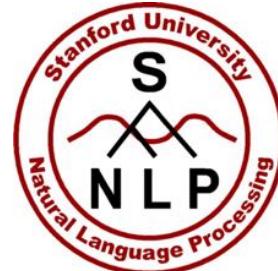


Neural Machine Translation



Thang Luong

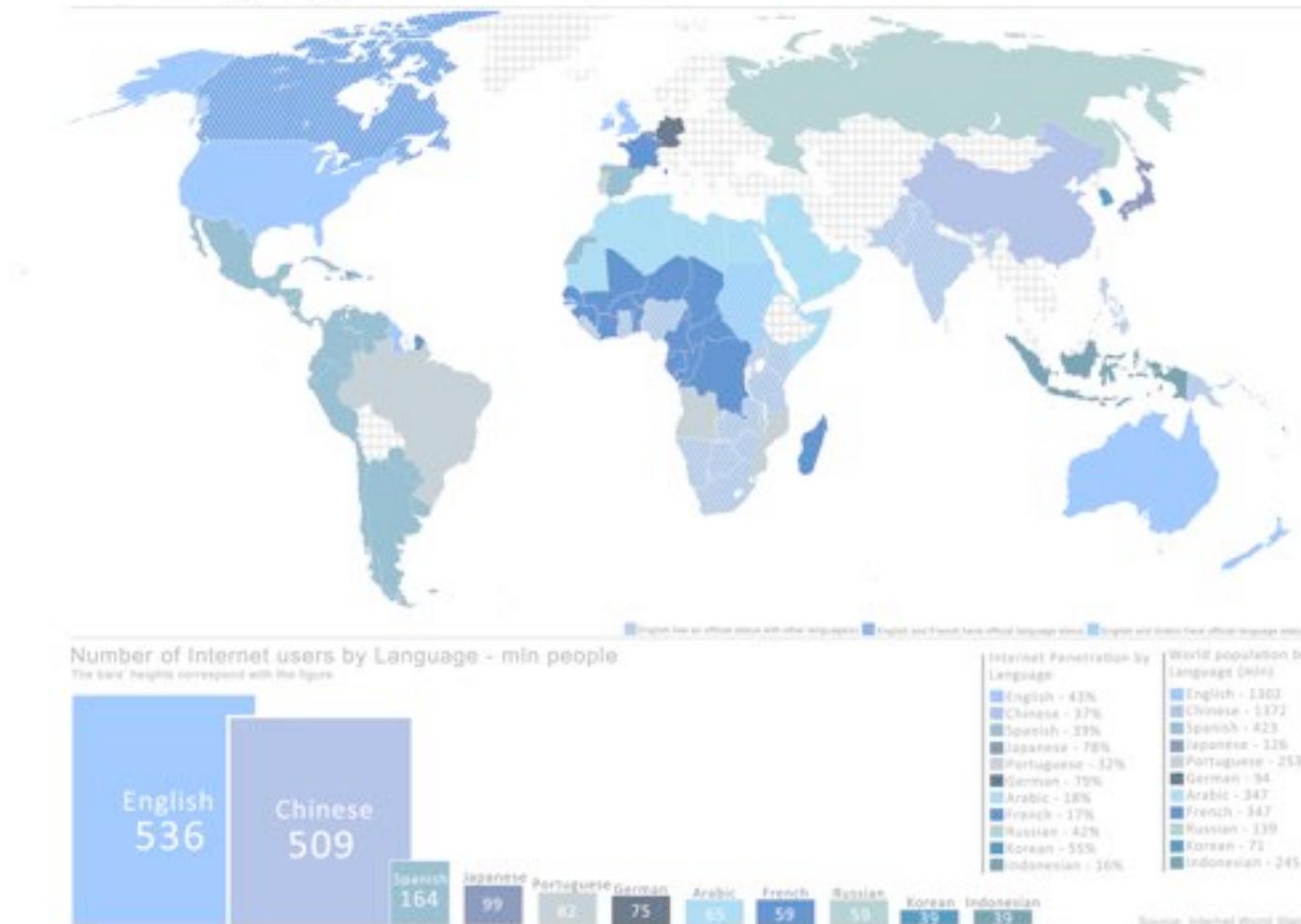
Lecture @ CS224D

Spring 2016

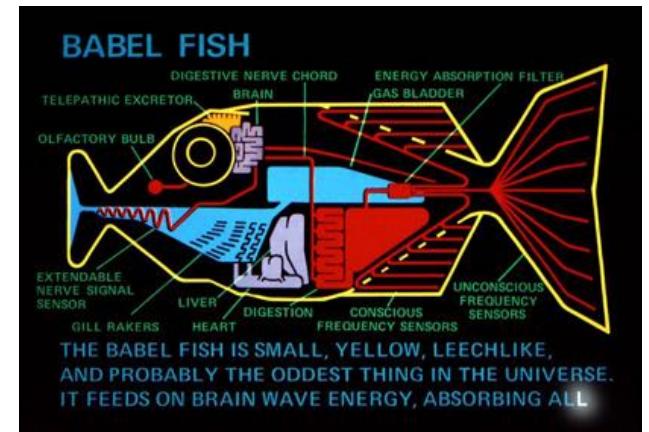
(Special thanks to Chris Manning for feedback!)

7 billion people, 7000 languages

Top Languages on the Internet



A universal translator



If you stick a Babel fish in your ear you can instantly understand anything said to you in any form of language.

Douglas Adams

(The Babel Fish from “the Hitchhiker's Guide to the Galaxy”)

Machine vs. Human Translation

serenpedia.com/quest-ce-que-le-deep-learning/

This page has been translated from French to English Show original

Google Translate

However, within the discipline of Machine Learning, developed specificity, that of Deep Learning.

Faithful translation

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



- Grammatically incorrect.

Machine vs. Human Translation

A screenshot of a web browser window. The address bar shows the URL: serenpedia.com/quest-ce-que-le-deep-learning/. To the right of the address bar is a "Google Translate" button, which includes a blue square icon with a white "A" and a stylized "文" character above it. Below the address bar, there is a toolbar with icons for back, forward, and home. A dropdown menu indicates the page has been translated from French to English, with options to "Show original".

However, within the discipline of Machine Learning, developed specificity, that of Deep Learning. Its **peculiarity** is to be "inspired by neurobiology.

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

serenpedia.com/quest-ce-que-le-deep-learning/

This page has been translated from French to English Show original

Google Translate

However, within the discipline of Machine Learning, developed specificity, that of Deep Learning. Its peculiarity is to be "inspired by neurobiology. Deep Learning aims to find IT elements allows a neural network to learn about the human brain model. "

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

serenpedia.com/quest-ce-que-le-deep-learning/

This page has been translated from French to English Show original

Google Translate

However, within the discipline of Machine Learning, developed specificity, that of Deep Learning. Its peculiarity is to be "inspired by neurobiology. Deep Learning aims to find IT elements allows a neural network to learn about the human brain model. "

Fluent translation

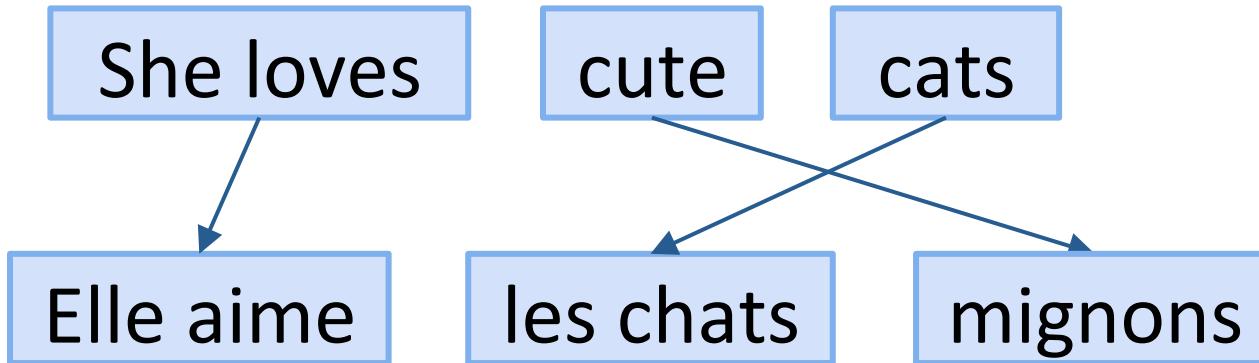
"However, in machine learning a specialization called deep learning has emerged. It can be recognized by its distinctive neurobiological influence. Deep learning is centered around networks of artificial neurons which can learn models of the human brain."



A big gap!

How has MT evolved?

Phrase-based MT



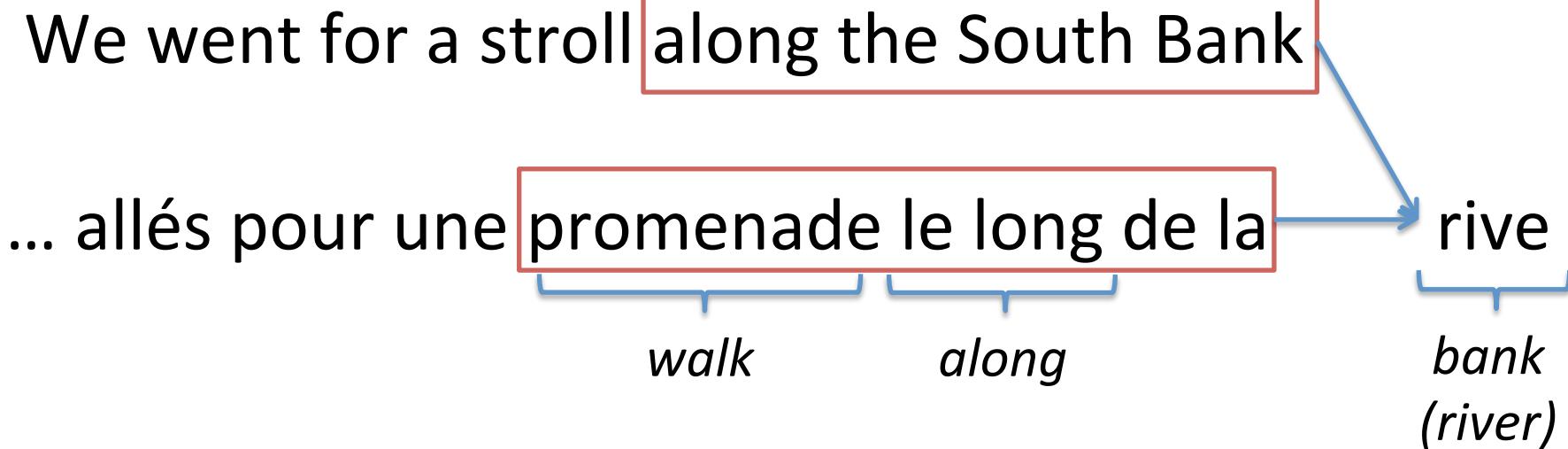
(Brown *et al.*, 1993; Koehn *et al.*, 2003; Och & Ney, 2004)

- 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

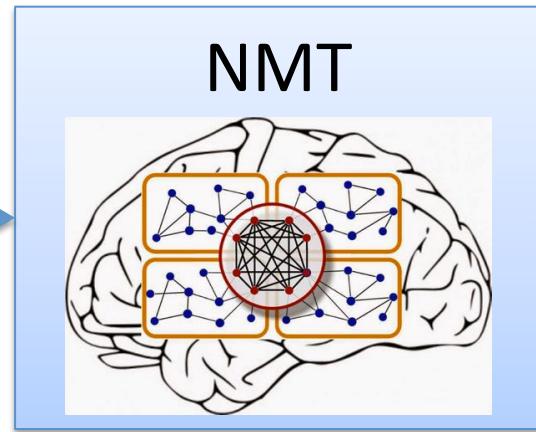


- Conditioned on **source words** (Devlin et al., 2014)
- Still translate **locally**.

