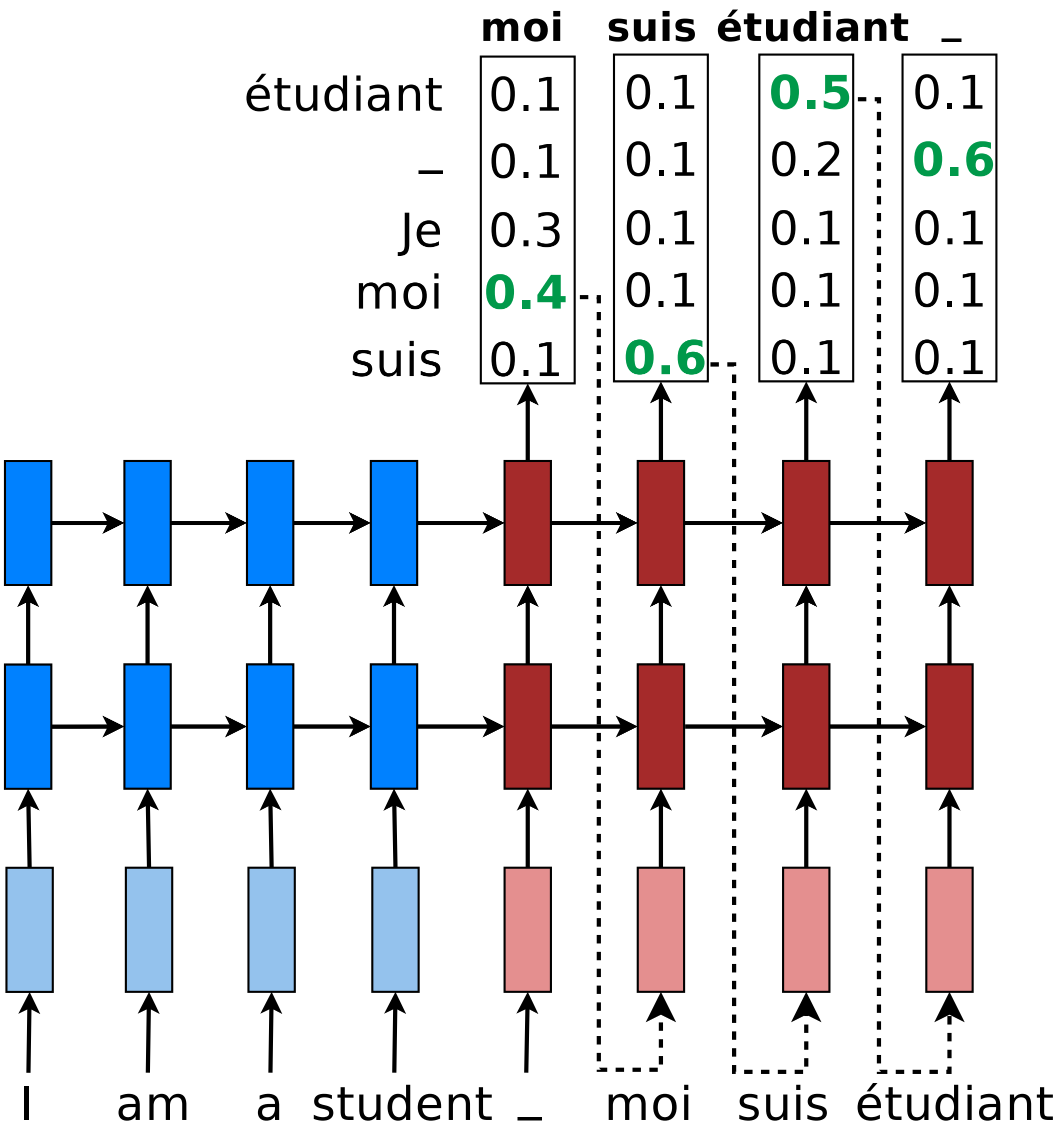
- Feed the most likely word

Testing



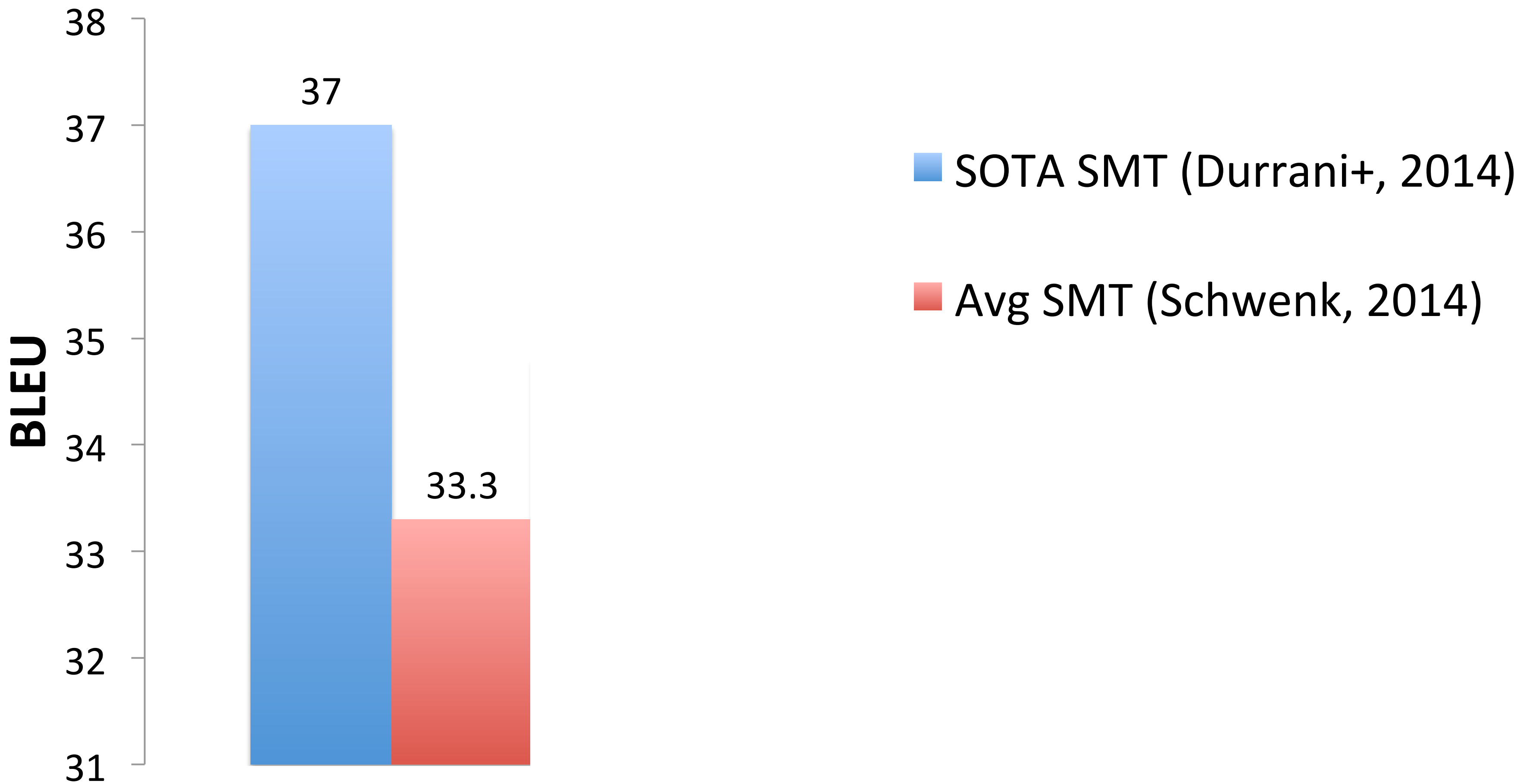
- Feed the most likely word

Testing

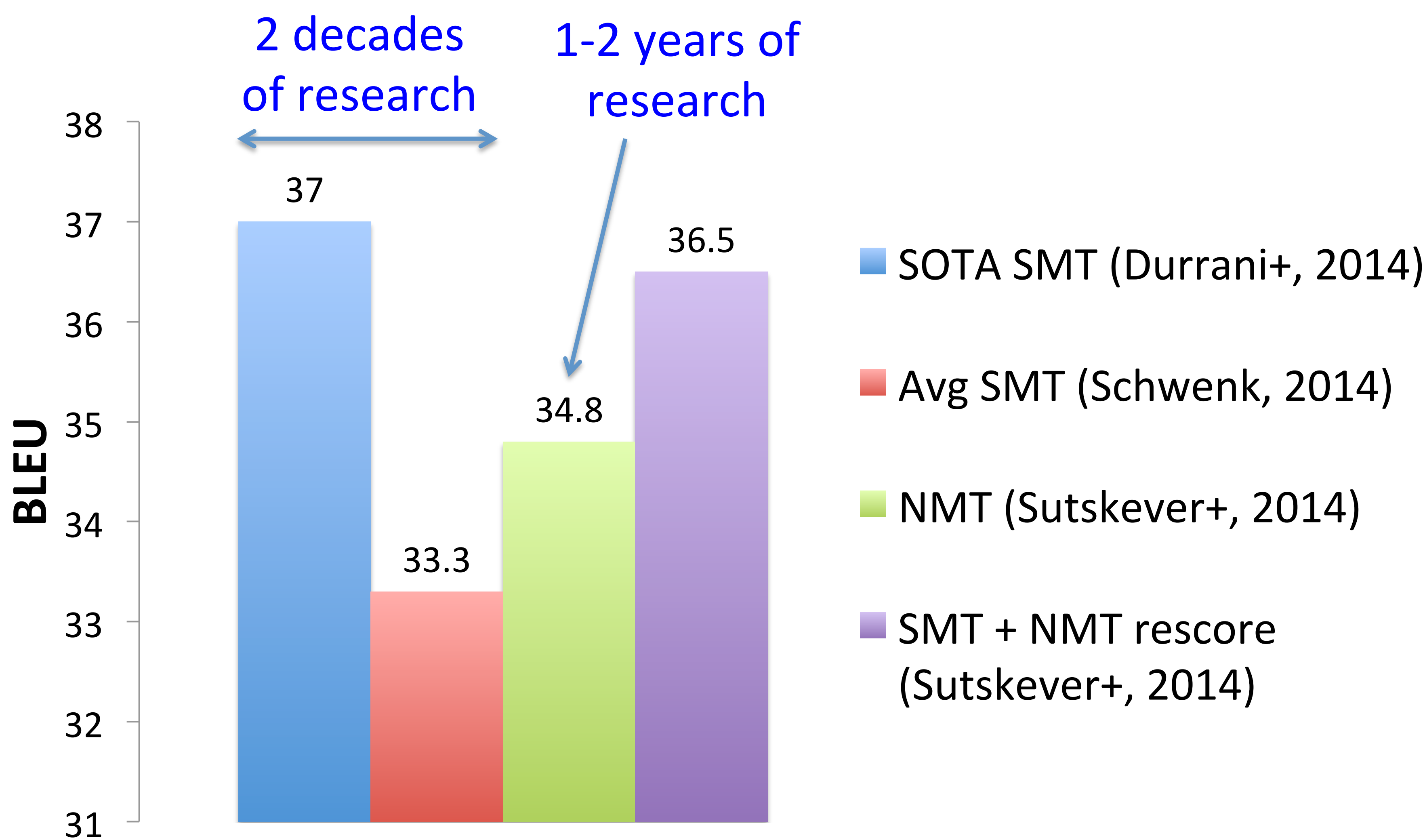


Simple beam-search decoders!

English-French WMT'14 results



English-French WMT'14 results



Encoder-decoder Variants

	Encoder	Decoder
(Sutskever et al., 2014) My NMT models	Deep LSTM	Deep LSTM
(Cho et al., 2014) (Bahdanau et al., 2015) (Jean et al., 2015)	(Bidirectional) GRU	GRU
(Kalchbrenner & Blunsom, 2013)	CNN	(Inverse CNN) RNN

Next, advanced NMT!

Break time: when MT fails …

Sale of chicken murder

Go back toward your behind

Deep fried baby

Meat muscle stupid bean sprouts

Limitations

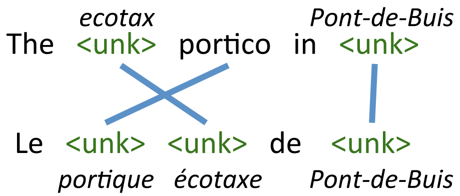
- #1: the ***vocabulary size*** problem
  - *Goal*: extend the vocabulary coverage.

- #2: the ***sentence length*** problem
  - *Goal*: translate long sentences better.

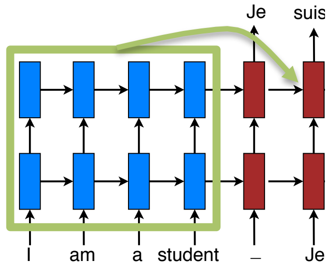
- #3: the ***language complexity*** problem
  - *Goal*: handle more language variations.

Advancing NMT

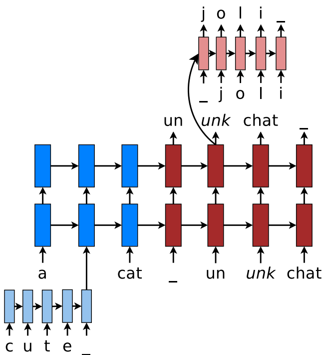
- #1: the ***vocabulary size*** problem
  - *Sol*: "copy" mechanism.



- #2: the ***sentence length*** problem
  - *Sol*: attention mechanism.



- #3: the ***language complexity*** problem
  - *Sol*: character-level translation.

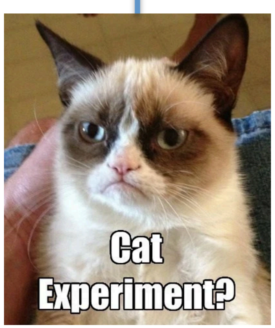


CV vs. NLP

1K categories

cat

Computer Vision



1M categories

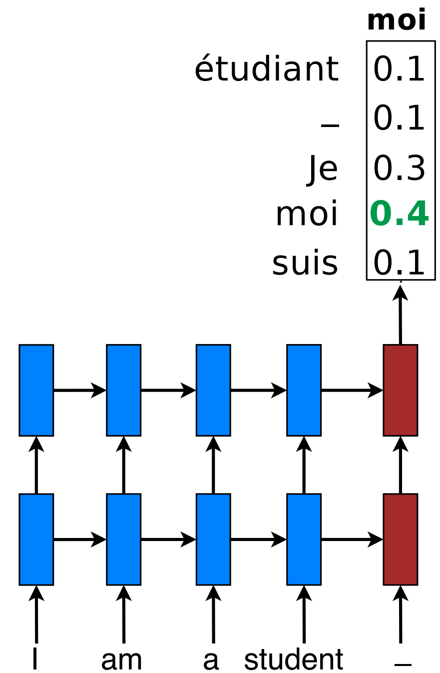
mat

NLP

The cat sat on a

#1 The Vocabulary Size Problem

- Word generation problem
  - Vocabs are modest: 50K.
  - Simple softmax: GPU friendliness.



The ecotax portico in Pont-de-Buis

Le portique écotaxe de Pont-de-Buis

The <unk> portico in <unk>

Le <unk> <unk> de <unk>

- Propose "copy" mechanisms for *<unk>*.

- Simple & effective
  - Treat any NMT as a black box.
  - Annotate training data.
  - Post-process translations.

SOTA for English-French translation.

Thang Luong, Ilya Sutskever*, Quoc Le*, Oriol Vinyals, and Wojciech Zaremba.* ***Addressing the Rare Word Problem in Neural Machine Translation****. ACL 2015.*

Our approach – *training annotation*

- Learn alignments.

The ecotax portico in Pont-de-Buis

Le portique écotaxe de Pont-de-Buis

- Add relative positions.