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)

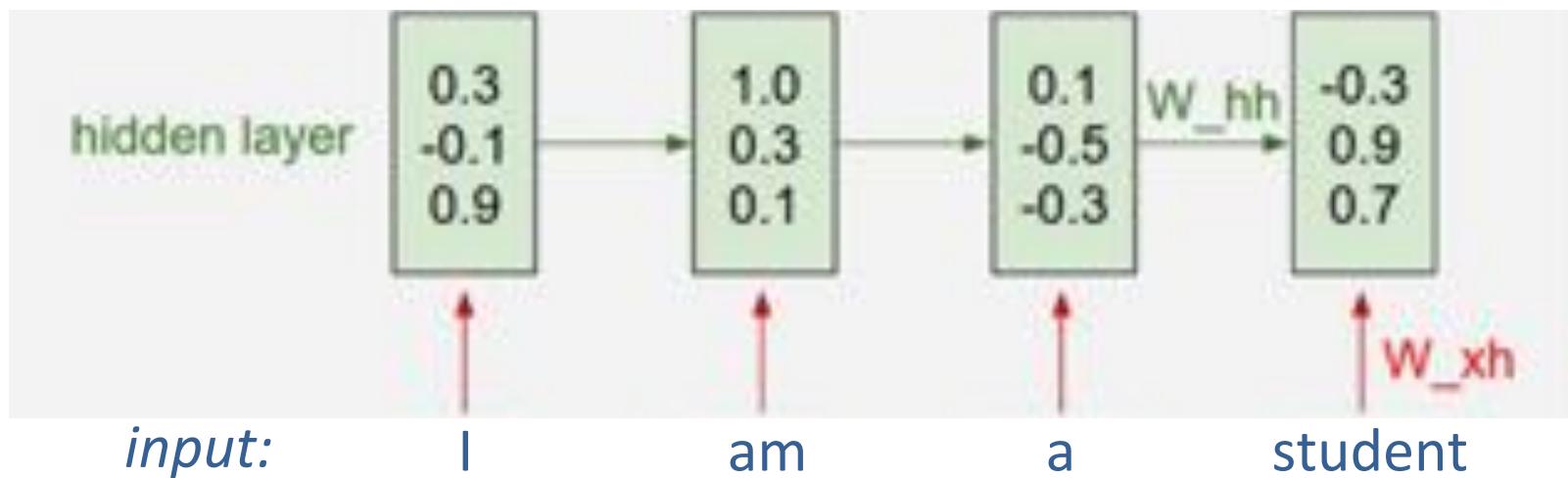
- *Sequence-to-sequence*: translate globally.
- *End-to-end*: simple & generalizable.

Let's find out!

Outline

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

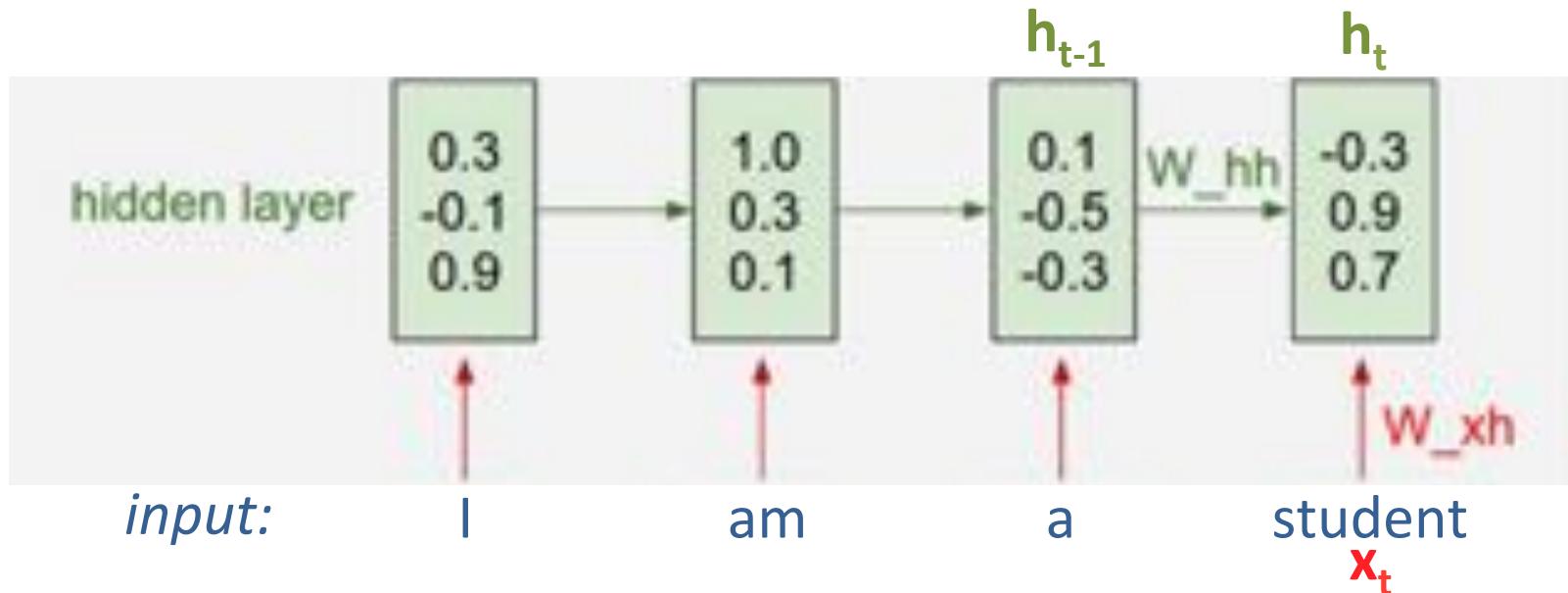
Recurrent Neural Networks (RNNs)



(Picture adapted from Andrej Karpathy)

Recurrent Neural Networks (RNNs)

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1})$$



RNNs to represent sequences!

(Picture adapted from Andrej Karpathy)

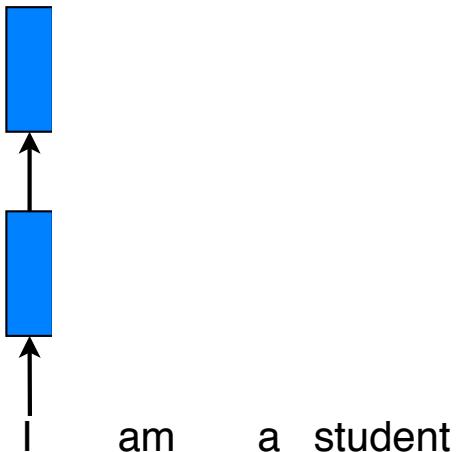
Neural Machine Translation (NMT)

I am a student

Je suis étudiant

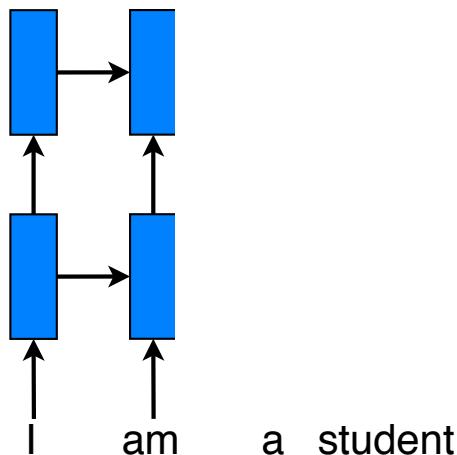
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



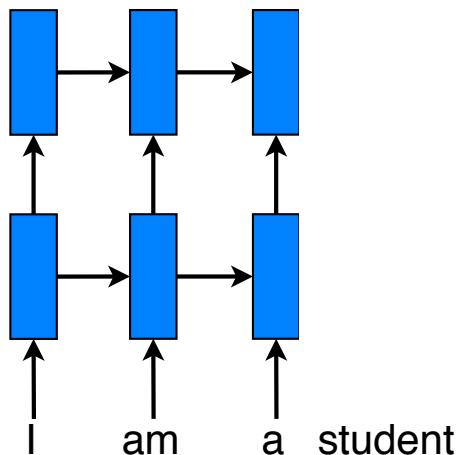
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



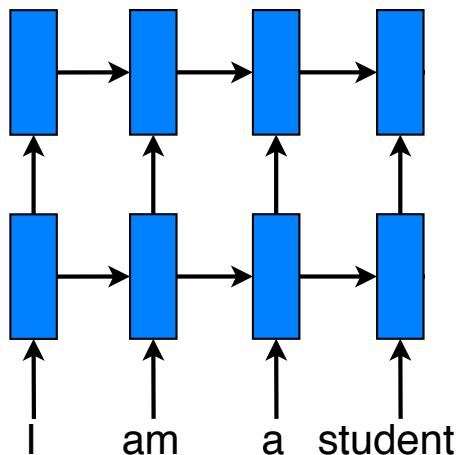
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



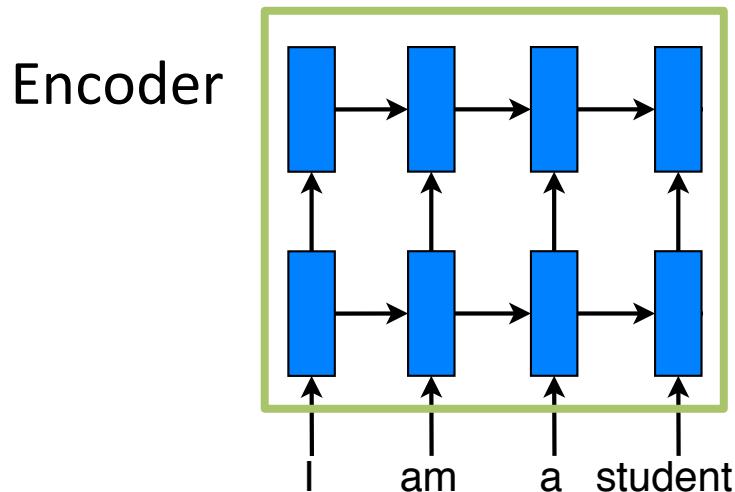
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



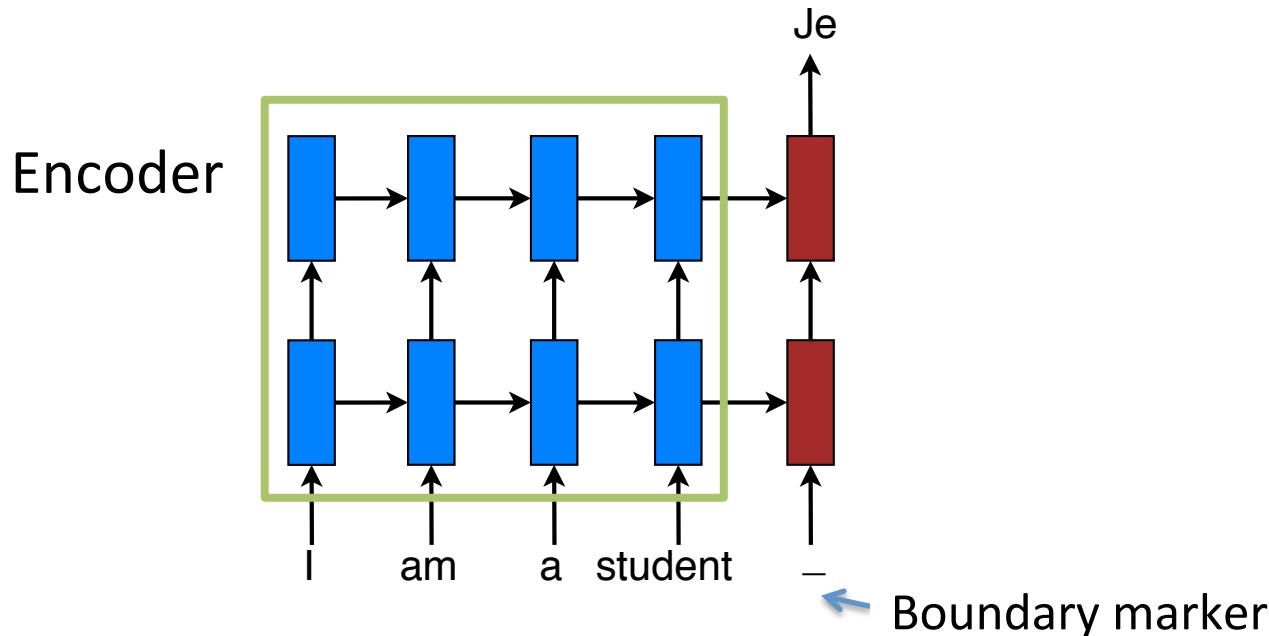
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



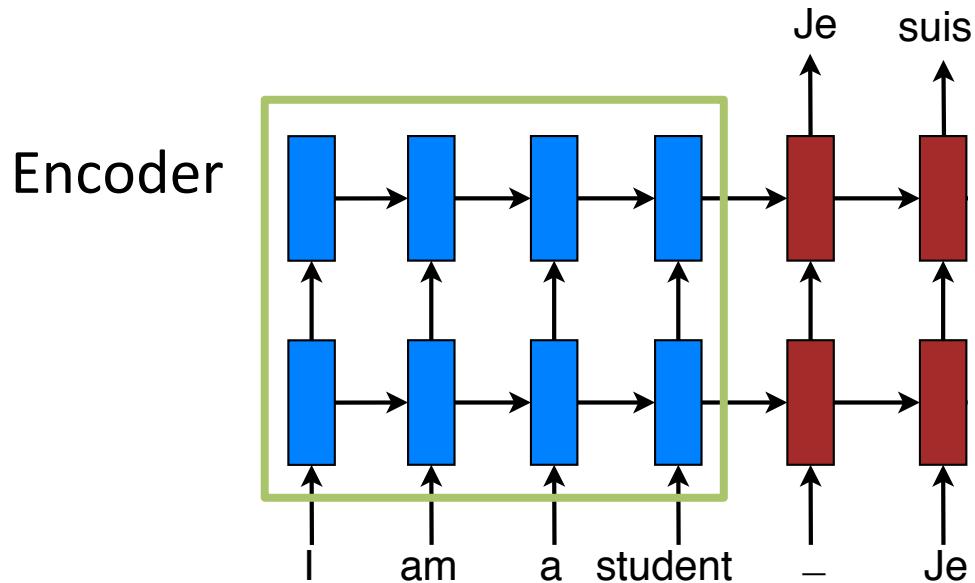
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



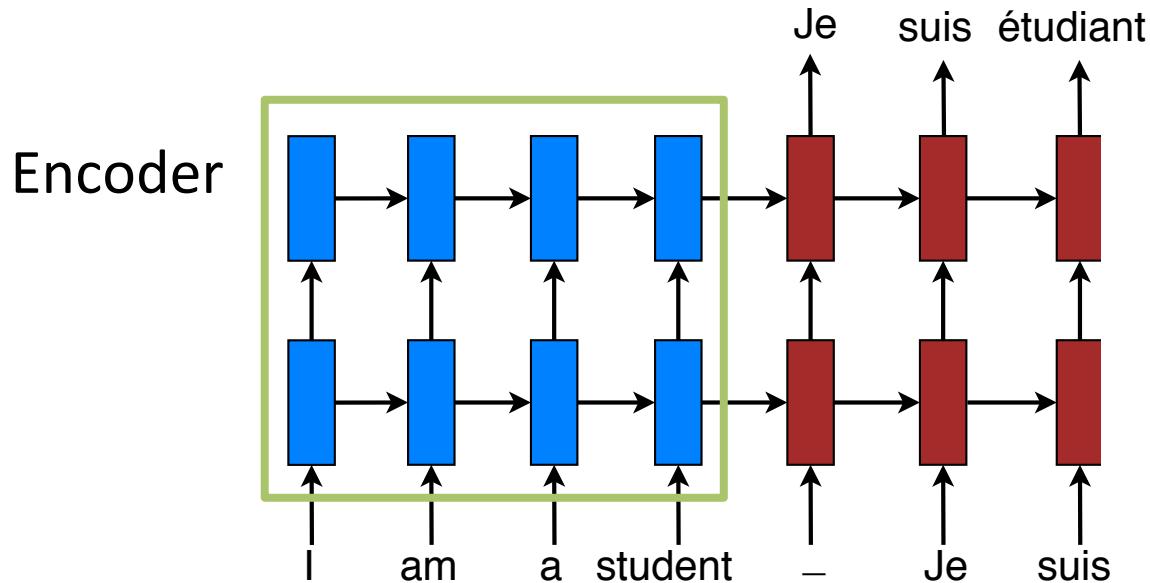
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



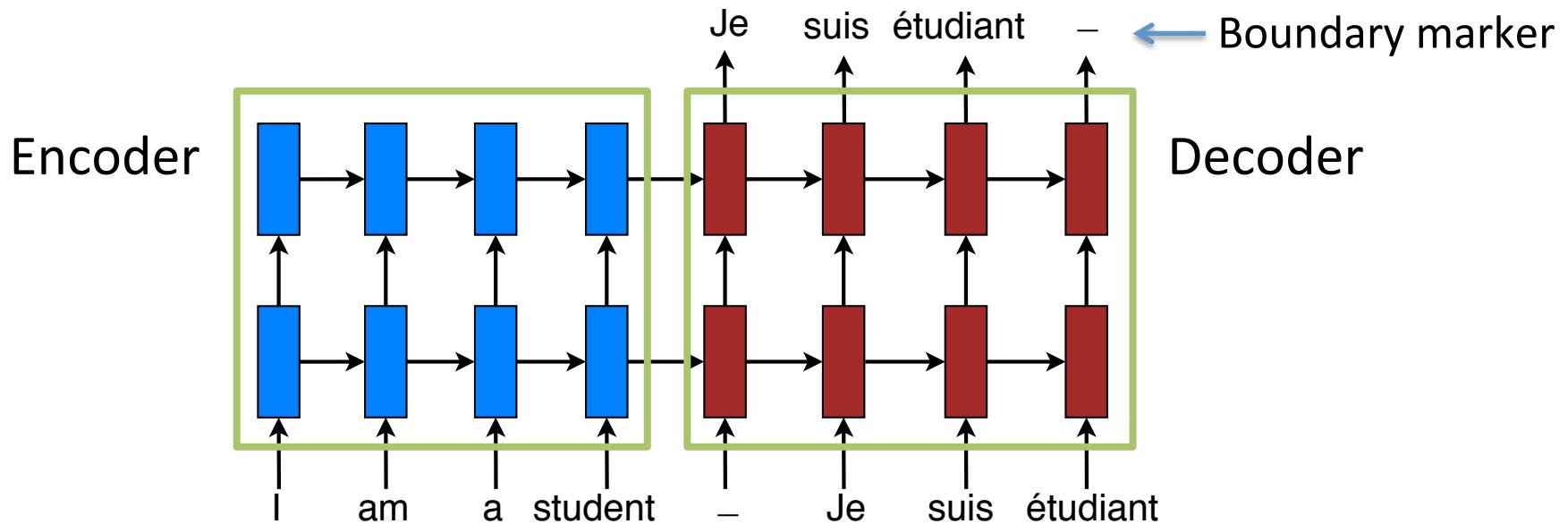
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



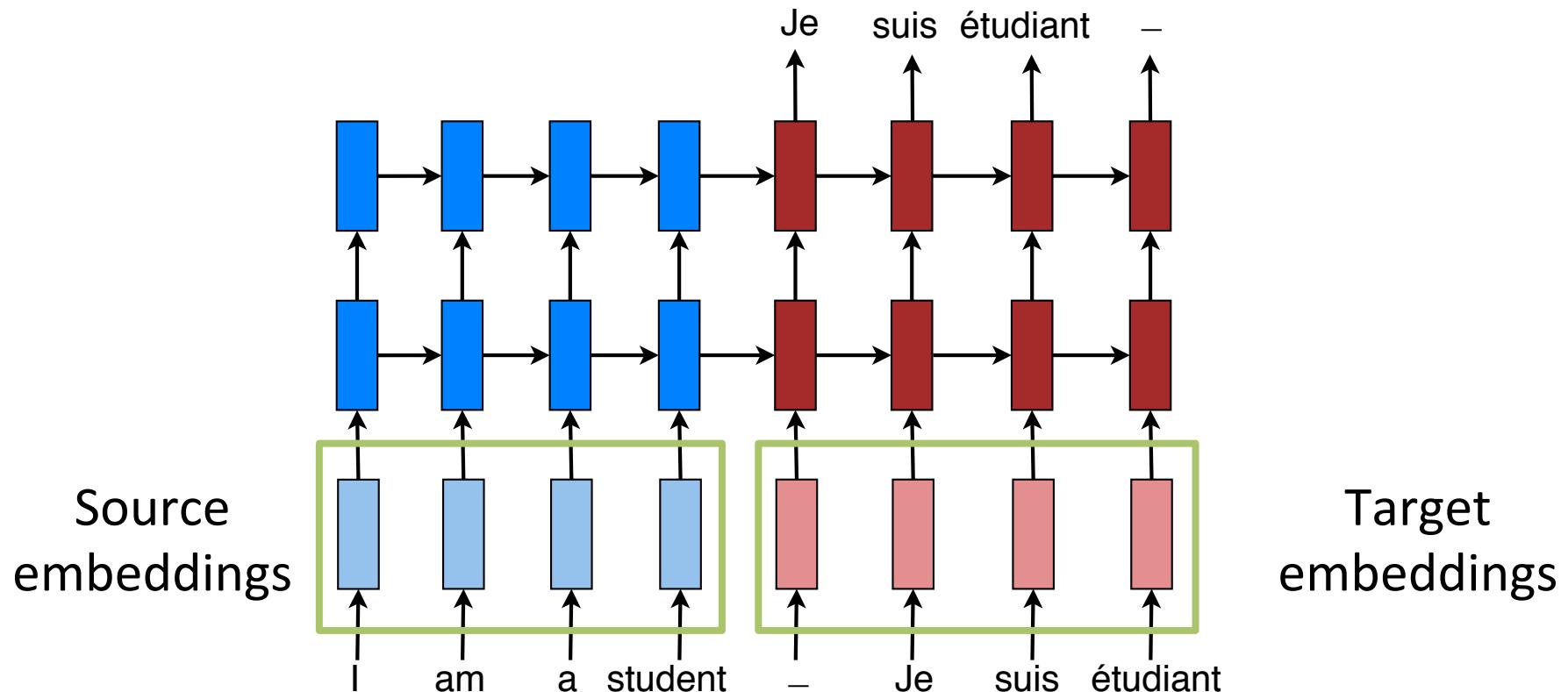
- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

Neural Machine Translation (NMT)



- Recurrent Neural Networks:
 - Model $P(\text{target} \mid \text{source})$ directly.
 - Can be trained end-to-end.

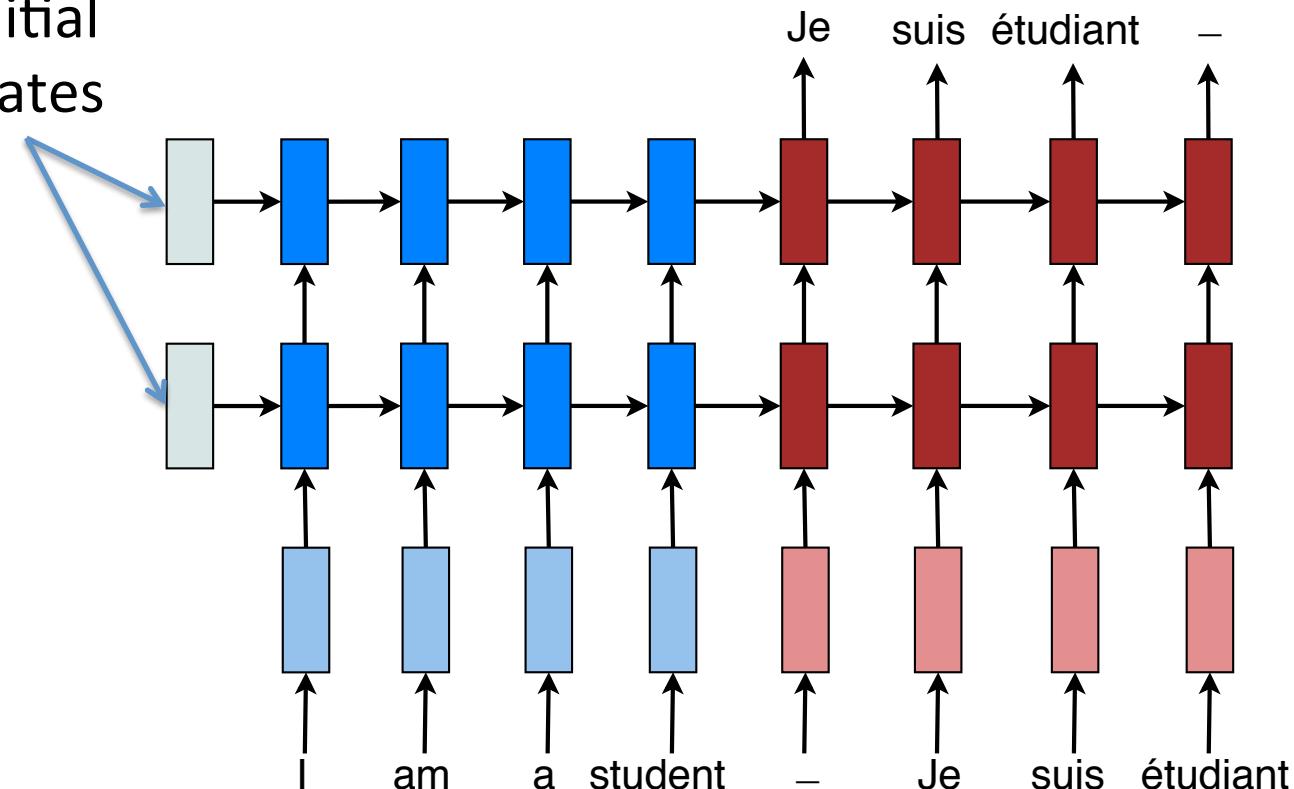
Word Embeddings



- One for each language: can learn from scratch.