The <unk> portico in <unk>

Le unk_1 unk_{-1} de unk_0

“Attention” for rare words

Our approach – *post-process*

		ecotax			*Pont-de-Buis*	
Test sentence	The	<unk>	portico	in	<unk>	
Translation	Le	portique	unk_{-1}	de	unk_{0}	

Our approach – *post-process*

		ecotax			*Pont-de-Buis*
Test sentence	The	<unk>	portico	in	<unk>
Translation	Le portique	unk$_{-1}$		de	unk$_0$

Dictionary translation

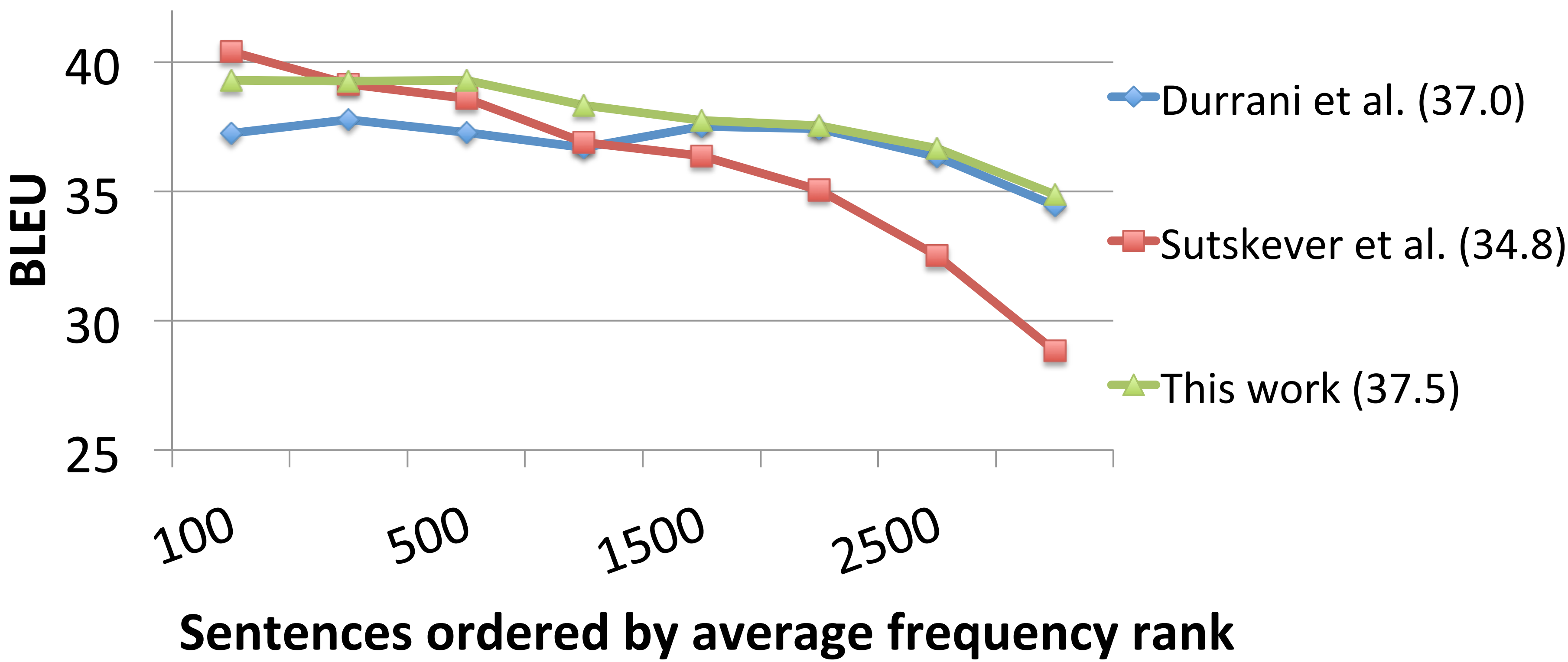
Post-edit Translation	Le portique	écotaxe	de	Pont-de-Buis

Our approach – *post-process*

		ecotax			*Pont-de-Buis*
Test sentence	The	<unk>	portico	in	<unk>
Translation	Le	portique	unk_{-1}	de	$\boxed{unk_0}$
					Identity copy
Post-edit Translation	Le	portique	écotaxe	de	Pont-de-Buis

Orthogonal to large-vocab techniques

Effects of Translating Rare Words



First SOTA NMT system!

Sample translations

source	An additional **2600** operations including **orthopedic** and $\boxed{\text{cataract}}$ surgery will help clear a backlog .
human	**2600** opérations supplémentaires , notamment dans le domaine de la chirurgie **orthopédique** et de la $\boxed{\textbf{cataracte}}$, aideront à rattraper le retard .
trans	En outre , **unk_1** opérations supplémentaires , dont la chirurgie **unk_5** et la $\boxed{\textbf{unk}_6}$, permettront de résorber l' arriéré .
trans +unk	En outre , **2600** opérations supplémentaires , dont la chirurgie **orthopédiques** et la $\boxed{\text{cataracte}}$, permettront de résorber l' arriéré .

- Predict well long-distance alignments.
  - Correct: **cataract** vs. ***cataracte***.

Sample translations

source	This **trader** , Richard **Usher** , left RBS in **2010** and is understand to have be given leave from his current position as European head of forex spot trading at **JPMorgan** .
human	Ce **trader** , Richard **Usher** , a quitté **RBS** en 2010 et aurait été mis suspendu de son poste de responsable européen du trading au comptant pour les devises chez **JPMorgan** .
trans	Ce **unk_0** , Richard **unk_0** , a quitté **unk_1** en 2010 et a compris qu' il est autorisé à quitter son poste actuel en tant que leader européen du marché des points de vente au **unk_5** .
trans +unk	Ce **négociateur** , Richard **Usher** , a quitté **RBS** en 2010 et a compris qu' il est autorisé à quitter son poste actuel en tant que leader européen du marché des points de vente au **JPMorgan** .

- Translate well long sentences.
  - Correct: **JPMorgan** vs. ***JPMorgan.***

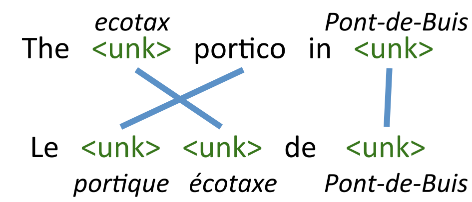
Sample translations

source	But concerns have grown after Mr Mazanga was quoted as saying Renamo was abandoning the 1992 peace accord .
human	Mais l' inquiétude a grandi après que M. **Mazanga** a déclaré que la **Renamo** ***abandonnait*** l' accord de paix de 1992 .
trans	Mais les inquiétudes se sont accrues après que M. **$unkpos_3$** a déclaré que la **unk_3** **unk_3** l' accord de paix de 1992 .
trans +unk	Mais les inquiétudes se sont accrues après que M. **Mazanga** a déclaré que la **Renamo** **était** l' accord de paix de 1992 .

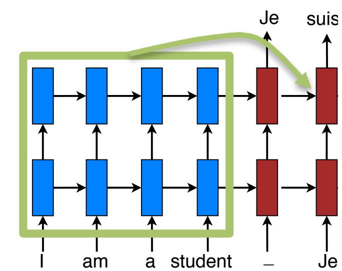
- Incorrect alignment prediction: **was** – **était** vs. ***abandonnait.***

Advancing NMT

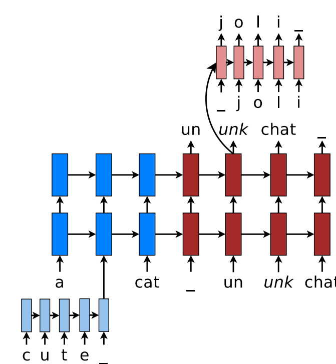
- #1: the *vocabulary size* problem
  - *Sol*: “copy” mechanism.



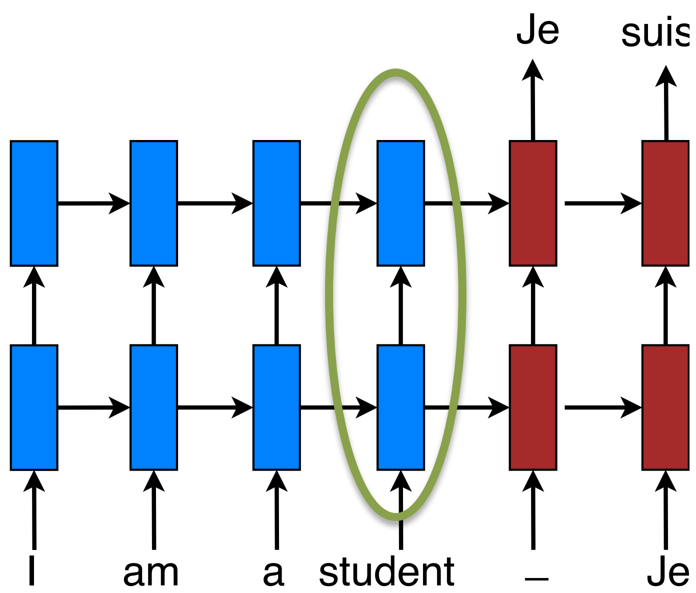
- #2: the ***sentence length*** problem
  - *Sol*: attention mechanism.



- #3: the ***language complexity*** problem
  - *Sol*: character-level translation.



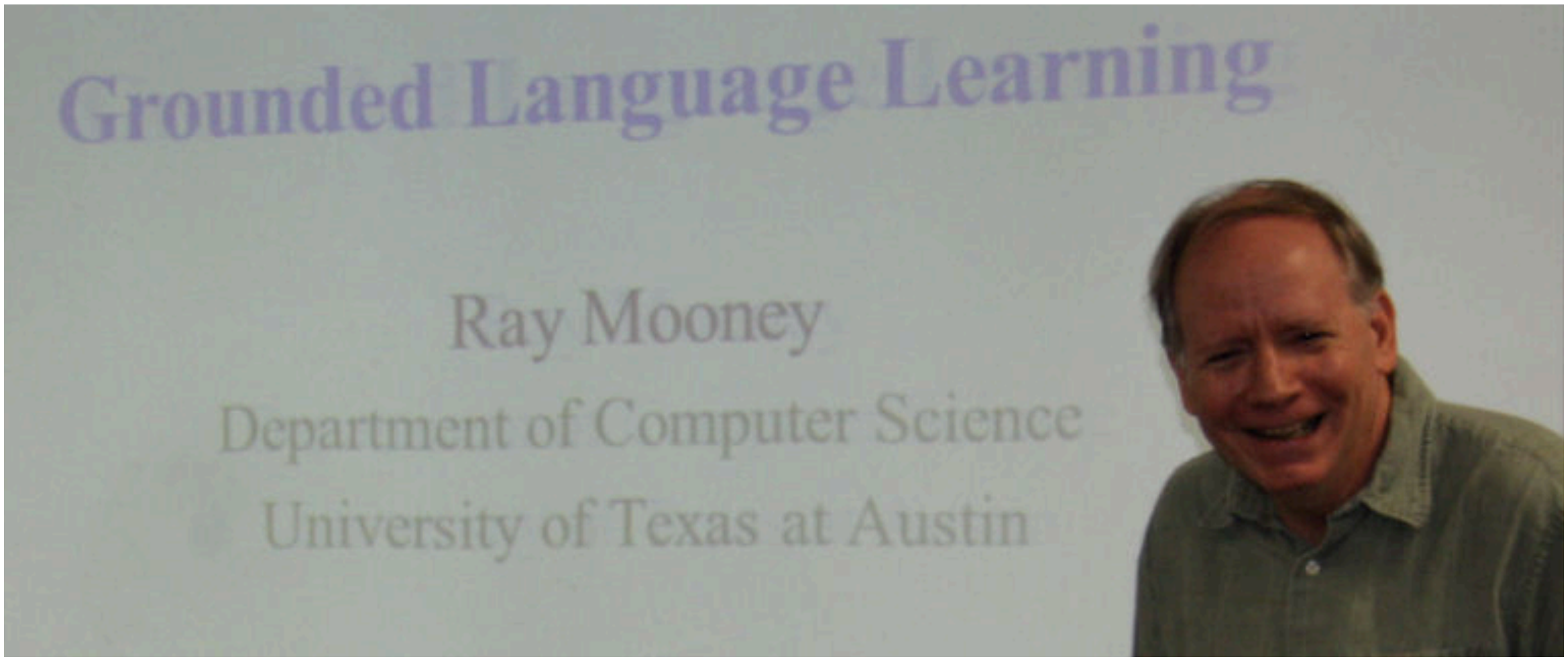
#2 The Sentence Length Problem



- Translation quality degrades with long sentences.

Problem: sentence meaning is represented by a fixed-dimensional vector.

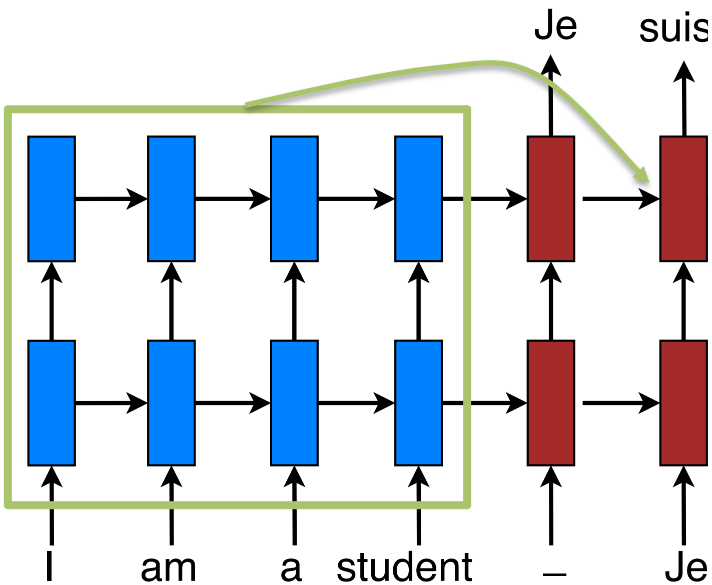
You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!



(Adapted from KyungHuyn Cho' talk)

Attention Mechanism

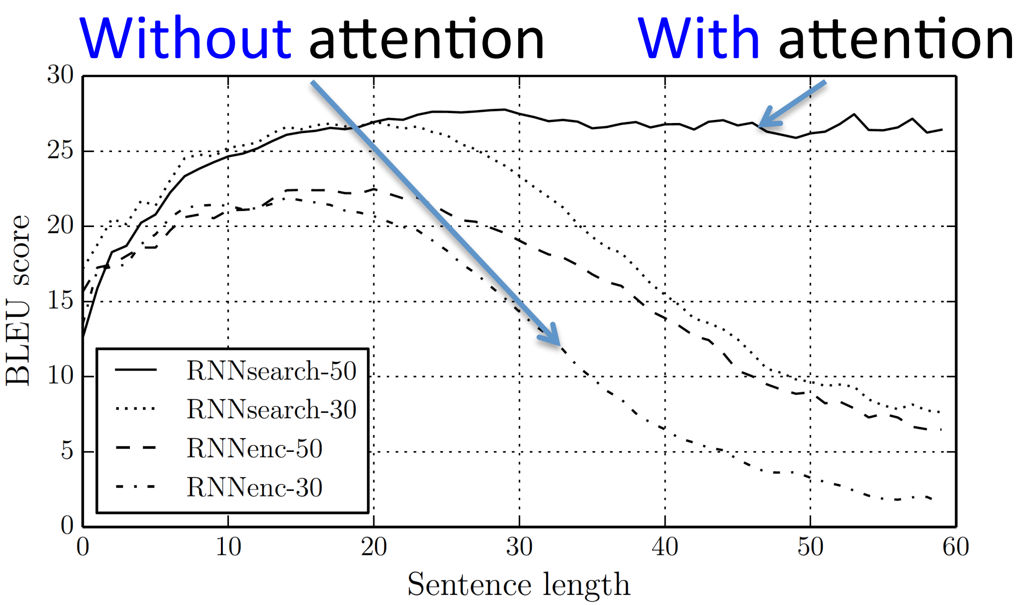
Pool of source states

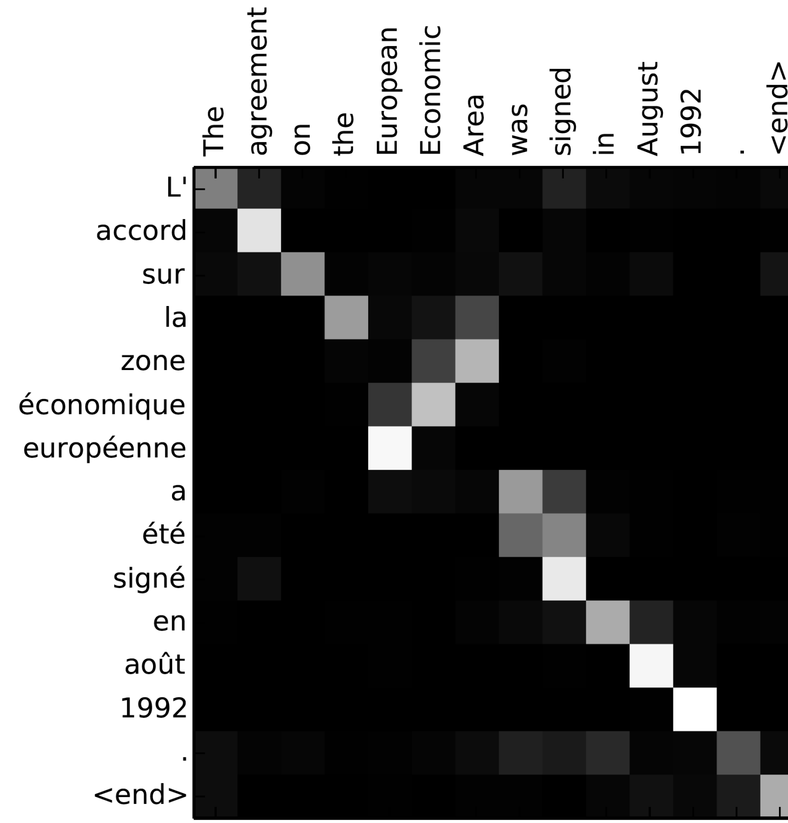


- Solution: random access memory
  - Retrieve as needed.