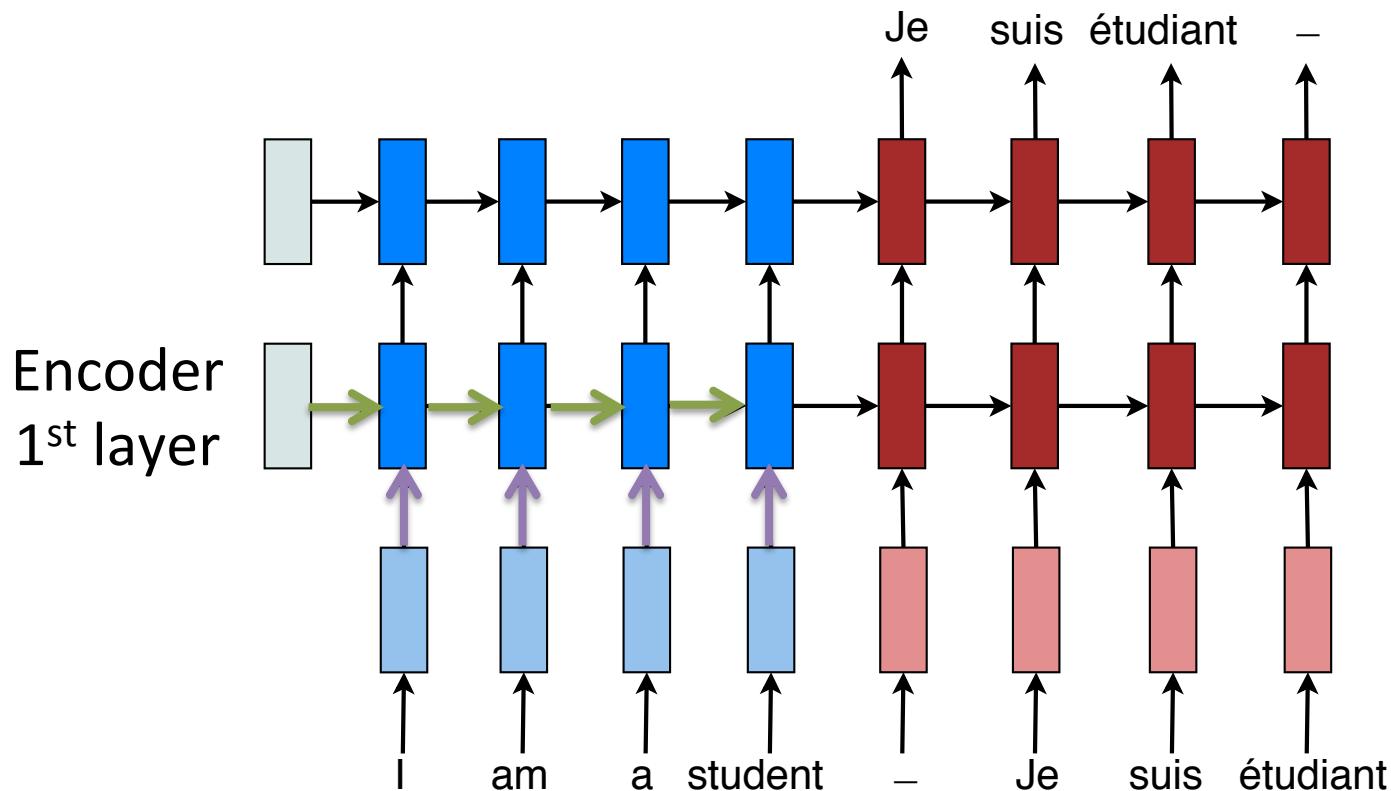
Recurrent Connections

Initial states



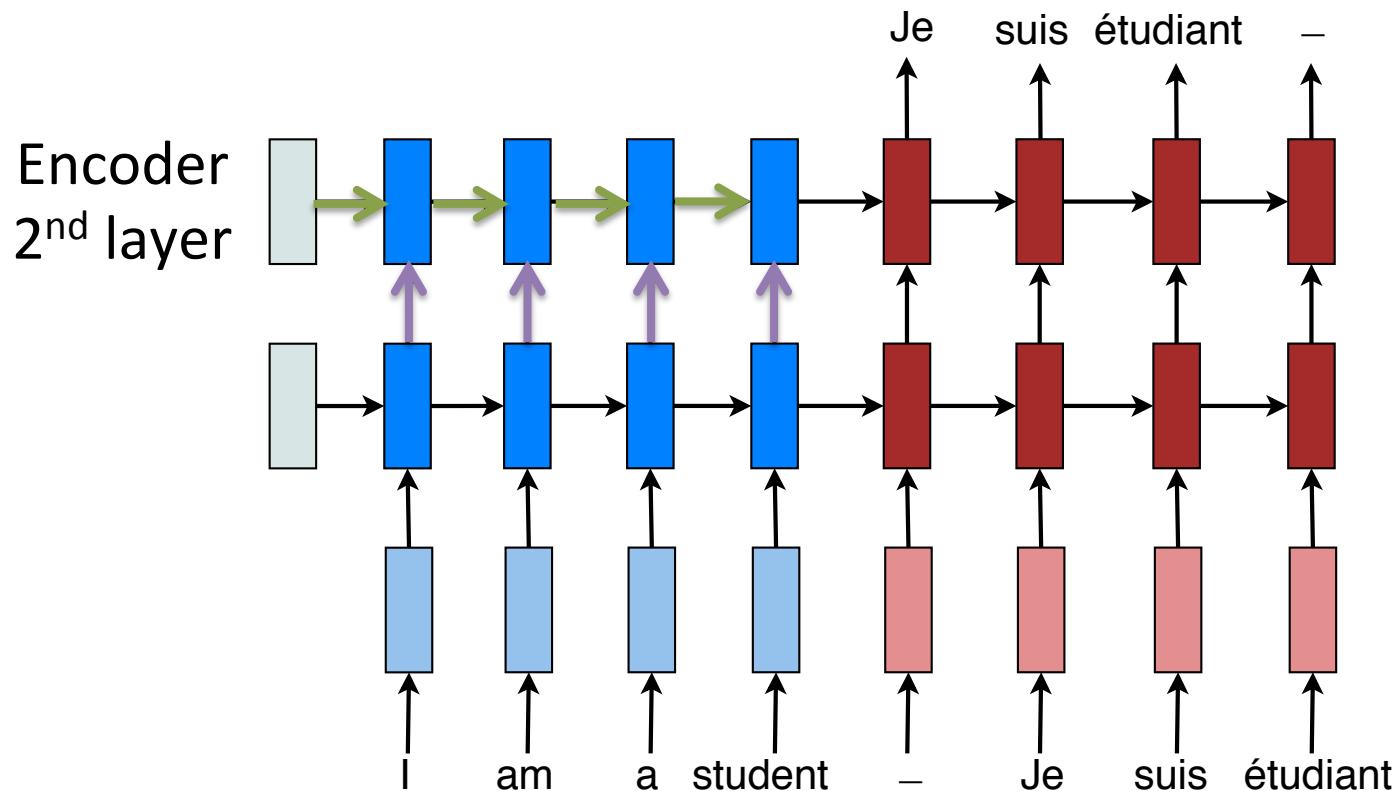
- Often set to 0.

Recurrent Connections



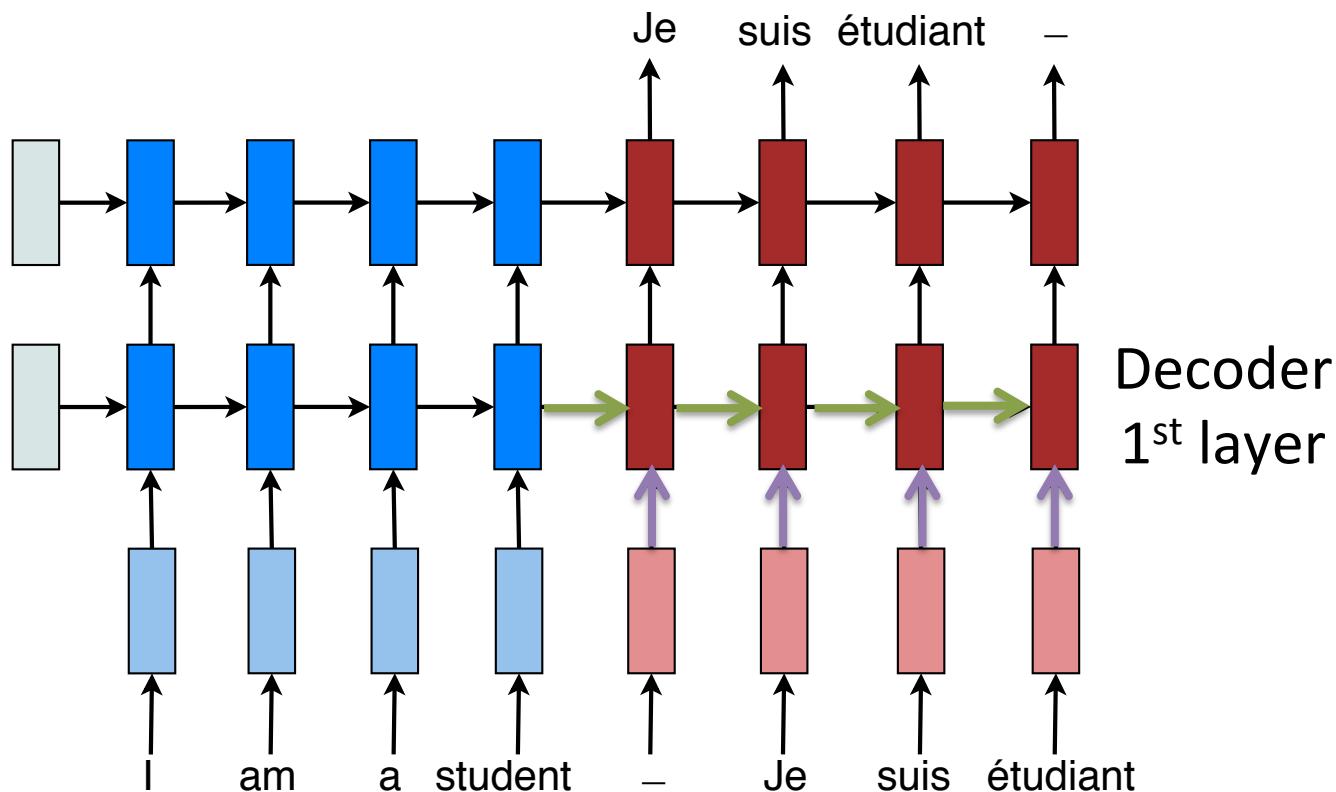
- **Different:** {1st layer, 2nd layer} x {encoder, decoder}.

Recurrent Connections



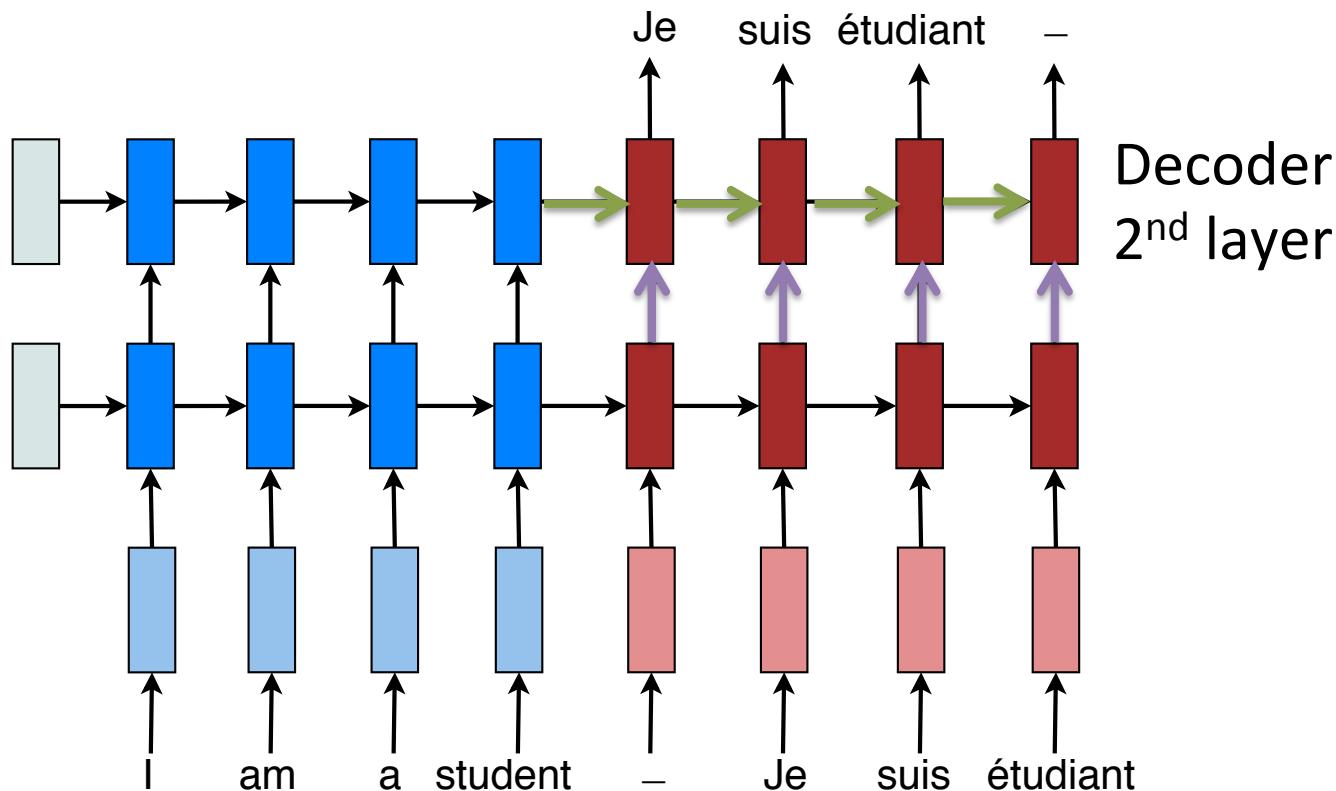
- **Different:** {1st layer, 2nd layer} x {encoder, decoder}.

Recurrent Connections



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

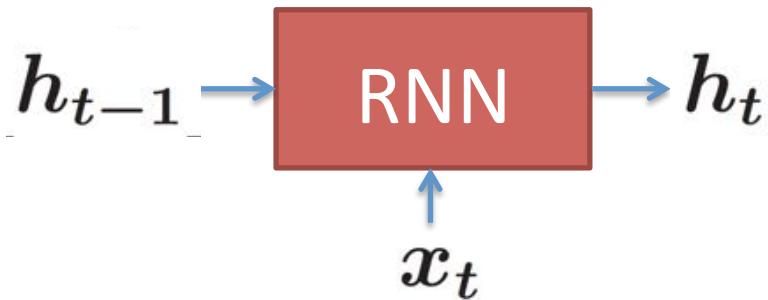
Recurrent Connections



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

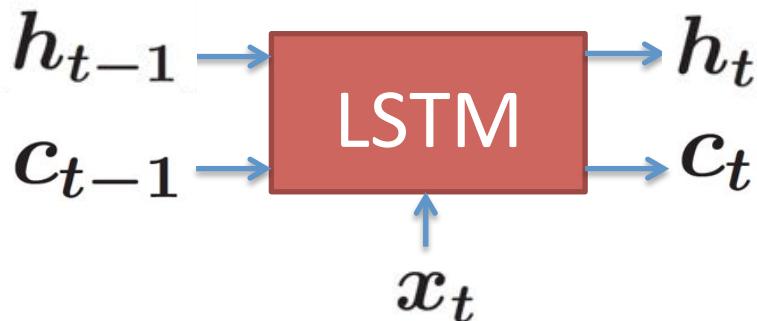
Recurrent Units

- Vanilla:



Vanishing
gradient problem!

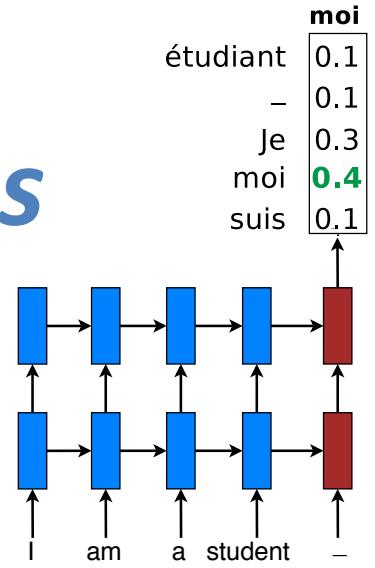
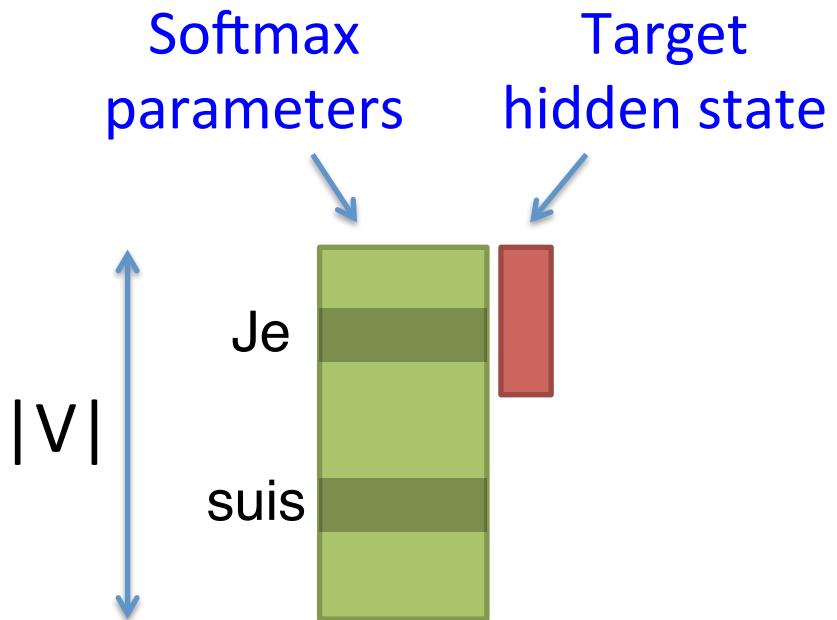
- LSTM:



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

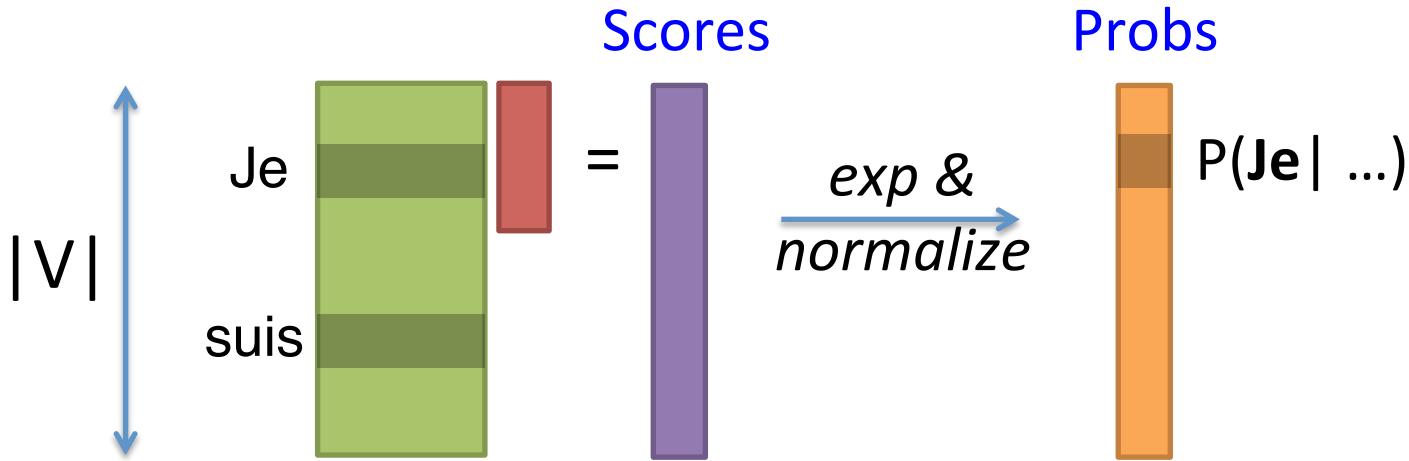
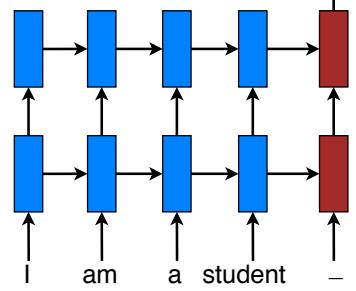


Softmax: vectors \mapsto categories



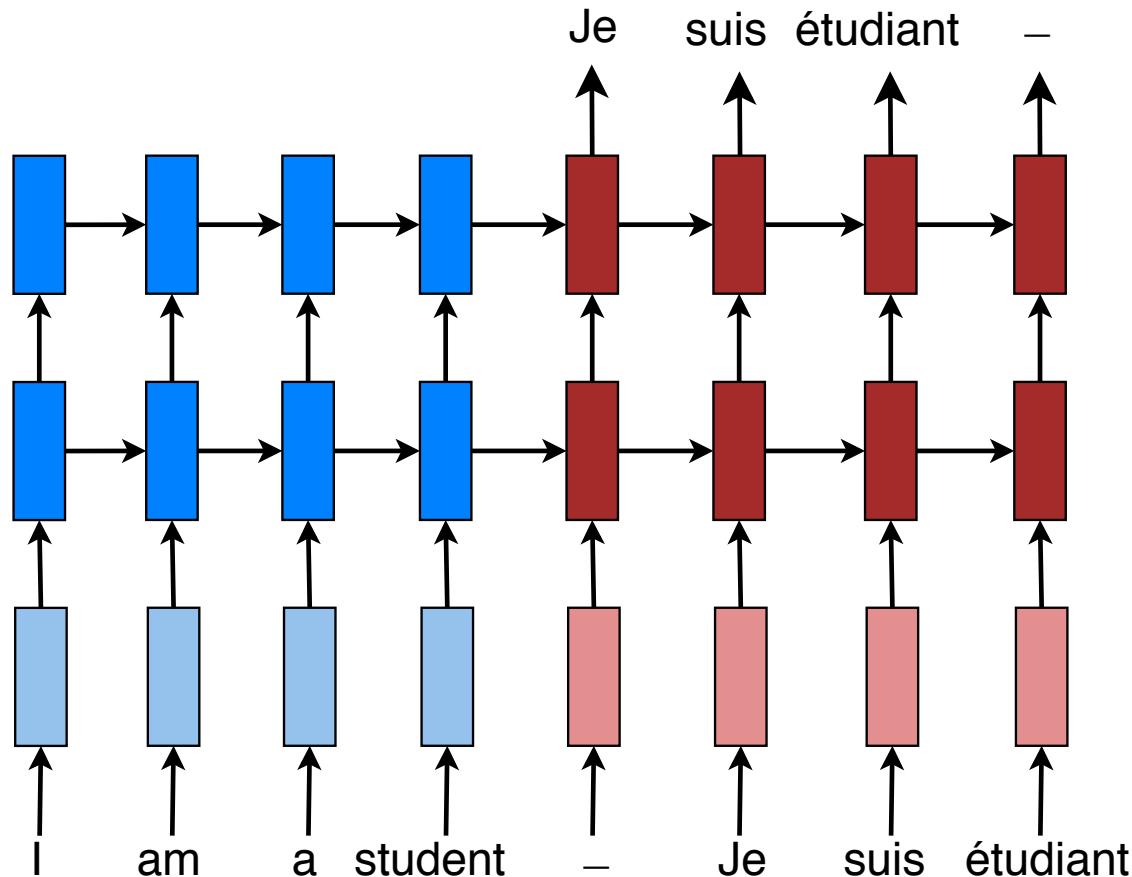
Softmax: vectors \mapsto categories

| moi | 0.1 |
|----------|-----|
| étudiant | 0.1 |
| - | 0.3 |
| Je | 0.3 |
| moi | 0.4 |
| suis | 0.1 |