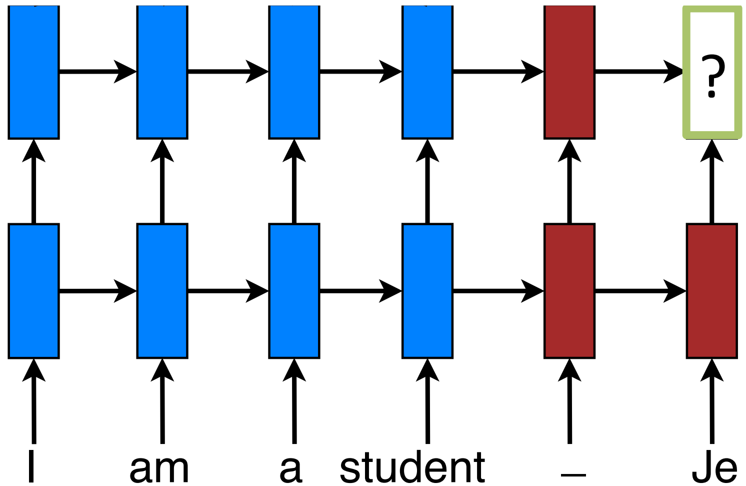
Université de Montréal





Dzmitry Bahdanau, KyungHuyn Cho, and Yoshua Bengio. ***Neural Machine Translation by Jointly Learning to Translate and Align****. ICLR 2015.*

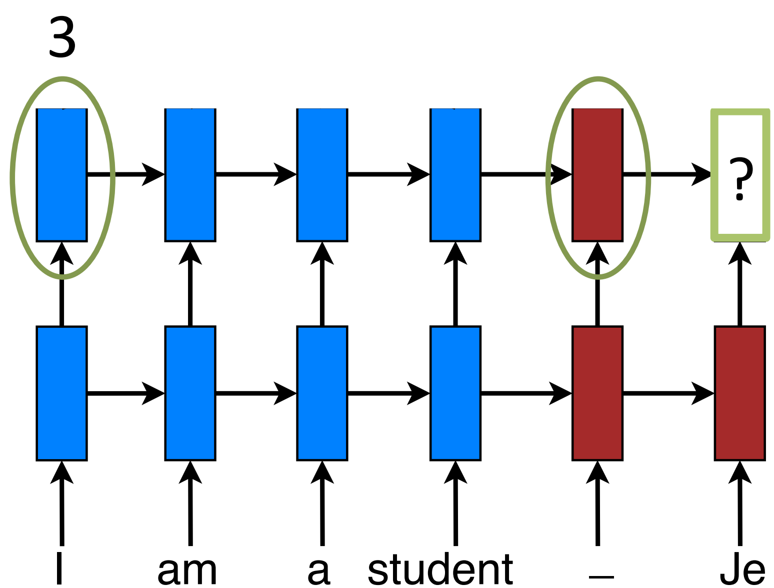
Attention Mechanism



A simplified version of (Bahdanau et al., 2015)

Attention Mechanism – *Scoring*

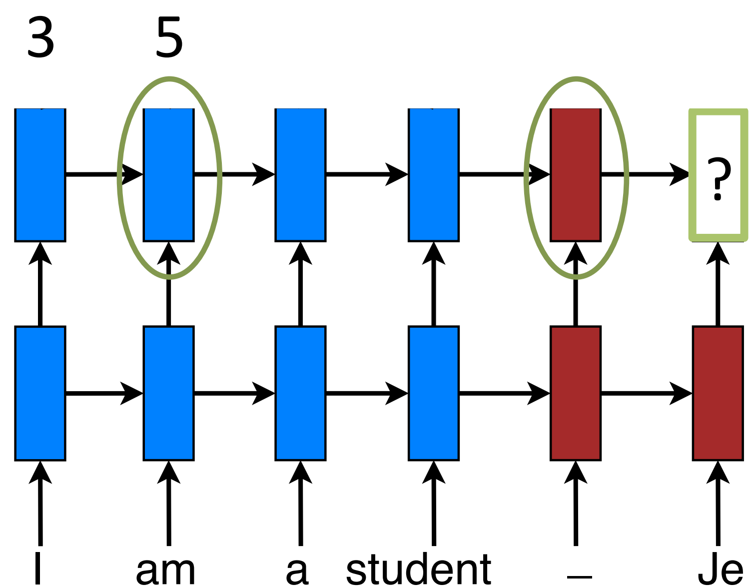
$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

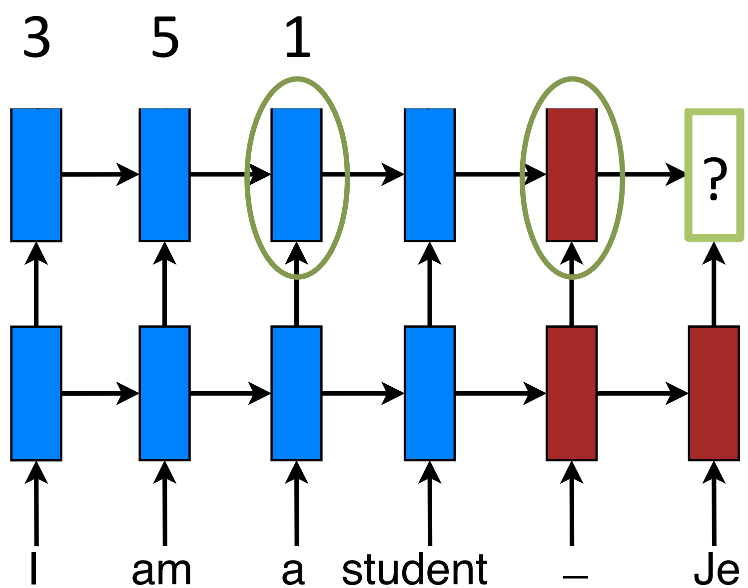
$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

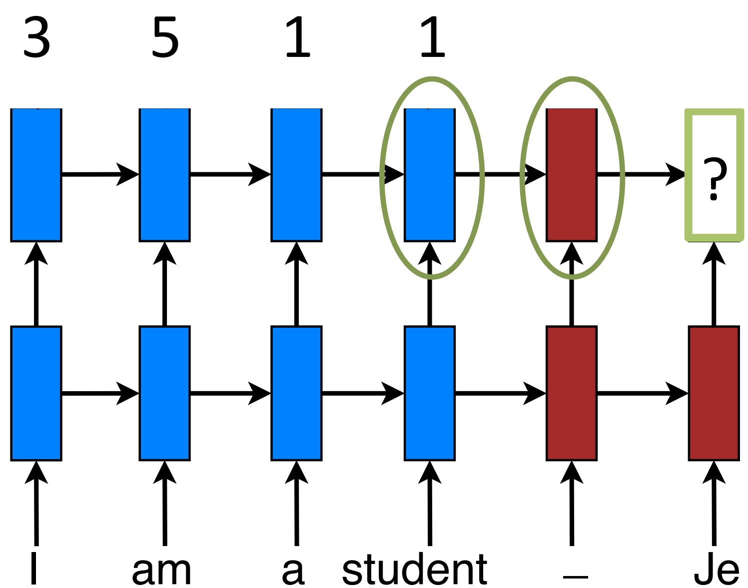
$$\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$$



- Compare target and source hidden states.

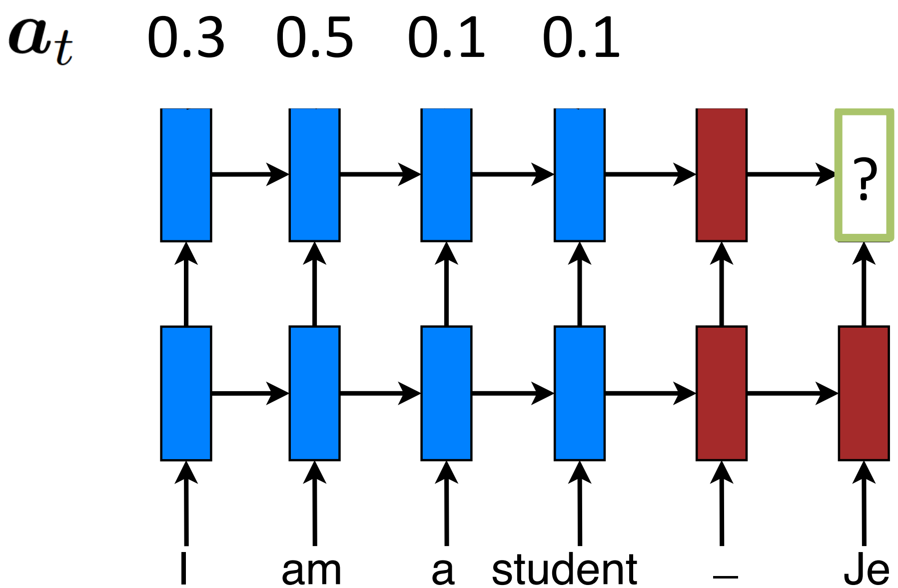
Attention Mechanism – *Scoring*

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



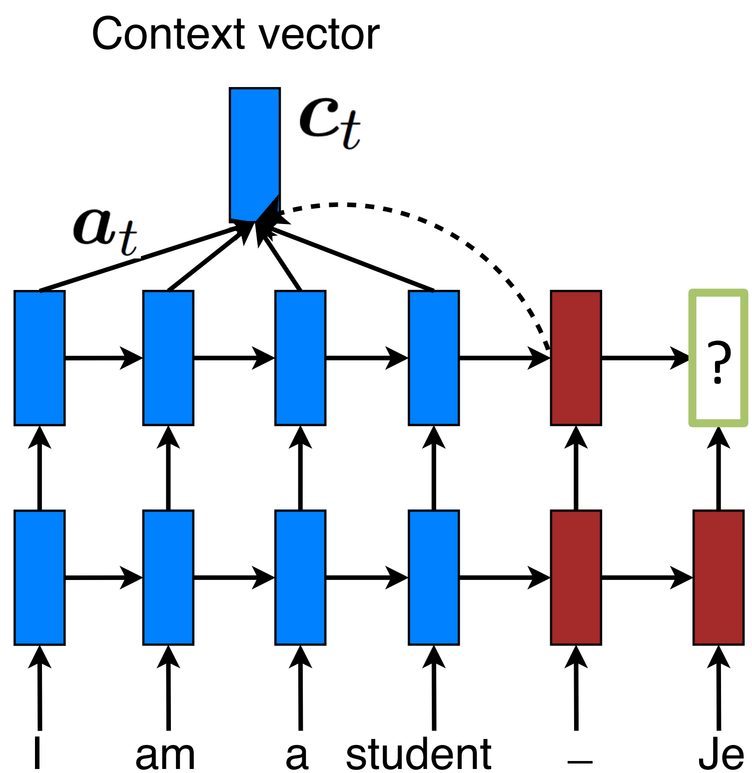
- Compare target and source hidden states.

Attention Mechanism – *Normalization*



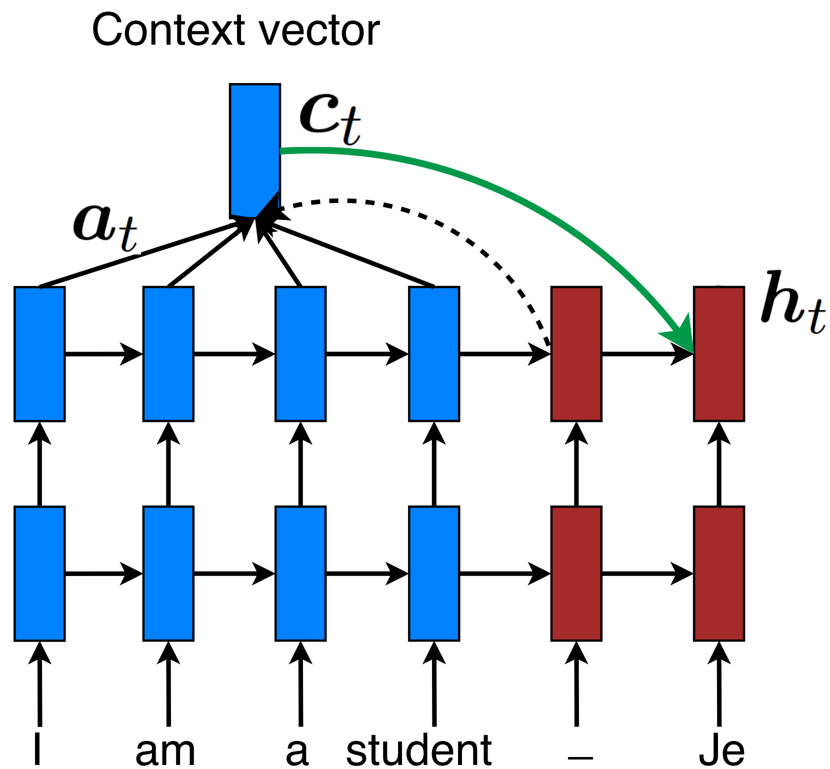
- Convert into alignment weights.

Attention Mechanism – *Context vector*



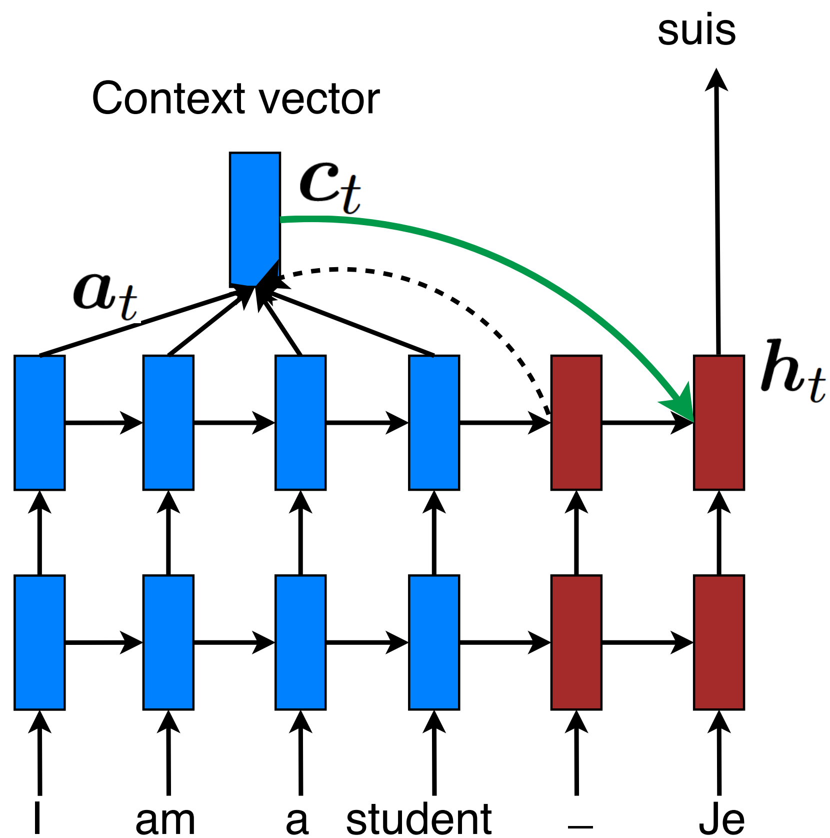
- Build context vector: weighted average.

Attention Mechanism – *Hidden state*



- Compute the next hidden state.

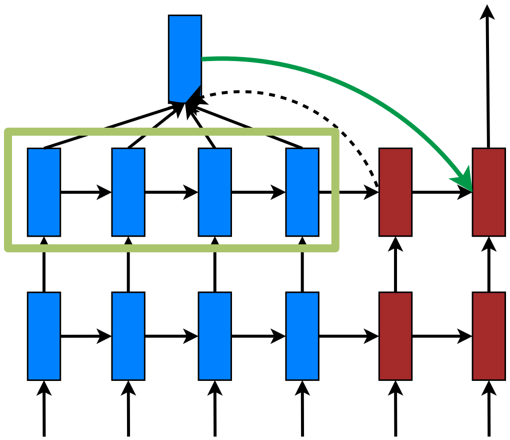
Attention Mechanism – *Predict*



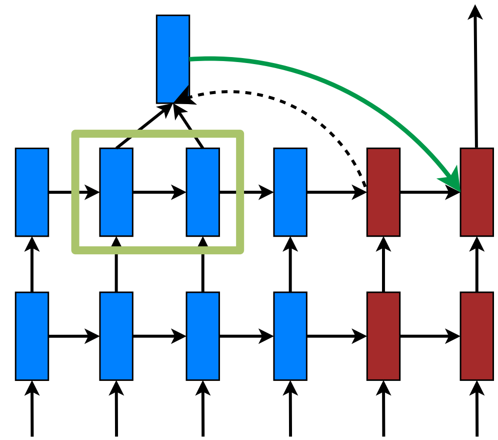
- Predict the next word.



- Examine various attention mechanisms:



Global: ***all*** source states.