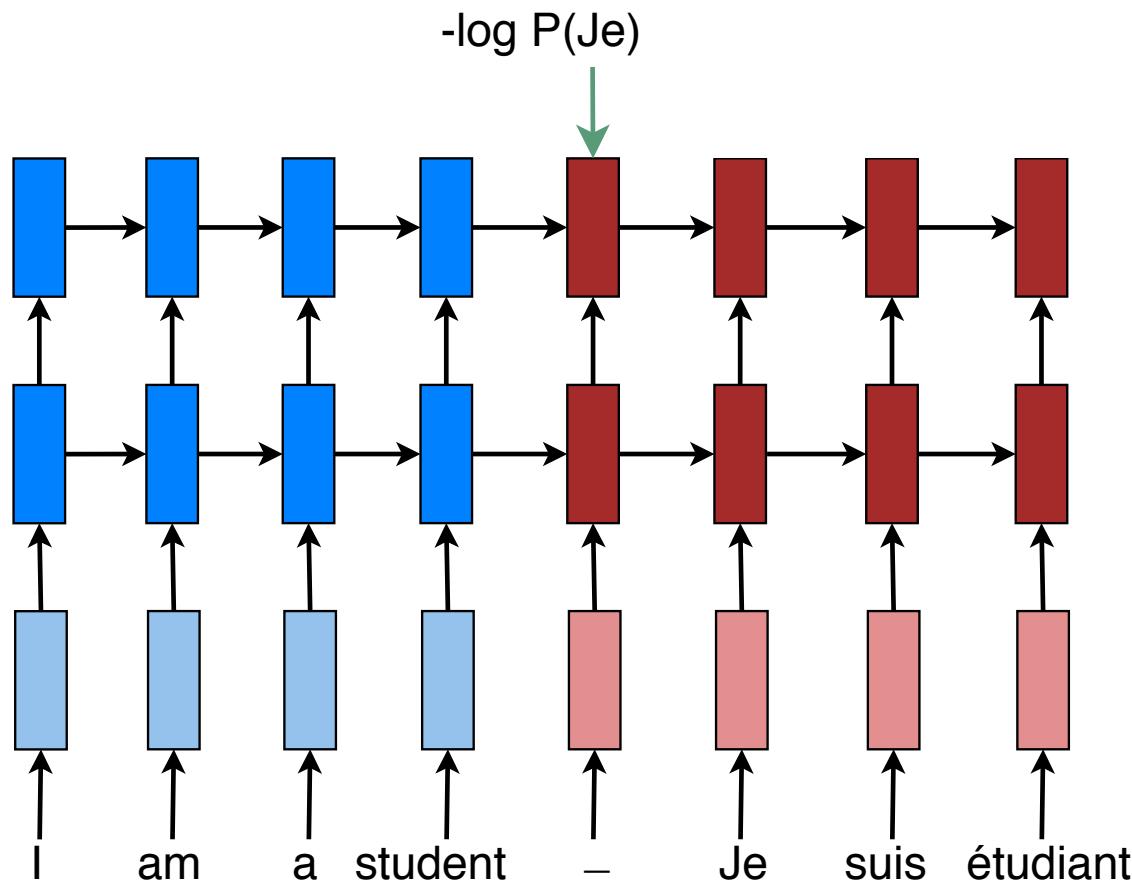
- Hidden states \mapsto scores \mapsto probabilities.

Training Loss



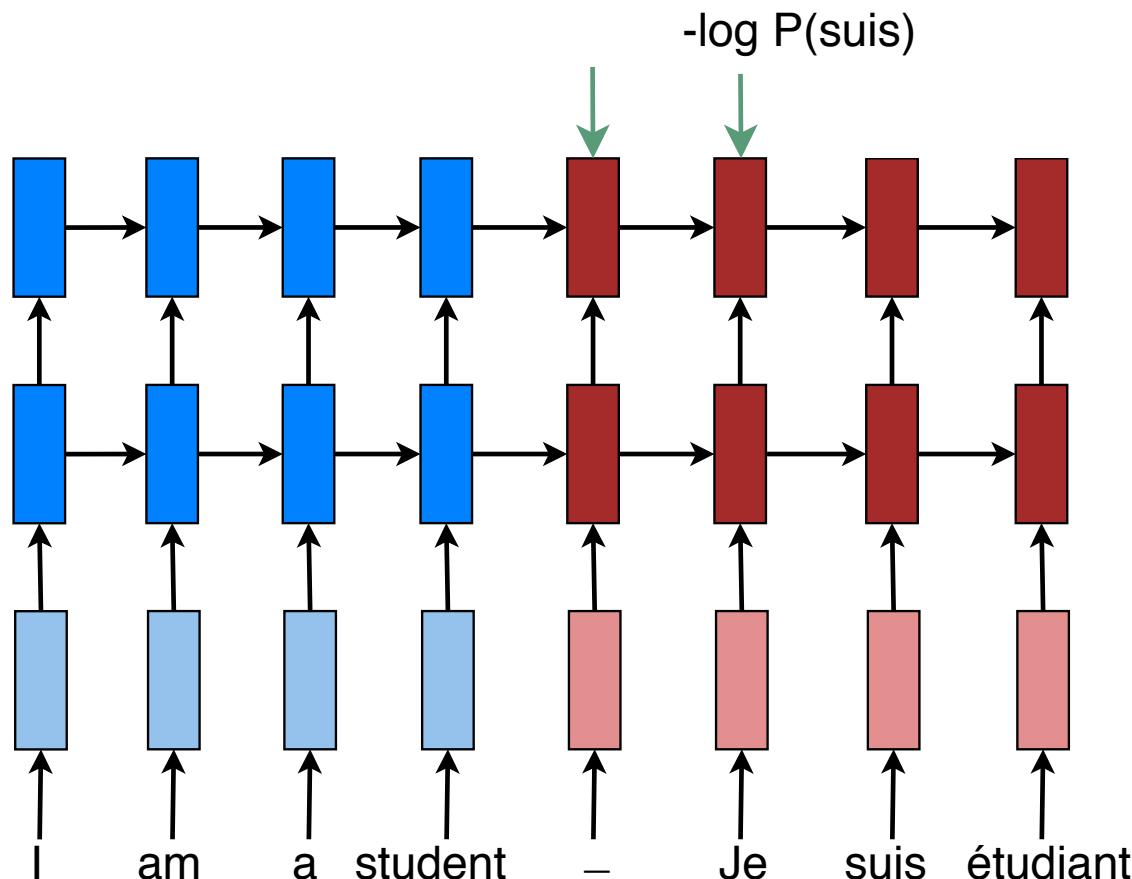
- Maximize $P(\text{target} \mid \text{source})$