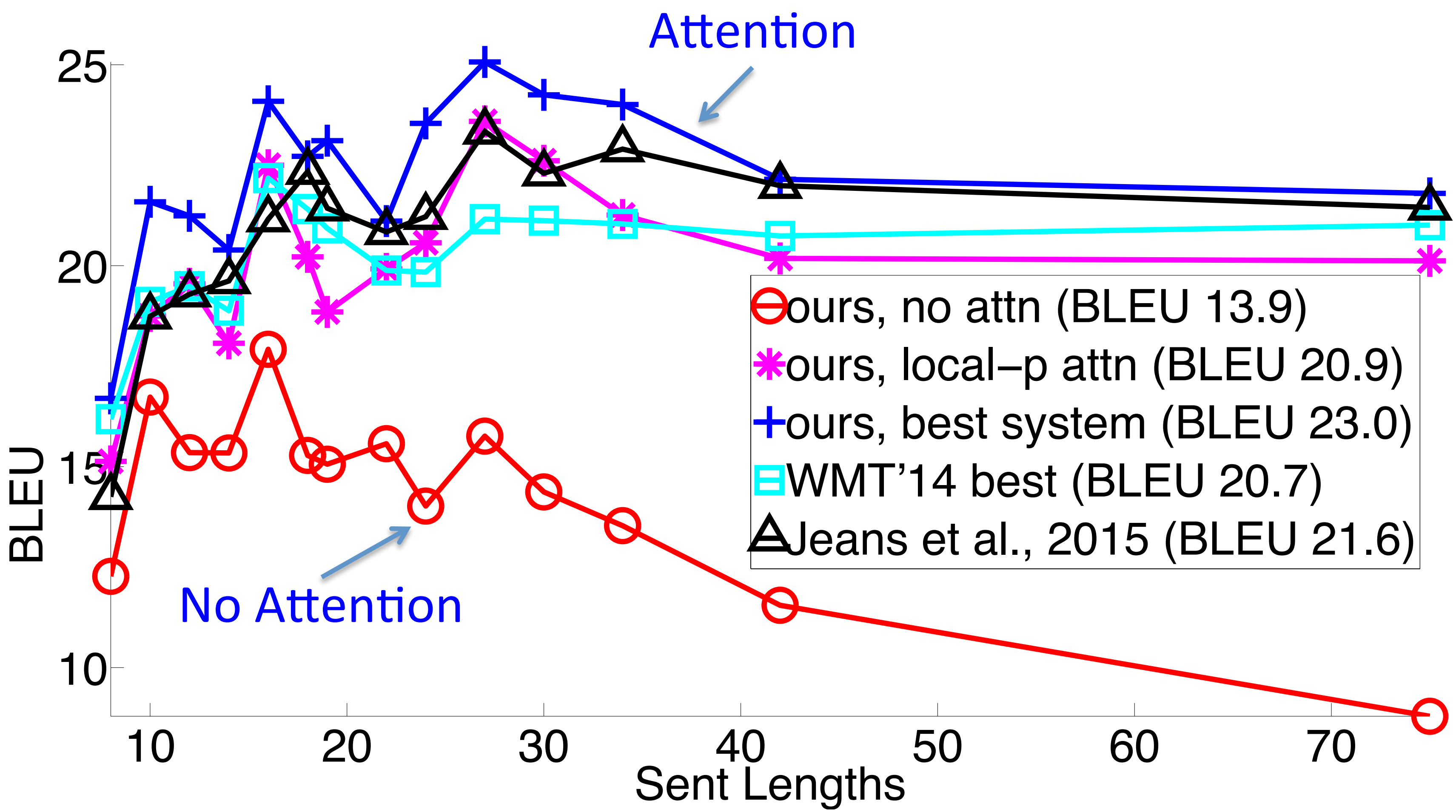


Local: ***subset*** of source states.

SOTA for English-German translation.

<u>Thang Luong</u>, Hieu Pham, and Chris Manning. ***Effective Approaches to Attention-based Neural Machine Translation****. EMNLP 2015.*

Translate Long Sentences



Sample English-German translations

source	Orlando Bloom and *Miranda Kerr* still love each other
human	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.

Sample English-German translations

source	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 .
human	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> .

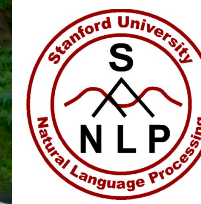
- Translate a doubly-negated phrase correctly

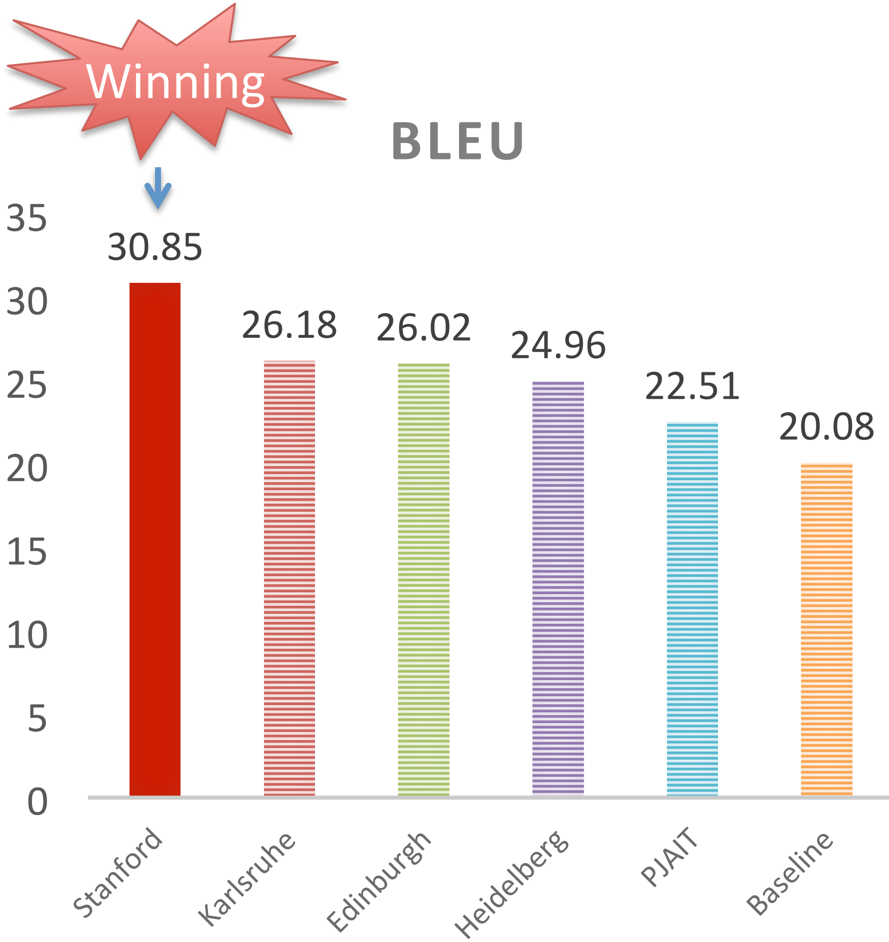
Sample English-German translations

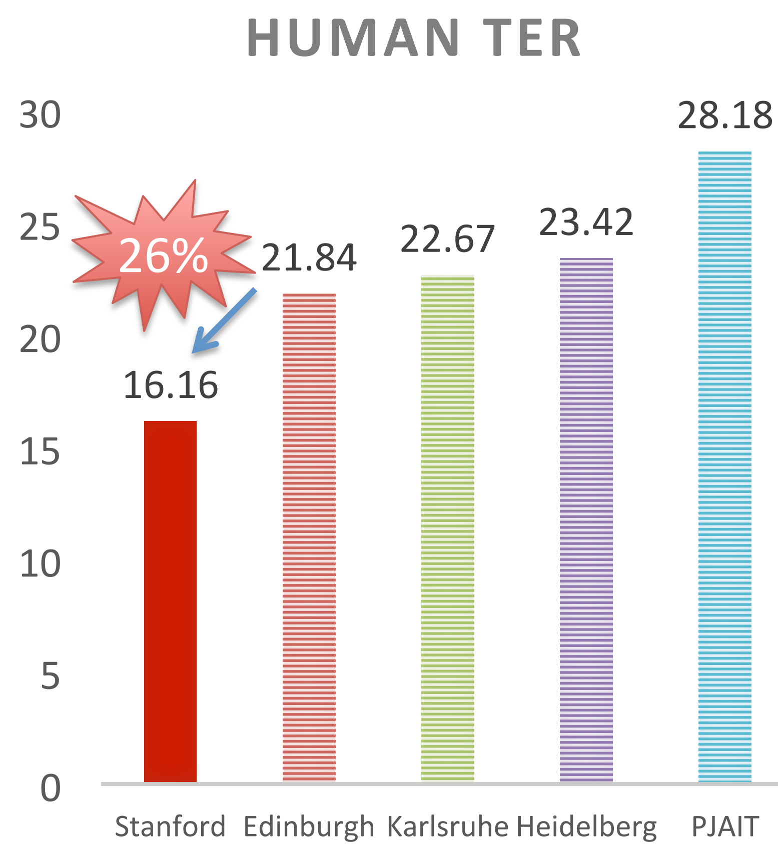
source	We ’ re pleased the FAA recognizes that an enjoyable passenger experience is ***not incompatible*** with safety and security , said Roger Dow , CEO of the U.S. Travel Association .
human	Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur 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

TED talk, English-German



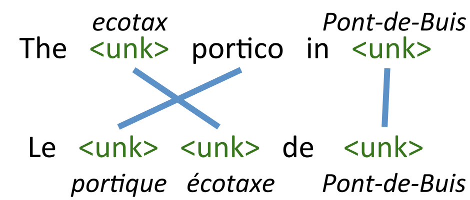




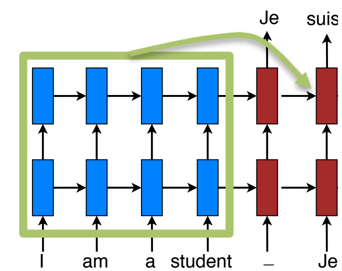
Thang Luong *and Chris Manning.* ***Stanford Neural Machine Translation Systems for Spoken Language Domain.*** *IWSLT 2015.*

Advancing NMT

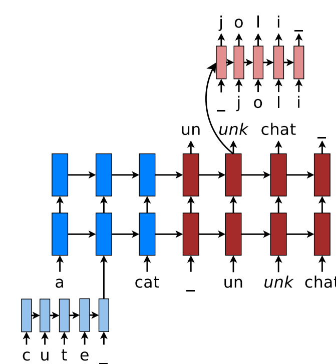
- #1: the *vocabulary size* problem
  - *Sol*: “copy” mechanism.



- #2: the *sentence length* problem
  - *Sol*: attention mechanism.



- #3: the ***language complexity*** problem
  - *Sol*: character-level translation.



#3 The rare word problem

- "Copying" mechanisms are **not sufficient**.
  - Different alphabets: *Christopher* ↦ *Kryštof*
  - Multi-word alignment: *Solar system* ↦ *Sonnensystem*

- Need to handle **large, open vocabulary**
  - *Rich morphology*: *nejneobhospodařovávatelnějšímu* *("to the worst farmable one")*
  - *Informal spelling*: *goooooood morning !!!!!*

Be able to generate at the character level.

Recent character-level NMT

- Unsatisfactory performance
  - (Wang Ling, Isabel Trancoso, Chris Dyer, Alan Black, arXiv 2015)
- Incomplete solution
  - Decoder only (Junyoung Chung, Kyunghyun Cho, Yoshua Bengio. arXiv 2016).


- A *best-of-both-worlds* architecture:
  - Translate mostly at the word level
  - Only go the character level when needed.

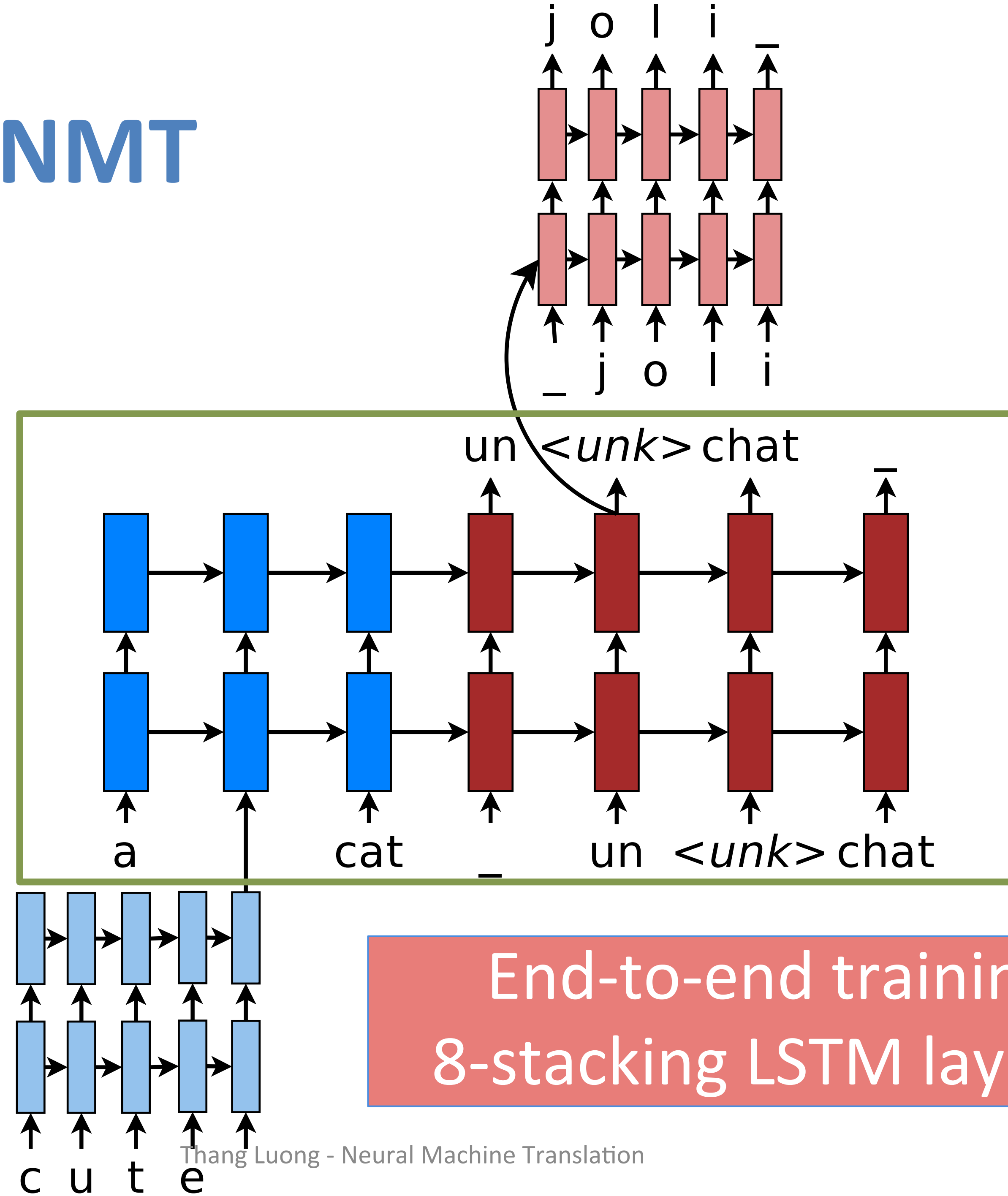
- Additional +2.1 ↦ +11.4 BLEU improvement.

SOTA for English-Czech translation.