Training Loss



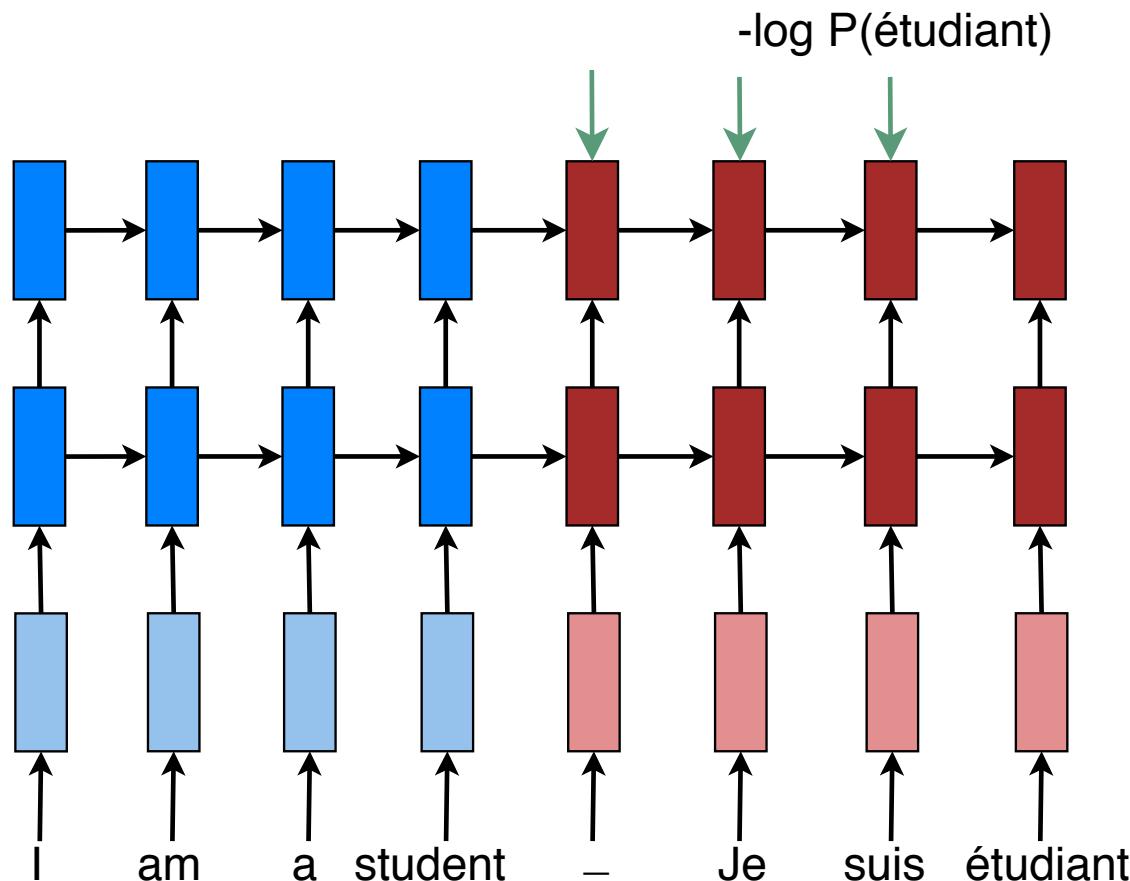
- Sum of all individual losses

Training Loss



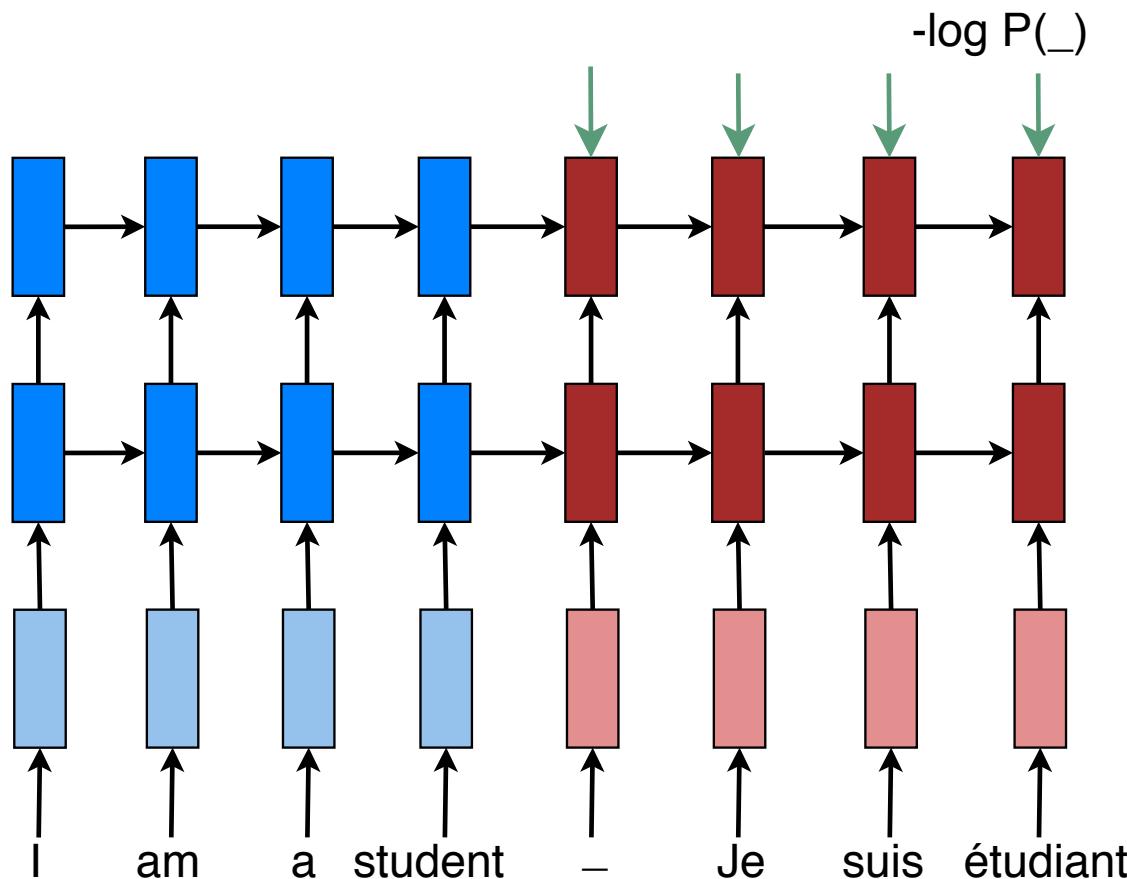
- Sum of all individual losses

Training Loss



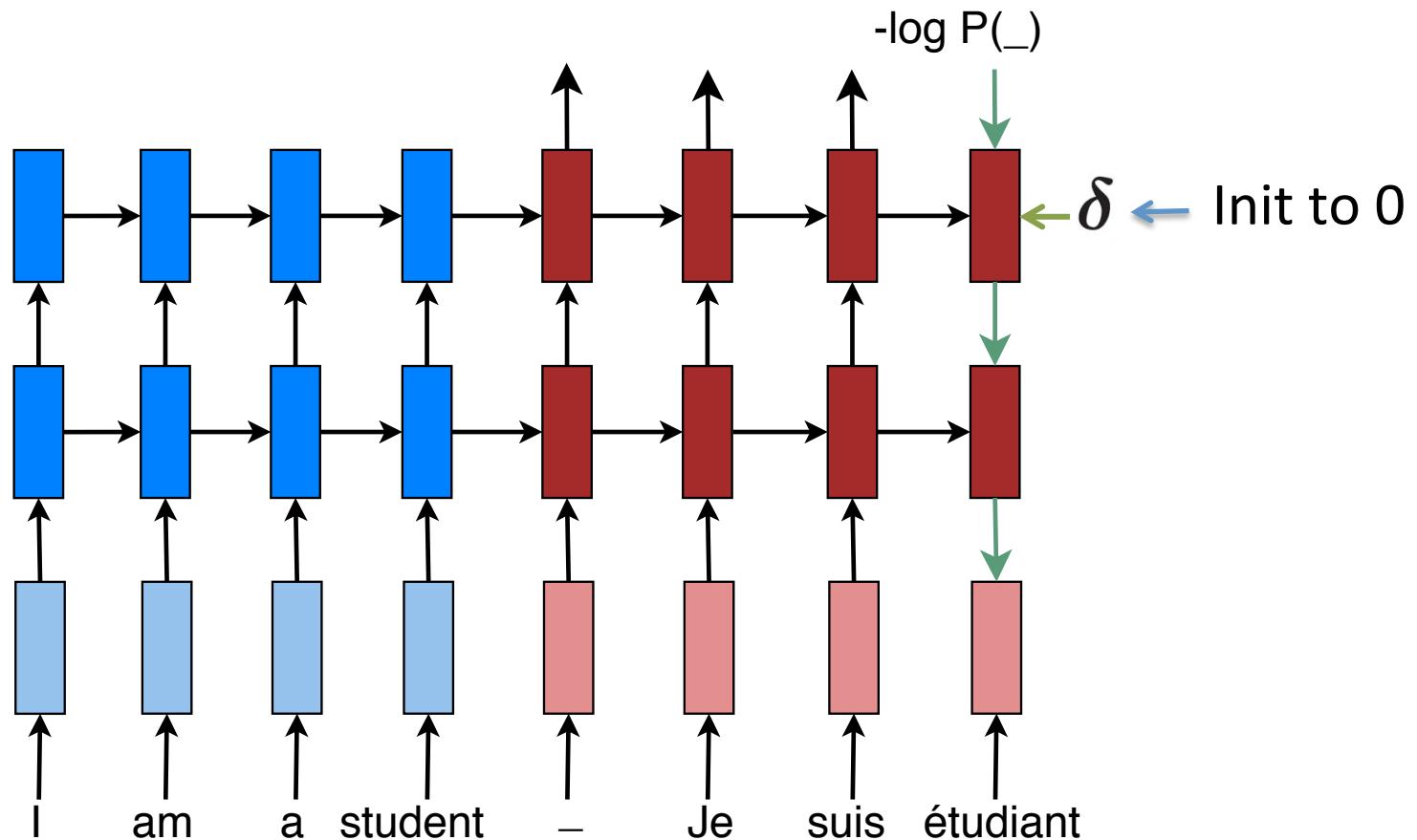
- Sum of all individual losses

Training Loss

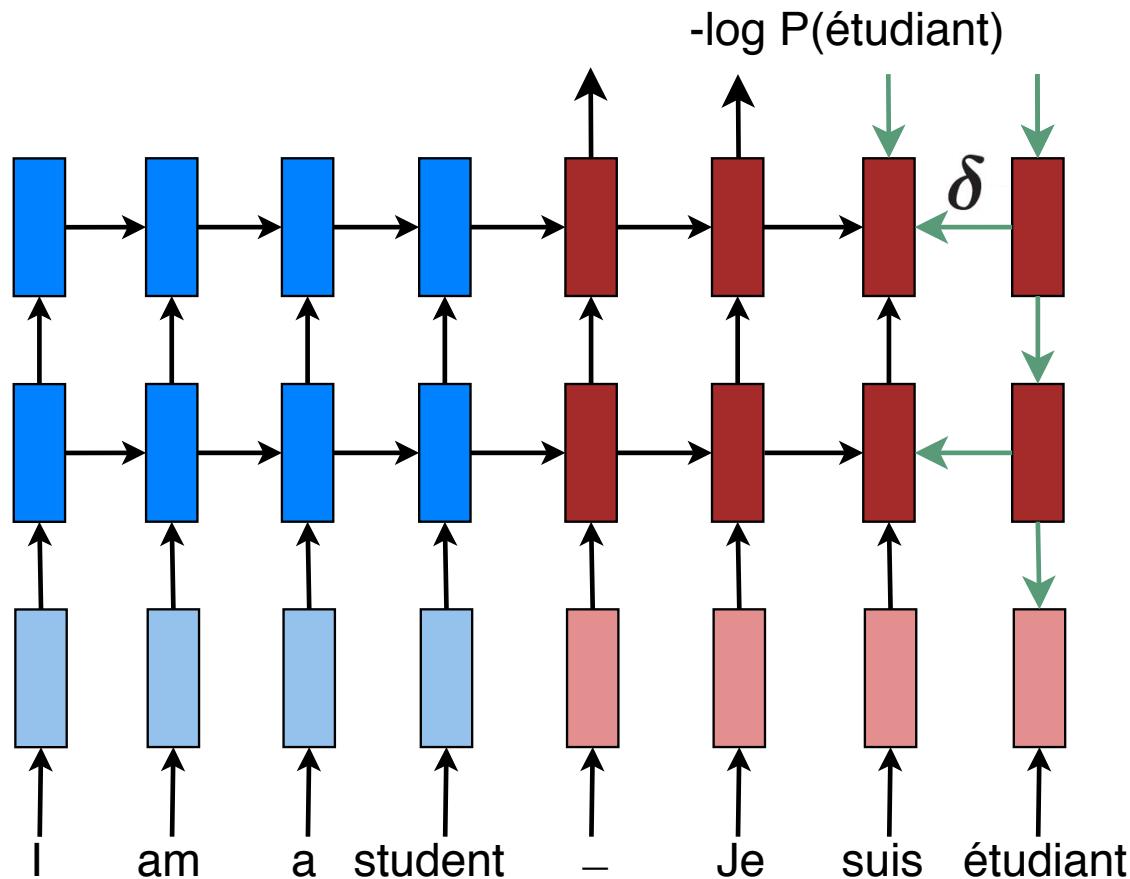


- Sum of all individual losses

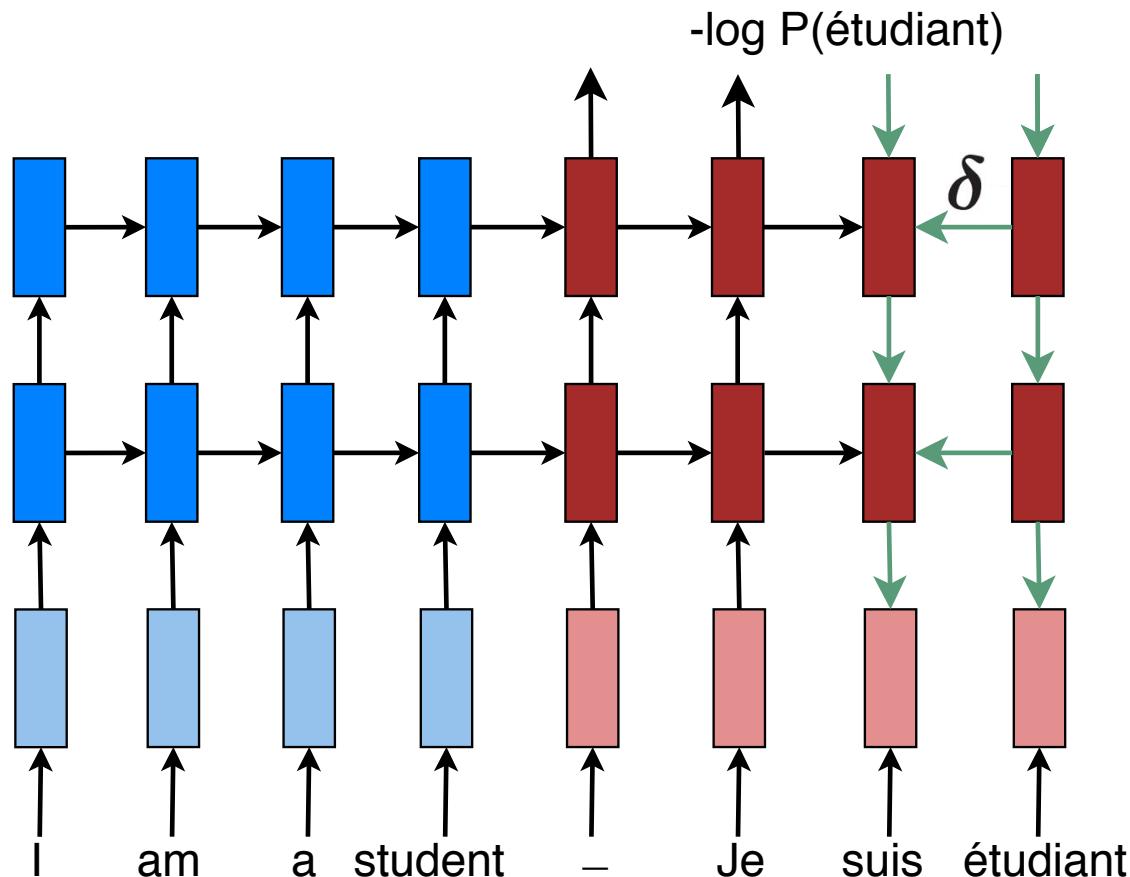