Thang Luong and Chris Manning. ***Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models****. In submission, ACL 2016.*

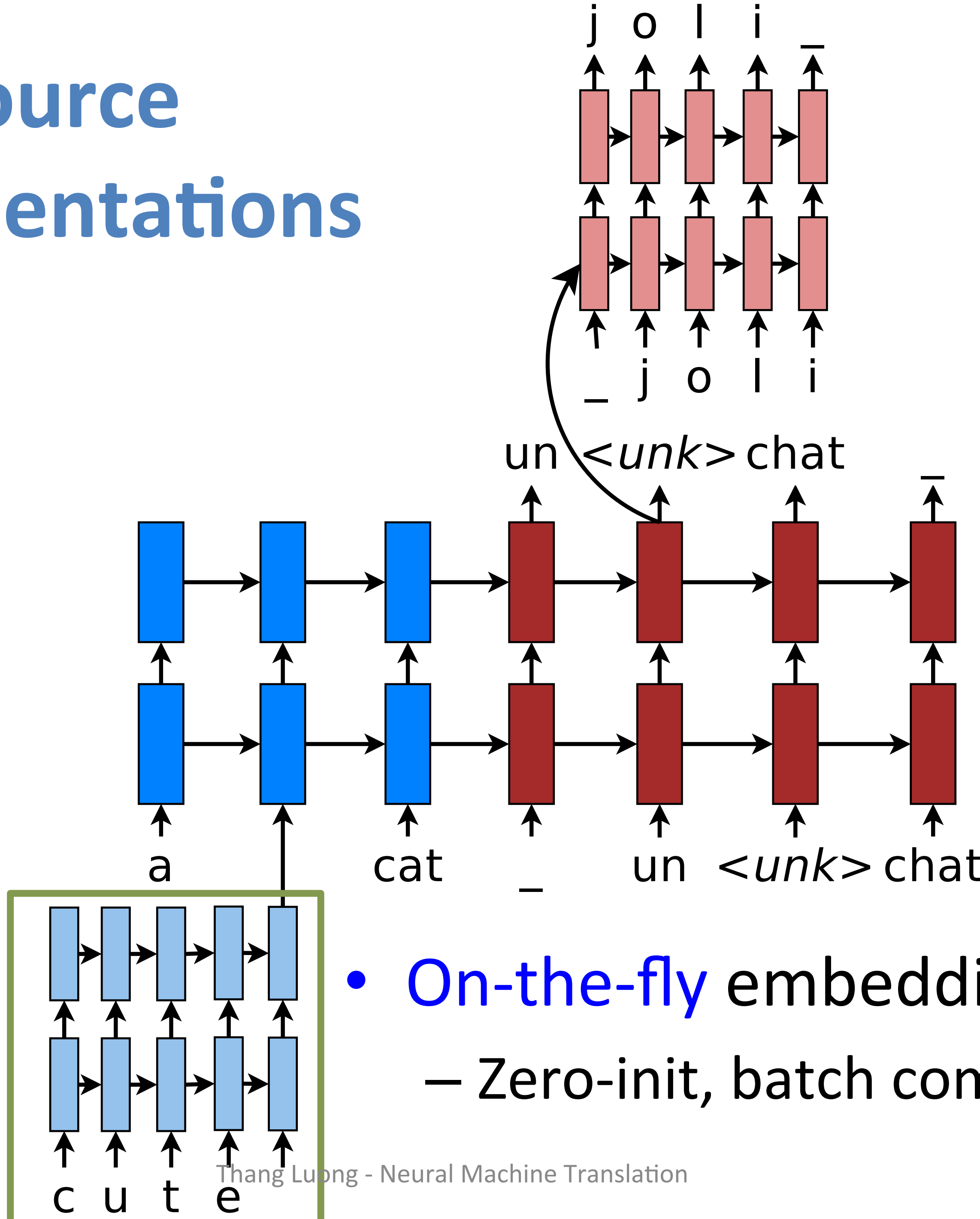
Hybrid NMT

Word-level
(4 layers)



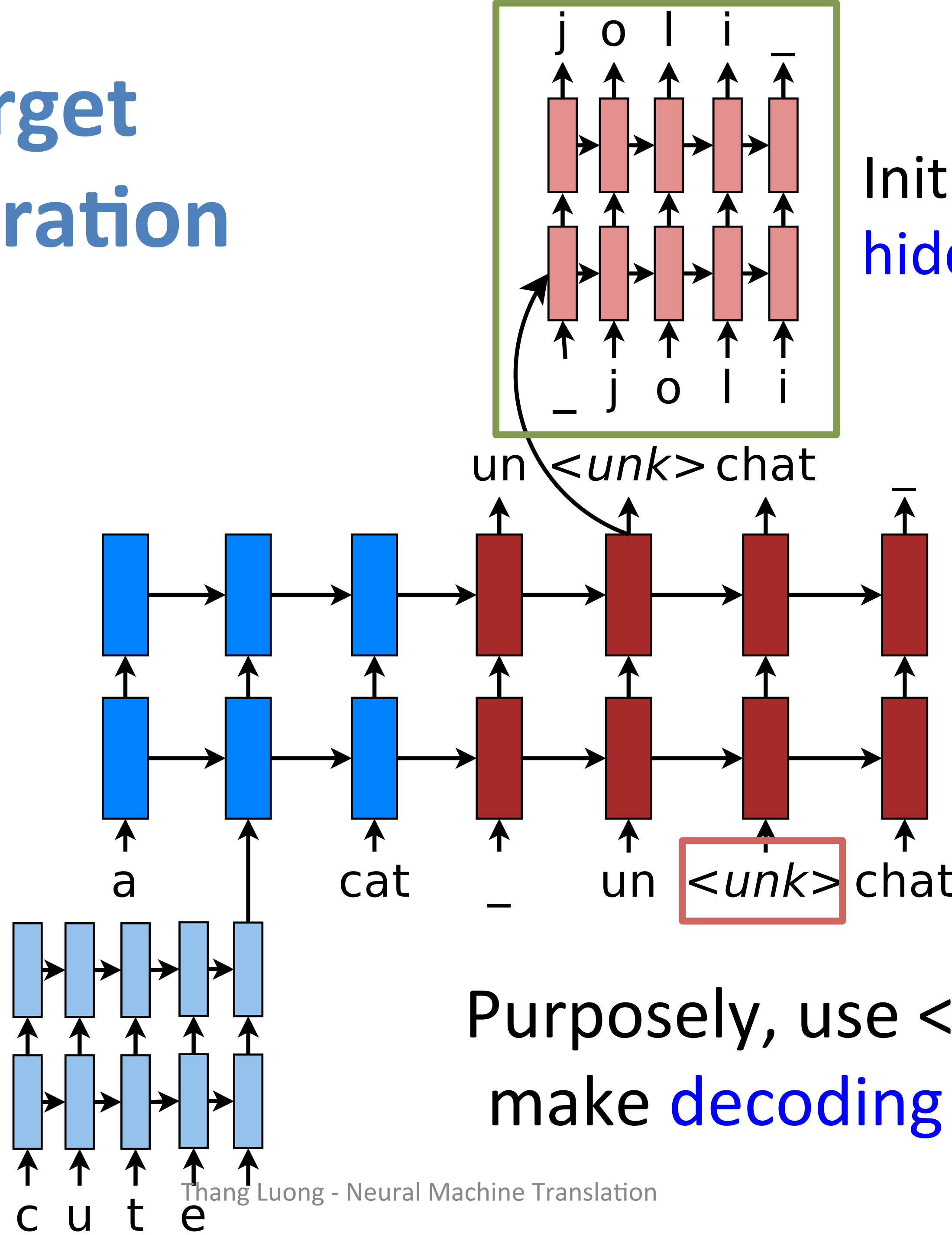
End-to-end training
8-stacking LSTM layers.

Source Representations



- On-the-fly embeddings.
  - Zero-init, batch computation.

Target Generation



Init with word hidden states.

Purposely, use <unk> to make decoding easier.

English-Czech WMT'15 Results

Systems	BLEU
Winning entry (Bojar & Tamchyna, 2015)	***18.8***
Existing ***word-level*** *NMT* (Jean et al., 2015)	
Single model	15.7
Ensemble 4 models	18.3

30x data
3 systems

Large vocab
+ unk replace

English-Czech WMT'15 Results

Systems	BLEU
Winning entry (Bojar & Tamchyna, 2015)	***18.8***
*Existing **word-level** NMT* (Jean et al., 2015)	
Single model	15.7
Ensemble 4 models	18.3
*Our **character-based** NMT*	
Single model (600-step backprop)	15.9

30x data
3 systems

Large vocab
+ unk replace

- Purely character-based: **slow but promising**!

English-Czech WMT’15 Results

Systems	BLEU
Winning entry (Bojar & Tamchyna, 2015)	***18.8***
*Existing **word-level** NMT* (Jean et al., 2015)	
Single model	15.7
Ensemble 4 models	18.3
*Our **character-based** NMT*	
Single model (600-step backprop)	15.9
*Our **hybrid** NMT*	
Single model	**19.6**

30x data
3 systems

Large vocab
+ unk replace

New
SOTA!

English-Czech WMT'15 Results

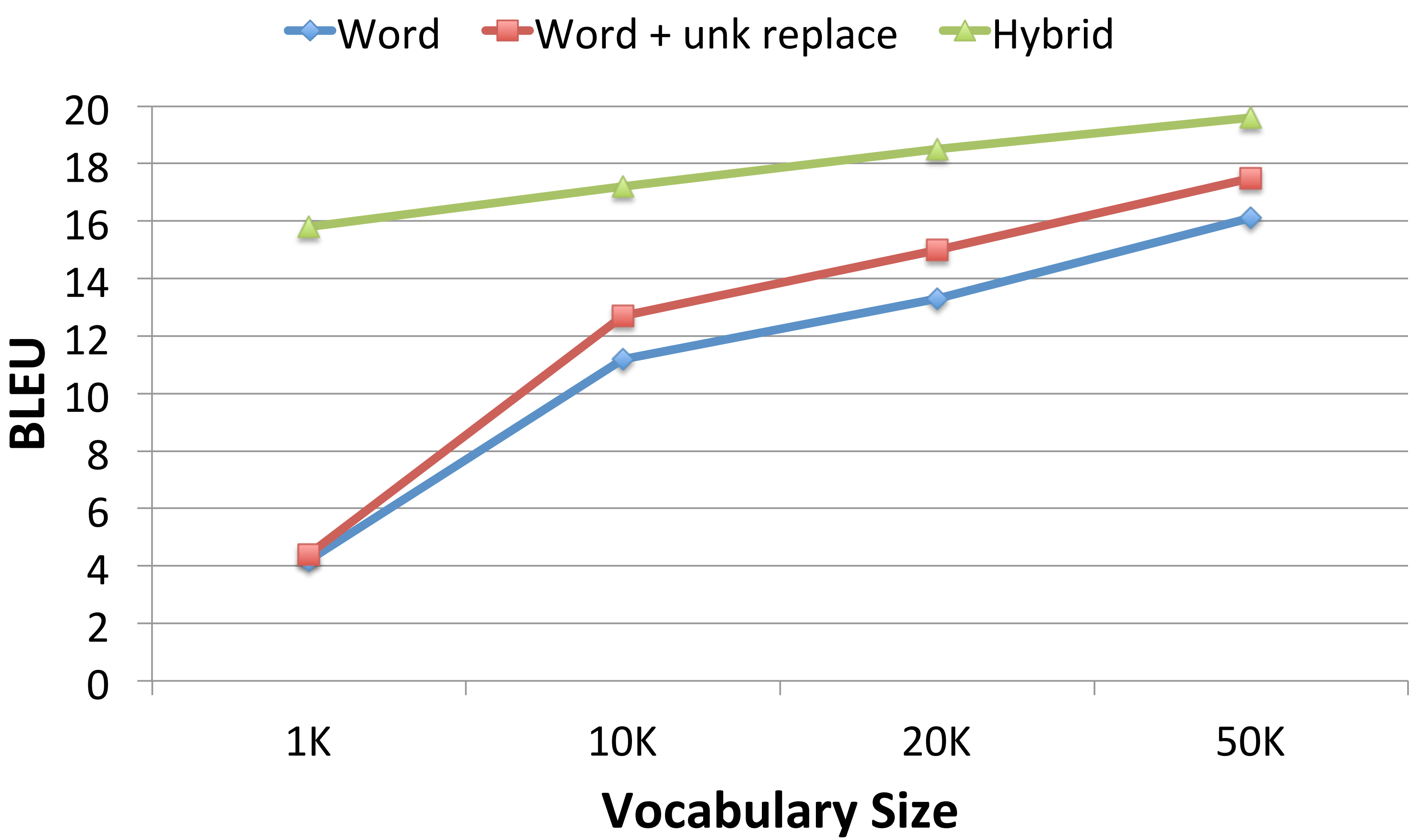
Systems	BLEU
Winning entry (Bojar & Tamchyna, 2015)	***18.8***
Existing ***word-level*** *NMT* (Jean et al., 2015)	
Single model	15.7
Ensemble 4 models	18.3
Our ***character-based*** *NMT*	
Single model (600-step backprop)	15.9
Our ***hybrid*** *NMT*	
Single model	19.6
Ensemble 4 models	**20.7**

30x data
3 systems

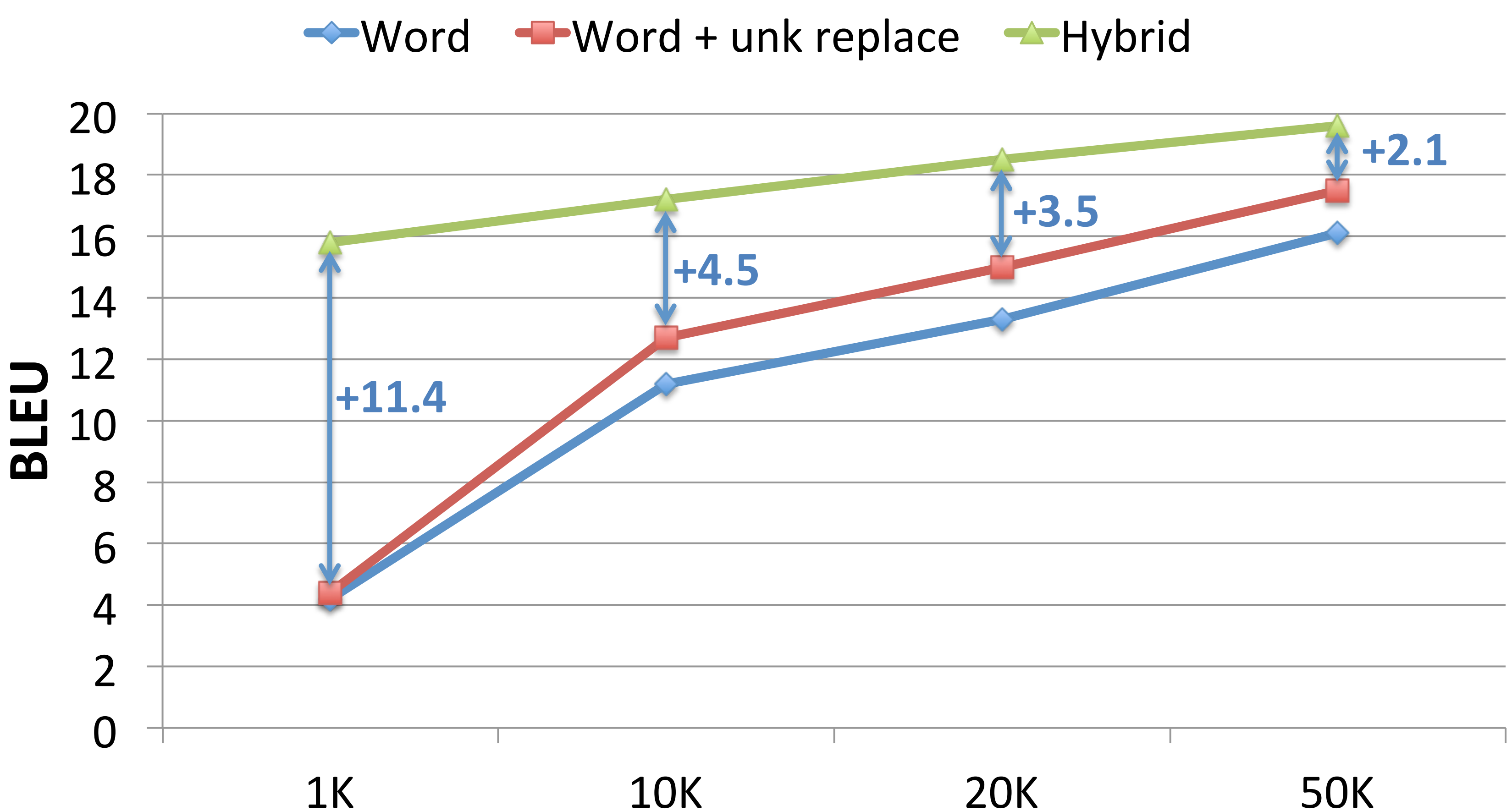
Large vocab
+ unk replace

Better
SOTA!

Effects of Vocabulary Sizes

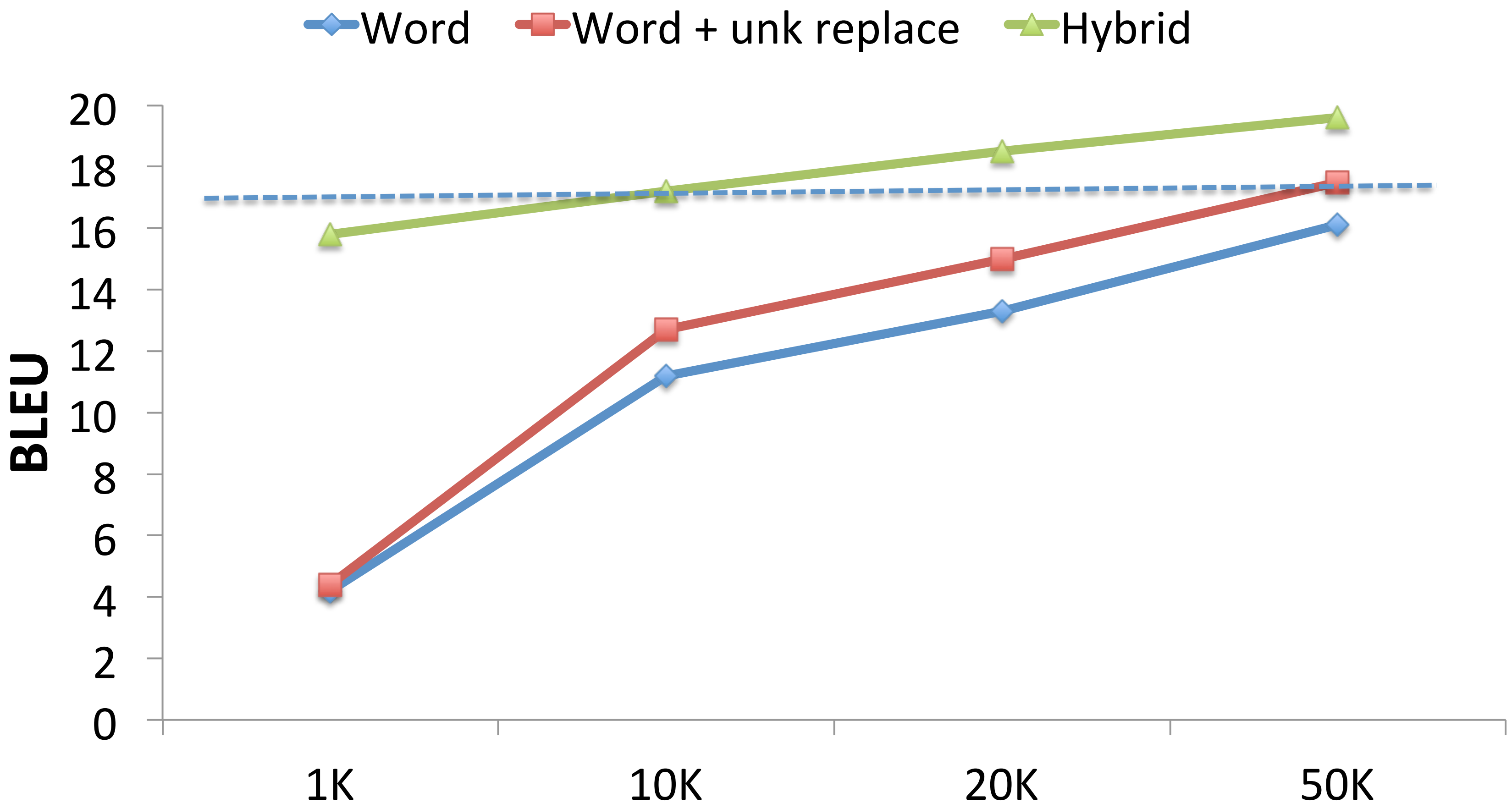


Effects of Vocabulary Sizes



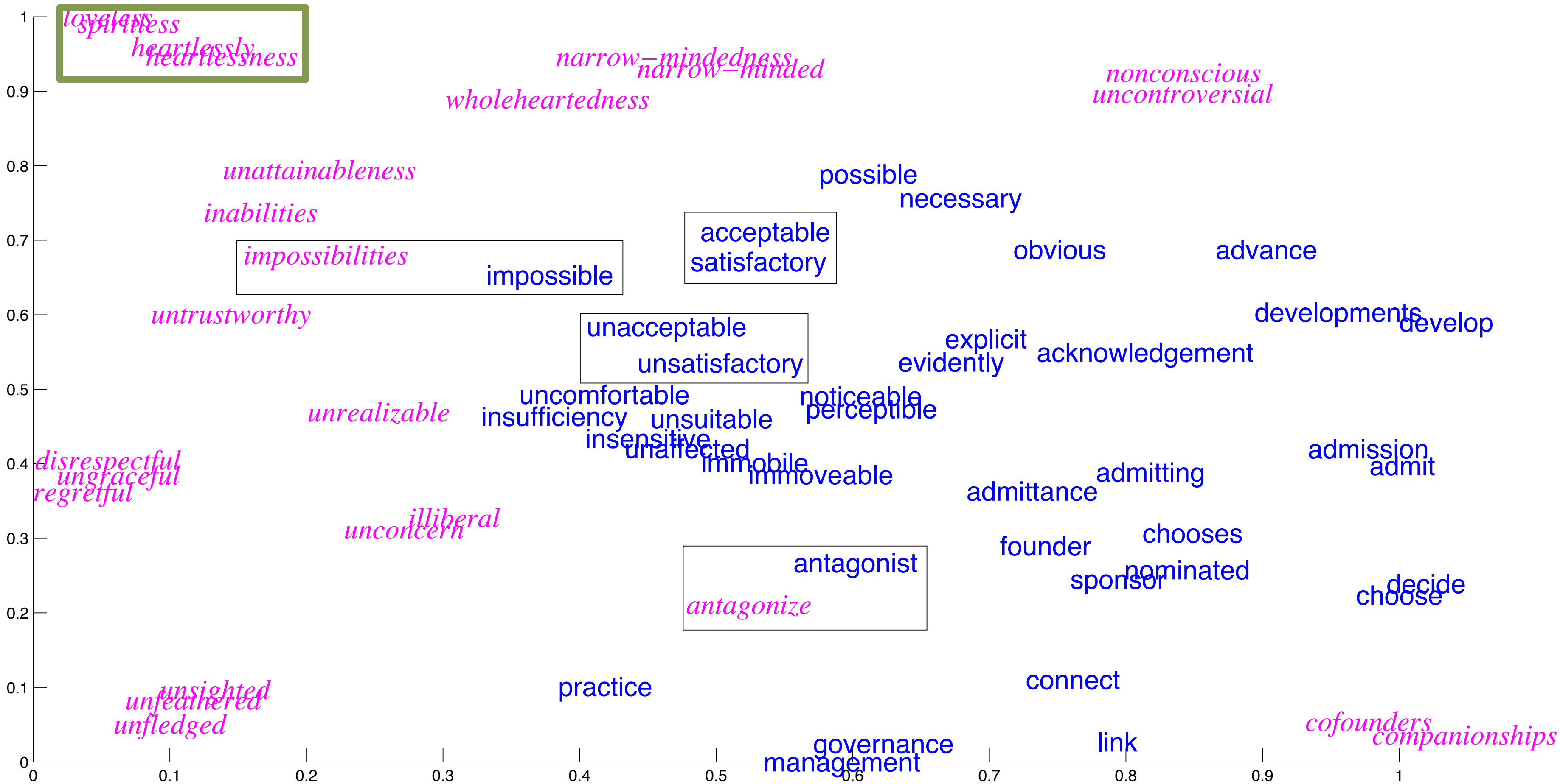
Additional gains of +2.1 ↦ +11.4 BLEU

Effects of Vocabulary Sizes



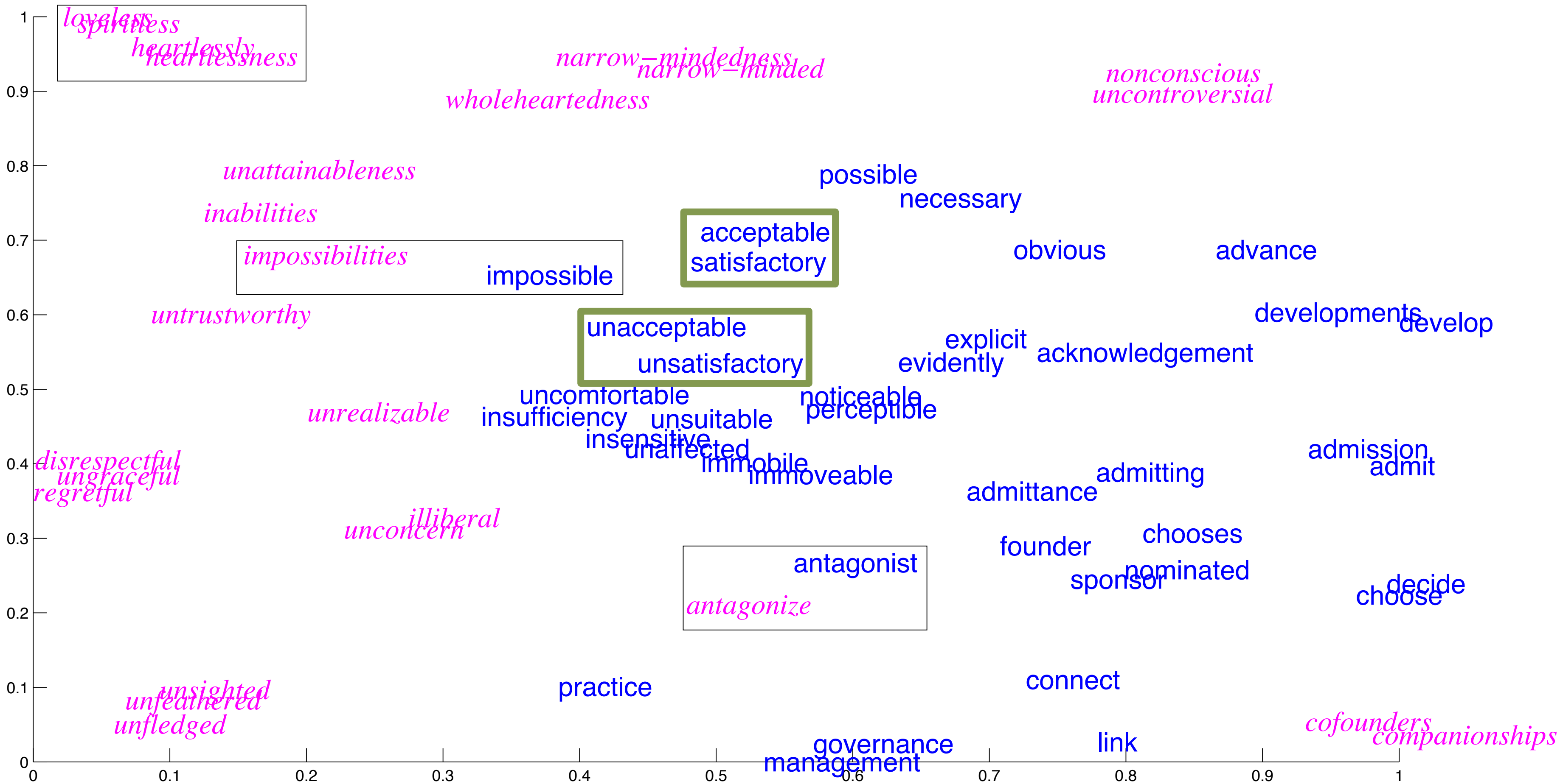
Small-vocab hybrid = *Large*-vocab word

Rare Word Embeddings



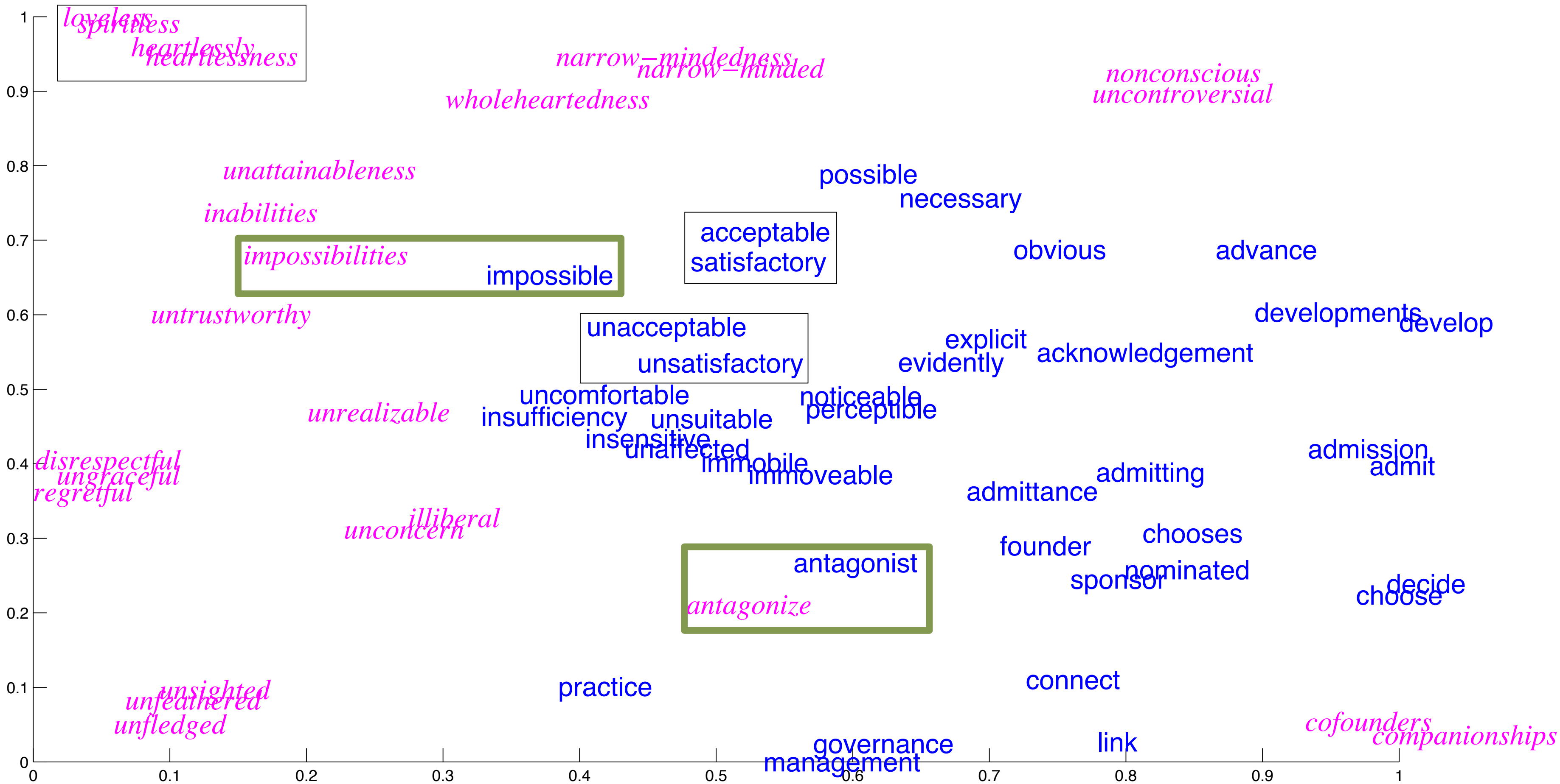
- Word & character-based embeddings.

Rare Word Embeddings



- Word & character-based embeddings.

Rare Word Embeddings



- Word & character-based embeddings.

Sample English-Czech translations

source	The author ***Stephen Jay Gould*** died 20 years after ***diagnosis*** .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

Perfect translation!

Sample English-Czech translations

source	The author ***Stephen Jay Gould*** died 20 years after ***diagnosis*** .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor $\boxed{\textbf{Stepher Stepher}}$ zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

- *Char*-based: wrong name translation.

Sample English-Czech translations

source	The author *Stephen Jay Gould* died 20 years after *diagnosis* .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po **diagnóze** .
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po $\boxed{\textbf{po}}$.
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .

- *Word*-based: incorrect alignment

Sample English-Czech translations

source	The author *Stephen Jay Gould* died 20 years after *diagnosis* .
human	Autor **Stephen Jay Gould** zemřel 20 let po **diagnóze** .
char	Autor **Stepher Stepher** zemřel 20 let po $\boxed{\textbf{diagnóze}}$.
word	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po **po** .
hybrid	Autor Stephen Jay <unk> zemřel 20 let po <unk> .
	Autor **Stephen Jay Gould** zemřel 20 let po $\boxed{\textbf{diagnóze}}$.

- *Char*-based & hybrid: correct translation of **diagnóze**.

Sample English-Czech translations

source	As the Reverend *Martin Luther King* Jr. *said fifty years ago* :
human	Jak **před padesáti lety** řekl reverend **Martin Luther King** Jr . :
char	Jako reverend **Martin Luther** **král říkal** ***před padesáti lety*** :
word	Jak řekl reverend Martin <unk> King <unk> před padesáti lety :
	Jak řekl reverend **Martin Luther King** **řekl** ***před padesáti lety*** :
hybrid	Jak řekl reverend Martin <unk> King <unk> před padesáti lety :
	Jak **před padesáti lety** řekl reverend **Martin Luther King** Jr. :

Correct reordering!

Sample English-Czech translations

source	As the Reverend *Martin Luther King* Jr. *said fifty years ago* :
human	Jak **před padesáti lety** řekl reverend **Martin Luther King** Jr . :
char	Jako reverend **Martin Luther** $\boxed{\textbf{král}}$ **říkal** ***před padesáti lety*** :
word	Jak řekl reverend Martin <unk> King <unk> před padesáti lety :
	Jak řekl reverend **Martin Luther King** **řekl** ***před padesáti lety*** :
hybrid	Jak řekl reverend Martin <unk> King <unk> před padesáti lety :
	Jak **před padesáti lety** řekl reverend **Martin Luther King** Jr. :

- *Char*-based: “král” means “king”.

Sample English-Czech translations

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
char	Její **jedenáctiletá** dcera , **Shani Bartová** , říkala , že cítí trochu ***divně***
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani** **,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

Generate complex words!

Sample English-Czech translations

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
char	Její **jedenáctiletá** dcera , **Shani Bartová** , říkala , že cítí trochu ***divně***
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani** **,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

- *Word*-based: identity copy fails.

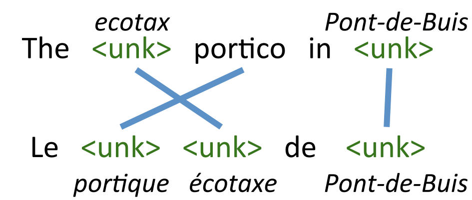
Sample English-Czech translations

source	Her *11-year-old* daughter , *Shani Bart* , said it felt a little bit *weird*
human	Její **jedenáctiletá** dcera **Shani Bartová** prozradila , že je to trochu **zvláštní**
char	Její **jedenáctiletá** dcera , **Shani Bartová** , říkala , že cítí trochu ***divně***
word	Její <unk> dcera <unk> <unk> řekla , že je to trochu divné
	Její **11-year-old** dcera **Shani** **,** řekla , že je to trochu ***divné***
hybrid	Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk>
	Její **jedenáctiletá** dcera , **Graham** ***Bart*** , řekla , že cítí trochu ***divný***

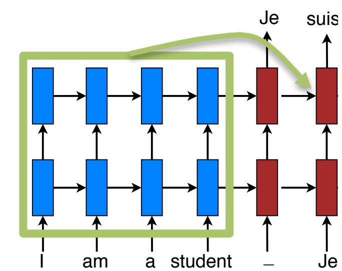
- *Hybrid*: translate names incorrectly.

We have advanced NMT

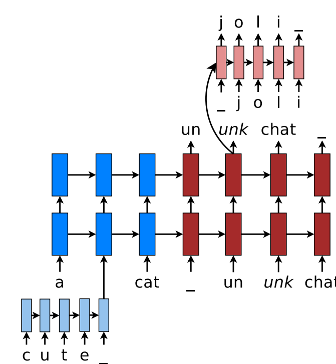
- #1: the ***vocabulary size*** problem
  - *Sol*: “copy” mechanism.
  - SOTA English-French



- #2: the ***sentence length*** problem
  - *Sol*: attention mechanism.
  - SOTA English-German



- #3: the ***language complexity*** problem
  - *Sol*: character-level translation.
  - SOTA English-Czech



NMT & beyond

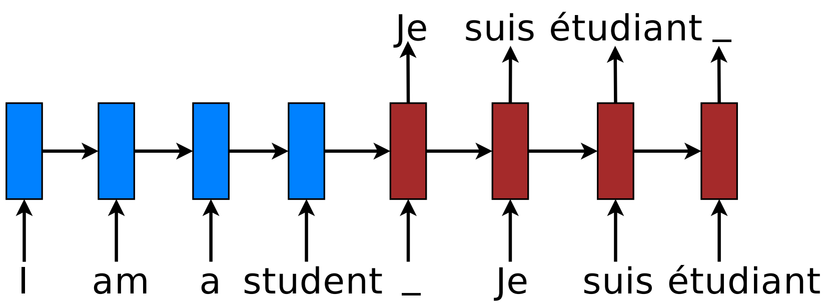
Thank you!

- Unsupervised learning for NMT
  - Utilize monolingual data.
- Long-context NMT
  - Translating an article / a book.
  - Smarter attention, longer sequences.
- Multi-modal Language Understanding System
  - Multi-lingual translation + speech recognition + more
  - Multi-task learning

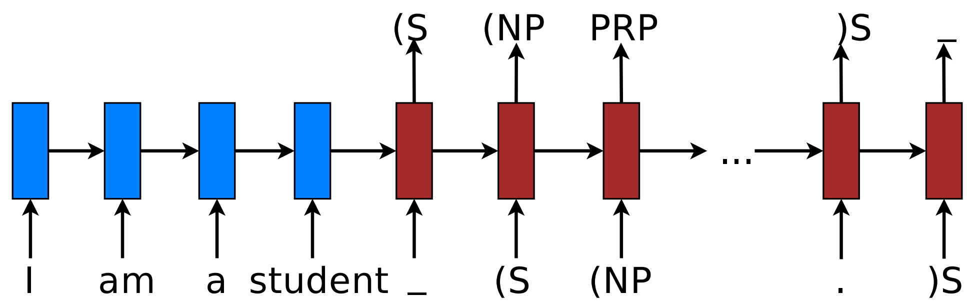
I’m done.
But if you are curious, read on!

#4 For the future of NMT

- Can we utilize all sequence-to-sequence data?



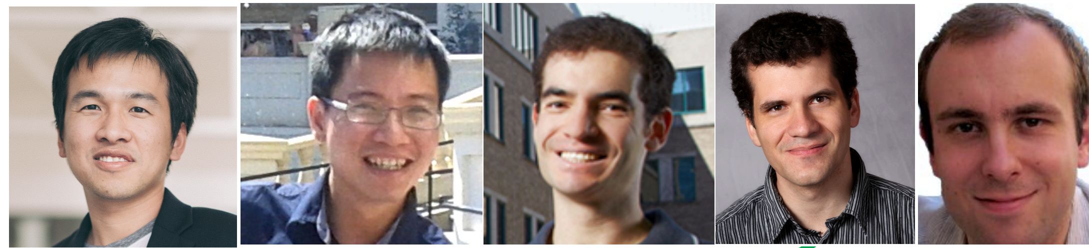
Machine translation



Constituent parsing

- Can we compress NMT for mobile devices?

Our work



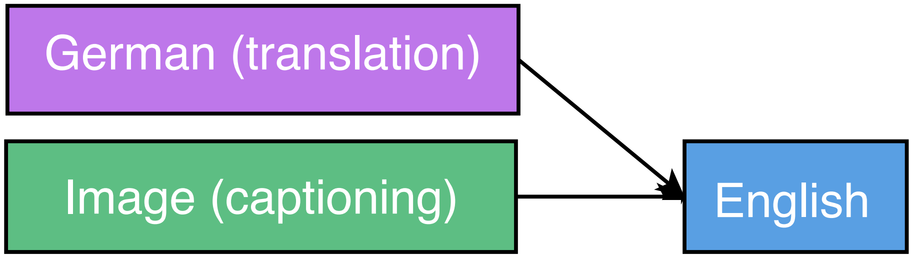
- Multi-task learning:
  - Machine translation
  - Constituent parsing
  - Image caption generation
  - Unsupervised learning

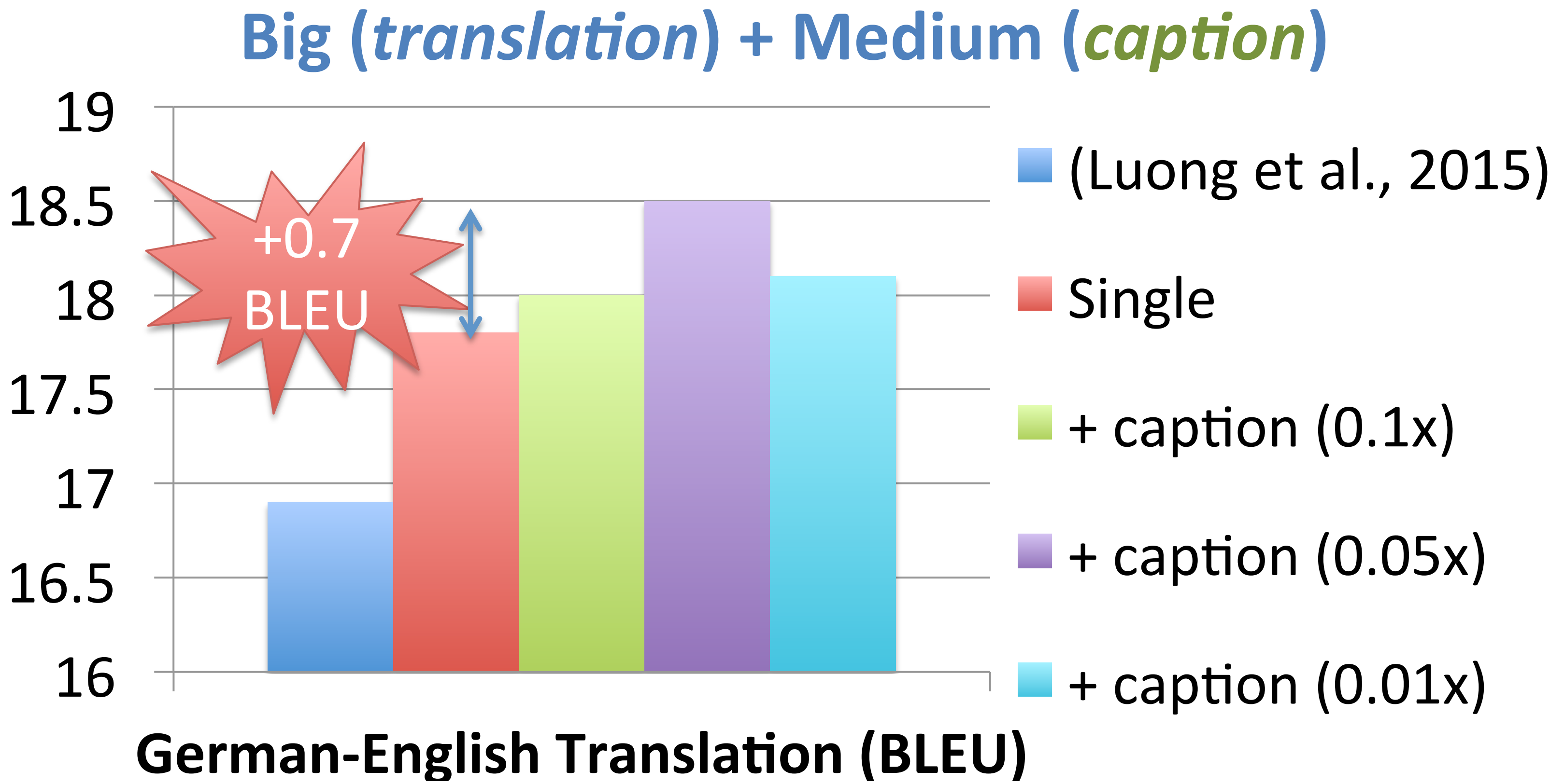
- Translation improvement: up to +1.5 BLEU.

- State-of-the-art in constituent parsing.

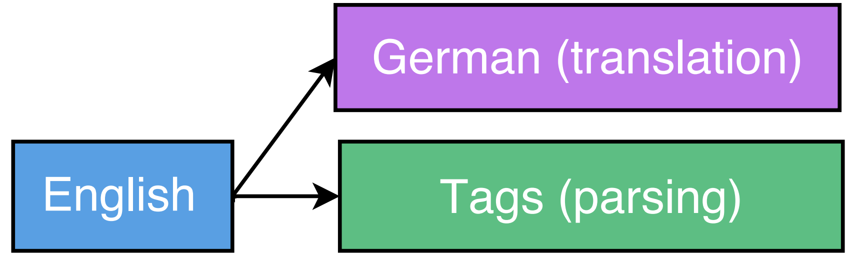
*Thang Luong, Quoc Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. **Multi-task sequence to sequence learning**. ICLR 2016.*

Many-to-one: shared decoder

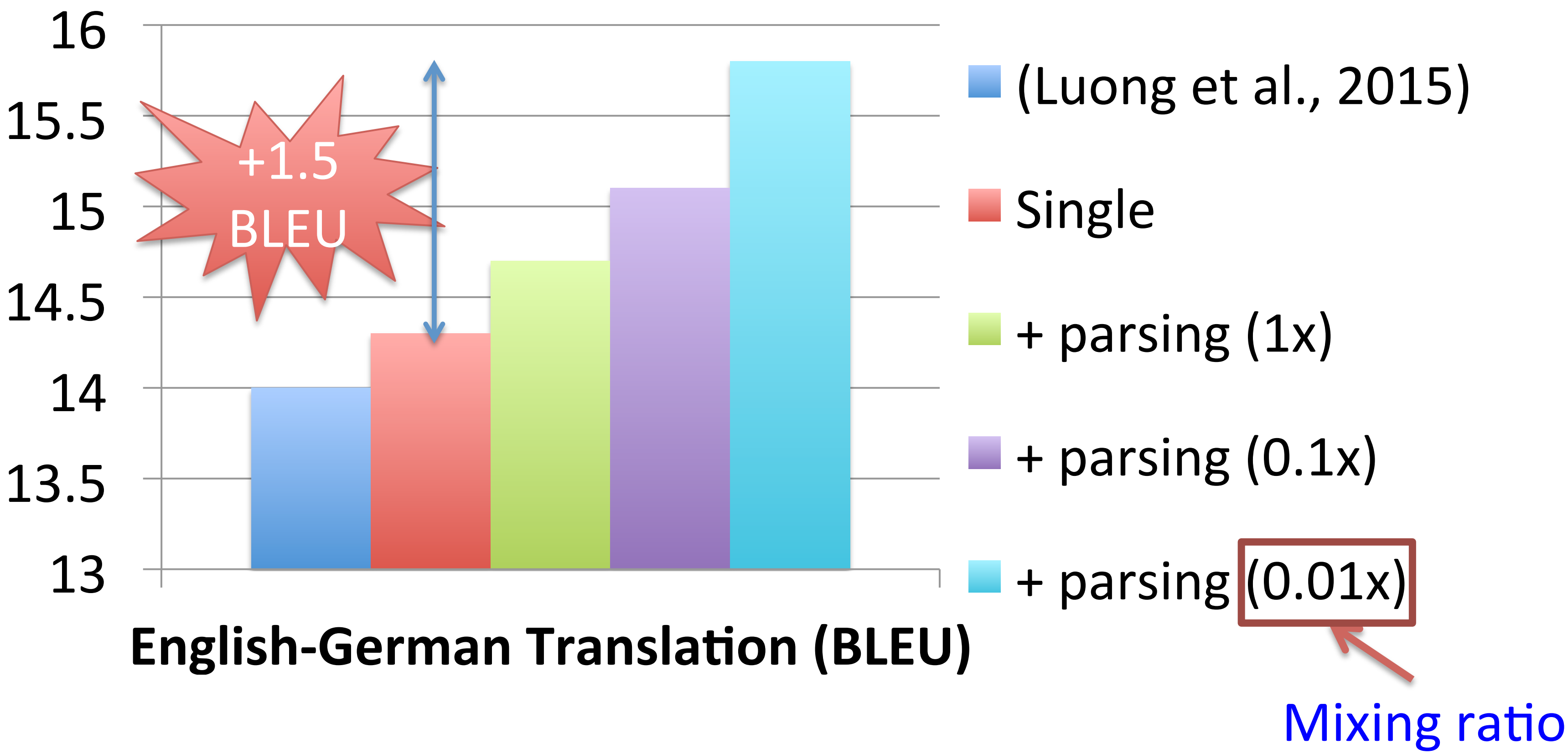




One-to-many: shared encoder

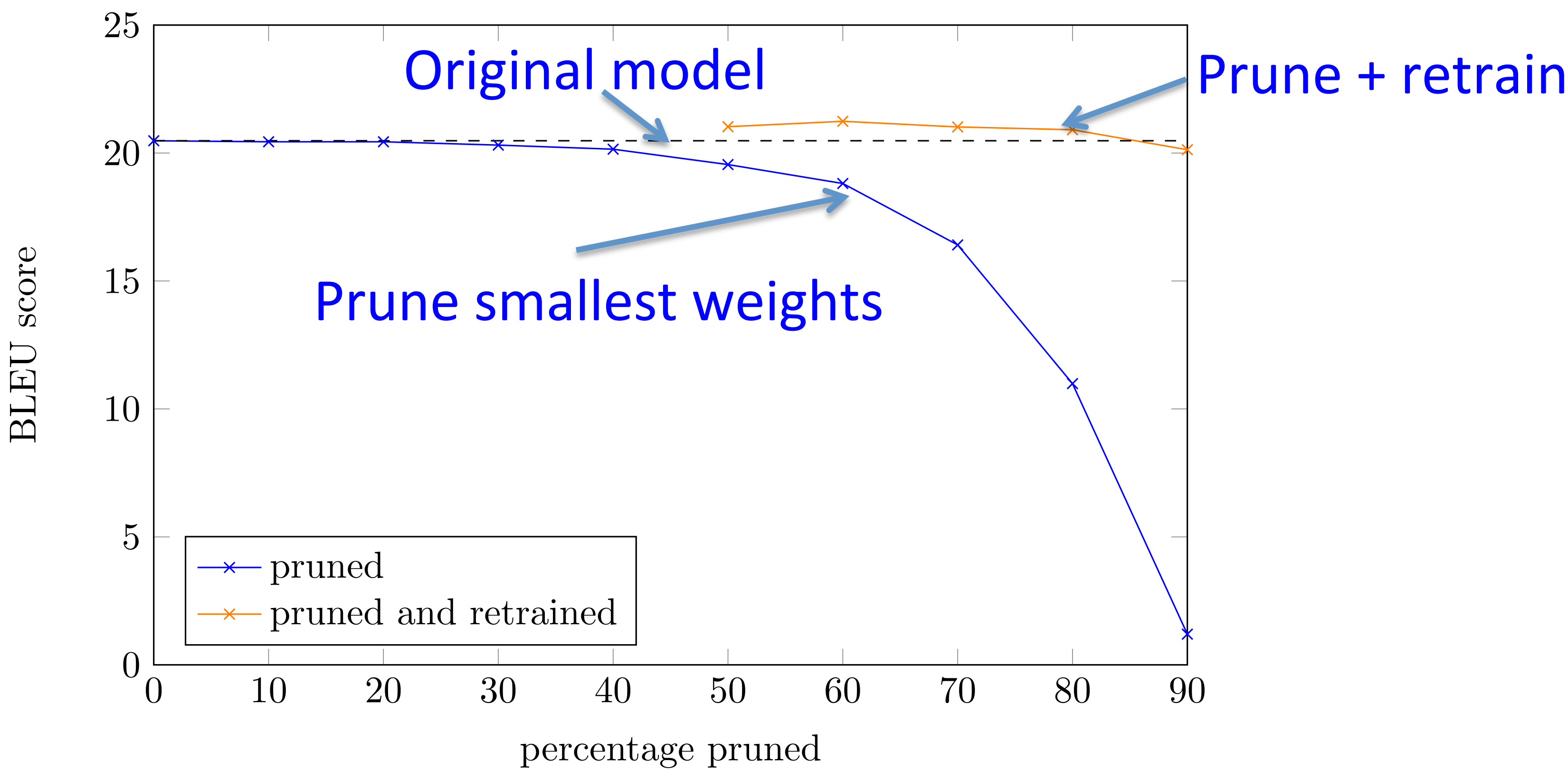


Big (*translation*) + Small (*PTB parsing*)



Our work

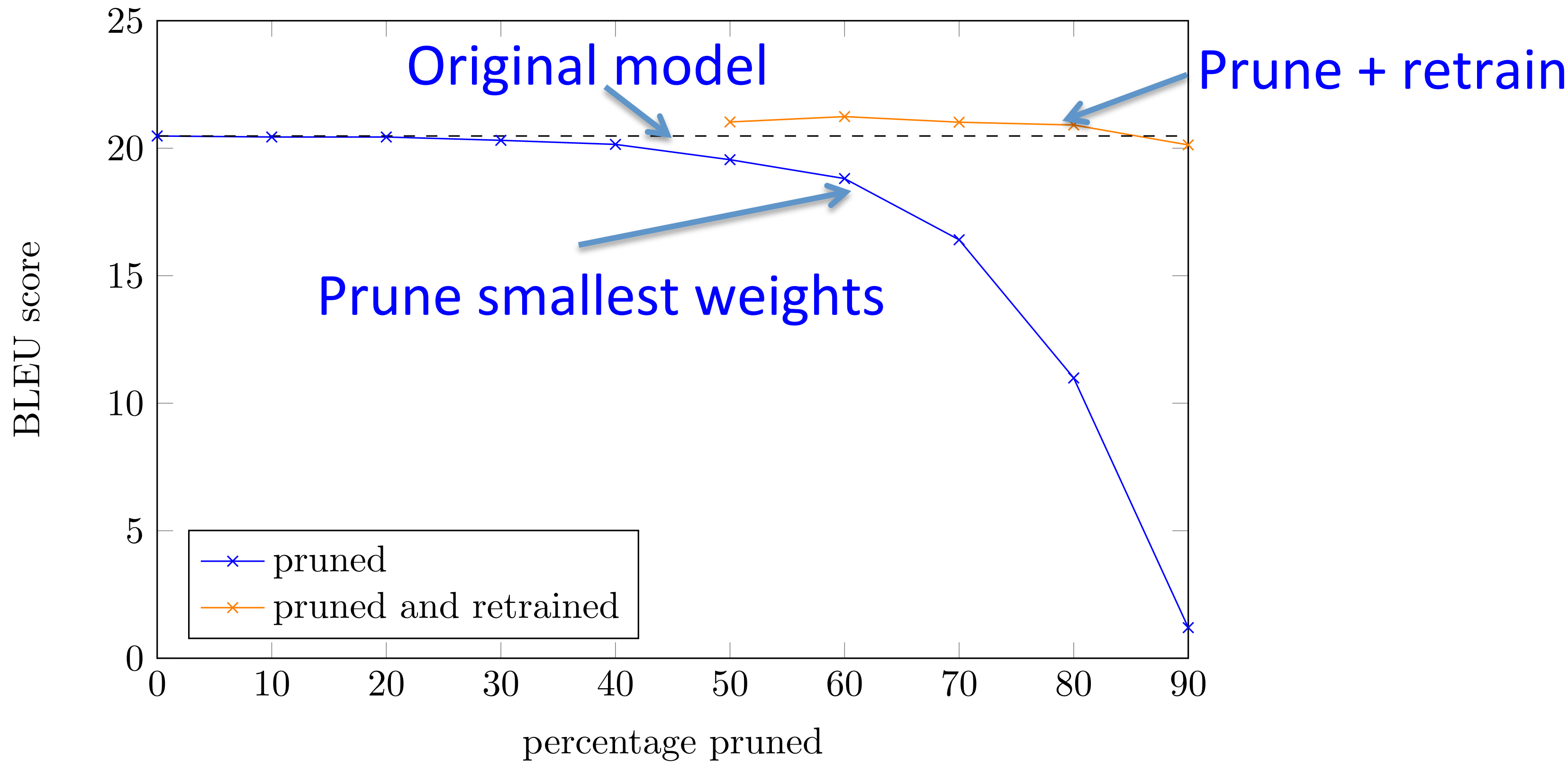

- Compress NMT via pruning & retraining:



Abigail See, Thang Luong*, and Chris Manning.* ***Compression of Neural Machine Translation Models via Pruning****. In submission.*

Our work

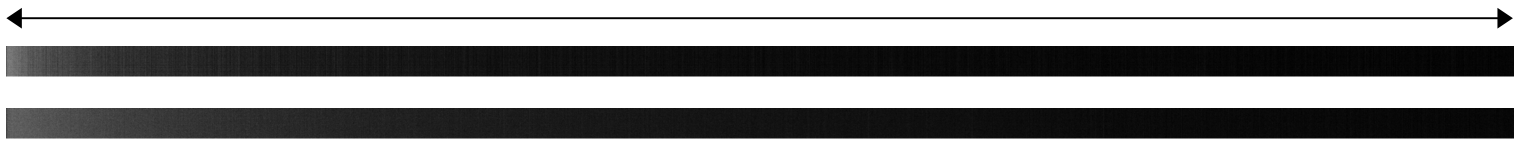
- Compress NMT via pruning & retraining:



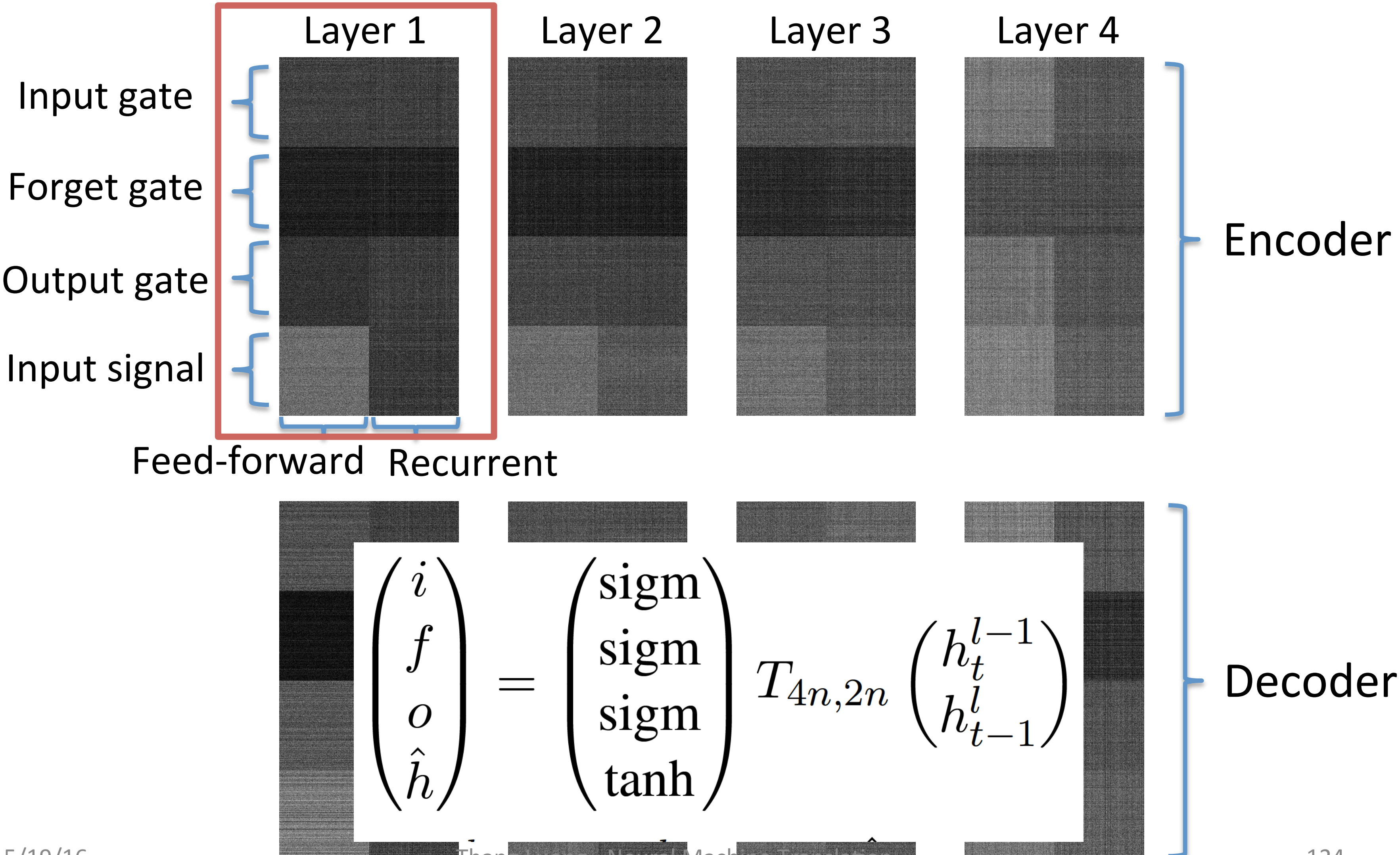
Prune 80% without loss of performance.

NMT Redundancy – *Embeddings*

- Frequent words have larger weights
  - white: large.
  - black: small.

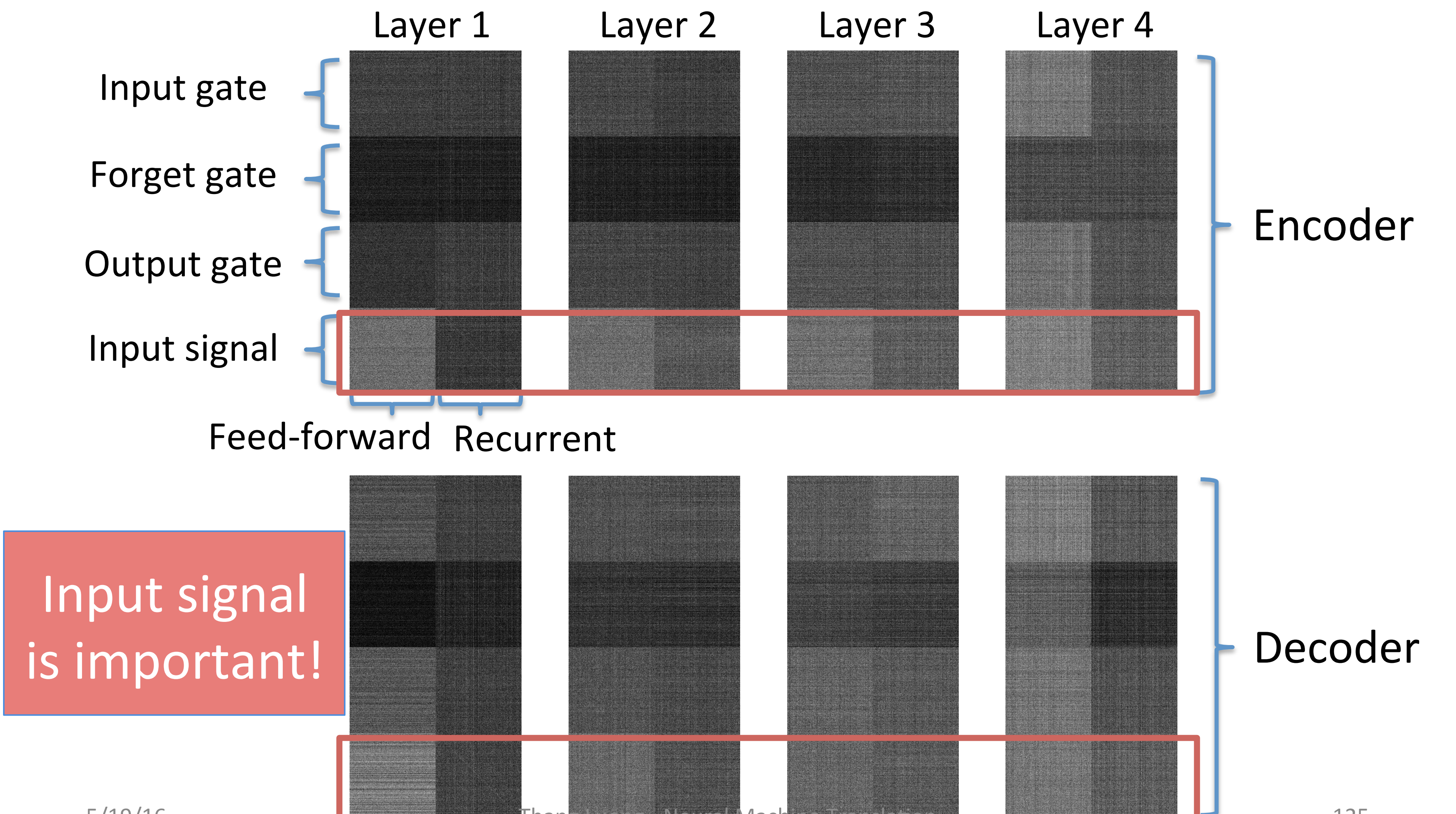


NMT Redundancy – *LSTM*

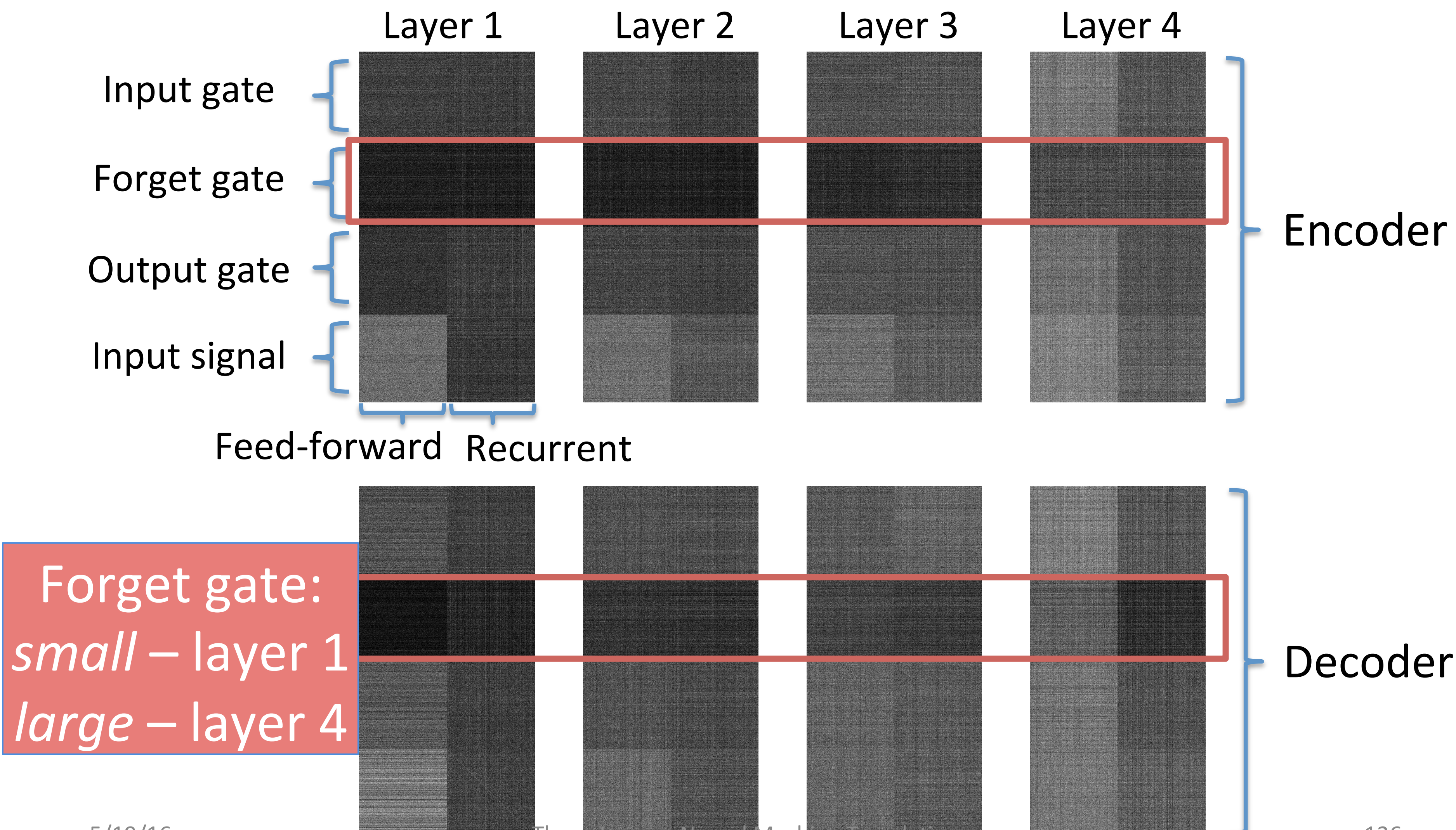


$$\begin{pmatrix} i \\ f \\ o \\ \hat{h} \end{pmatrix} = \begin{pmatrix} \text{sigm} \\ \text{sigm} \\ \text{sigm} \\ \tanh \end{pmatrix} T_{4n,2n} \begin{pmatrix} h_t^{l-1} \\ h_{t-1}^{l} \end{pmatrix}$$

NMT Redundancy – *LSTM*



NMT Redundancy – *LSTM*



Future Challenges

Thank you!

She saw an elephant in **her** dress.

She saw an elephant in **her** dress.

The elephant must have a good sense of fashion!

Needs to understand
common sense & larger context.