Backpropagation Through Time



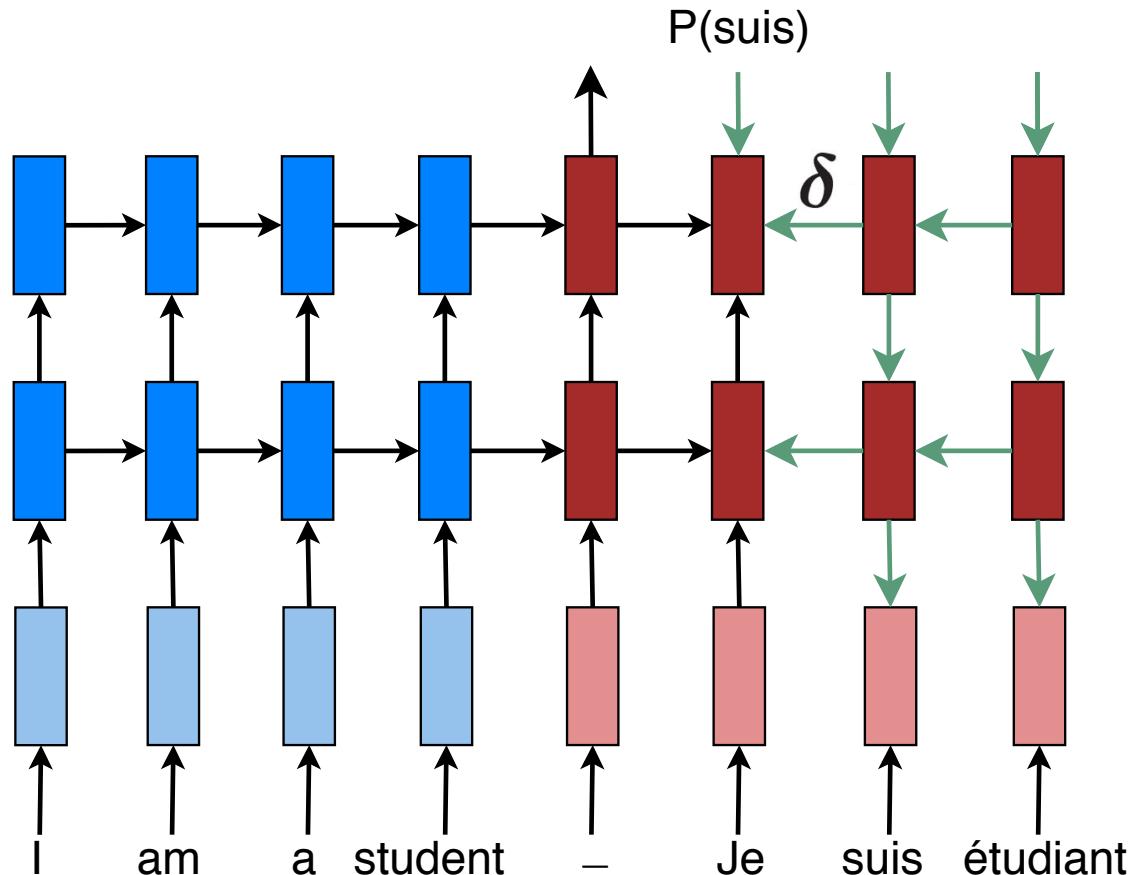
Backpropagation Through Time



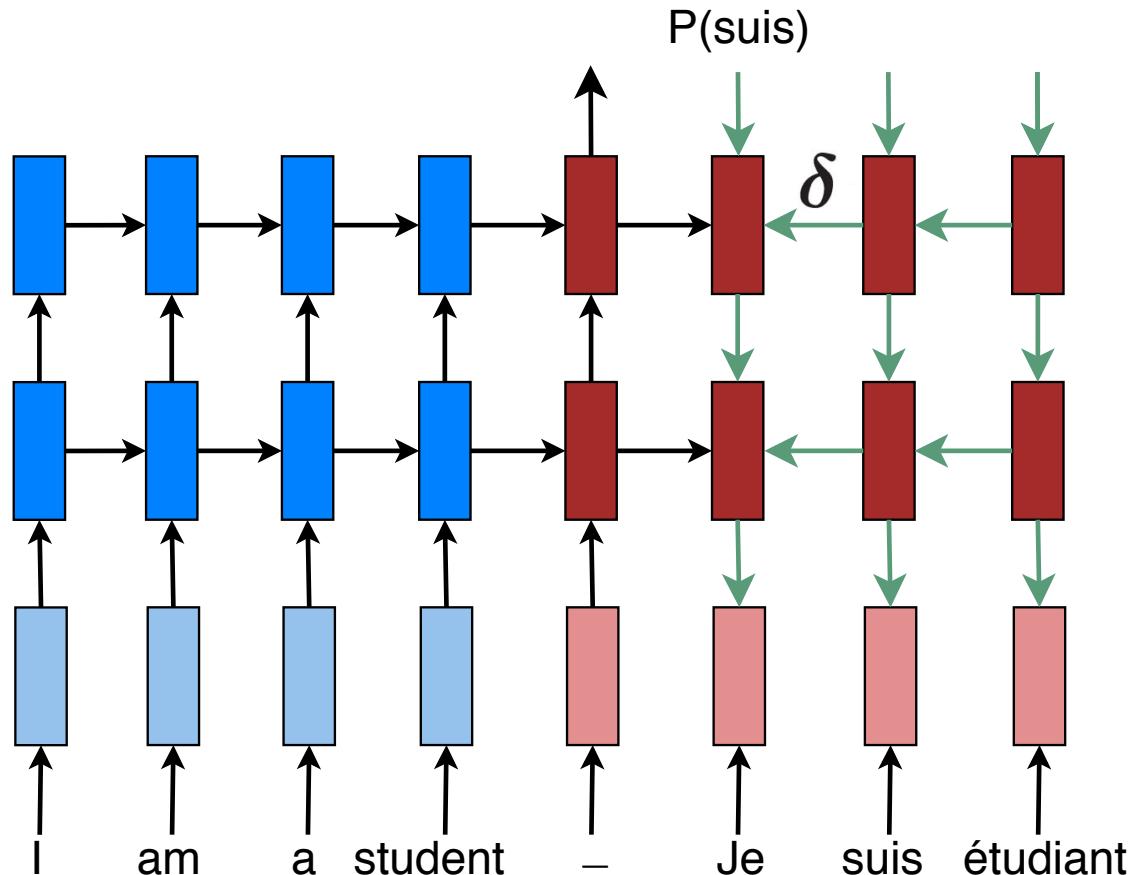
Backpropagation Through Time



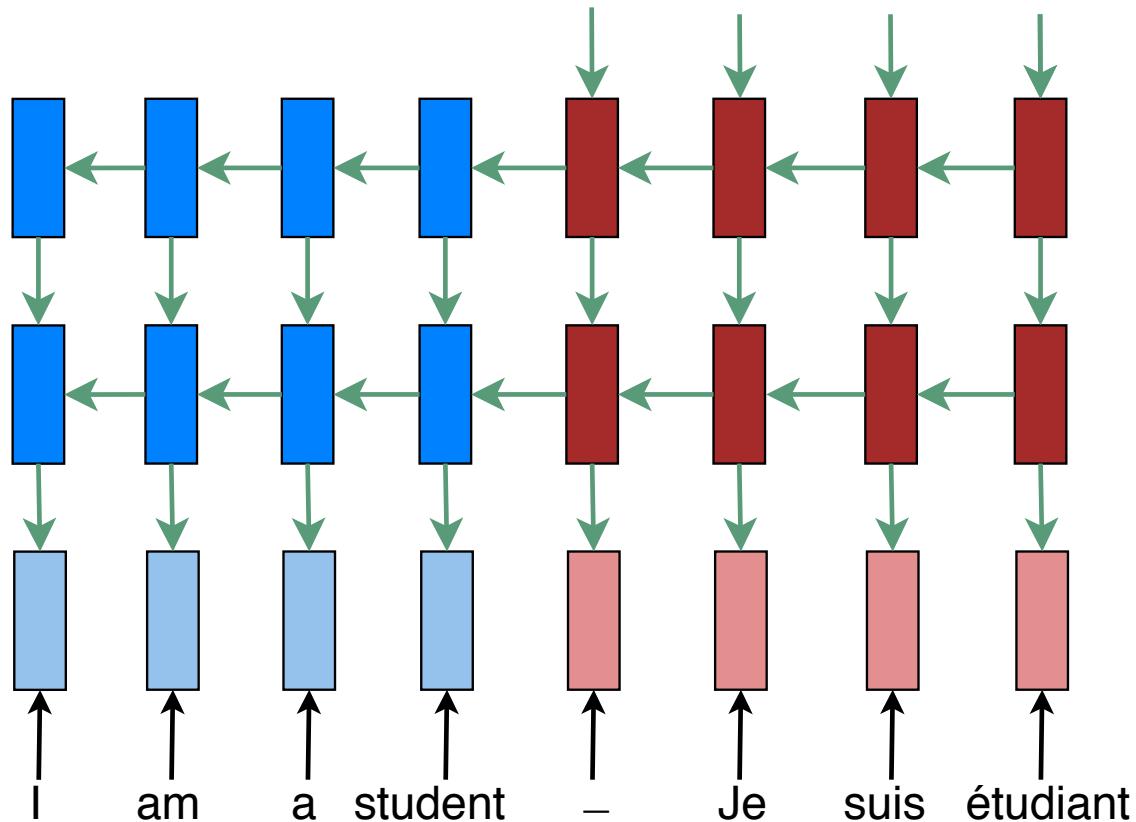
Backpropagation Through Time



Backpropagation Through Time



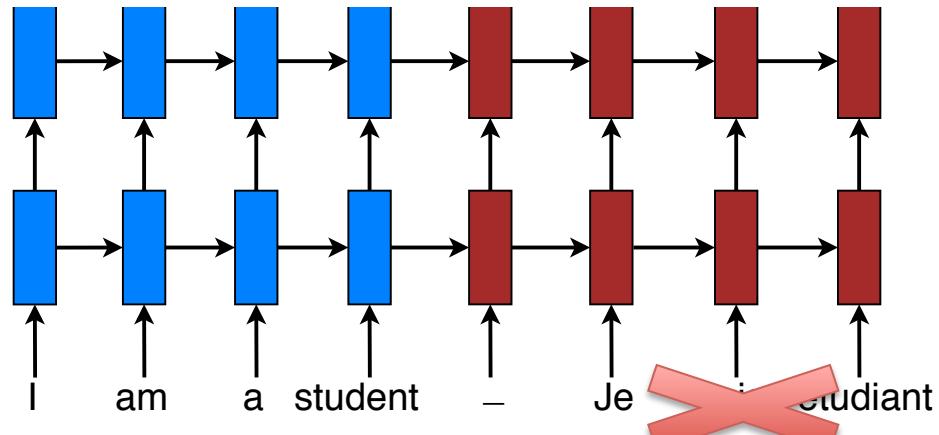
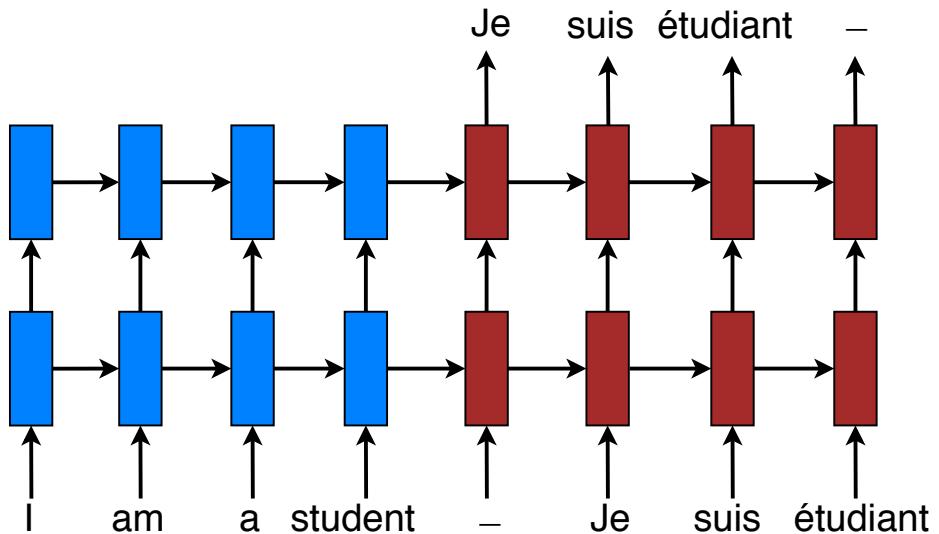
Backpropagation Through Time



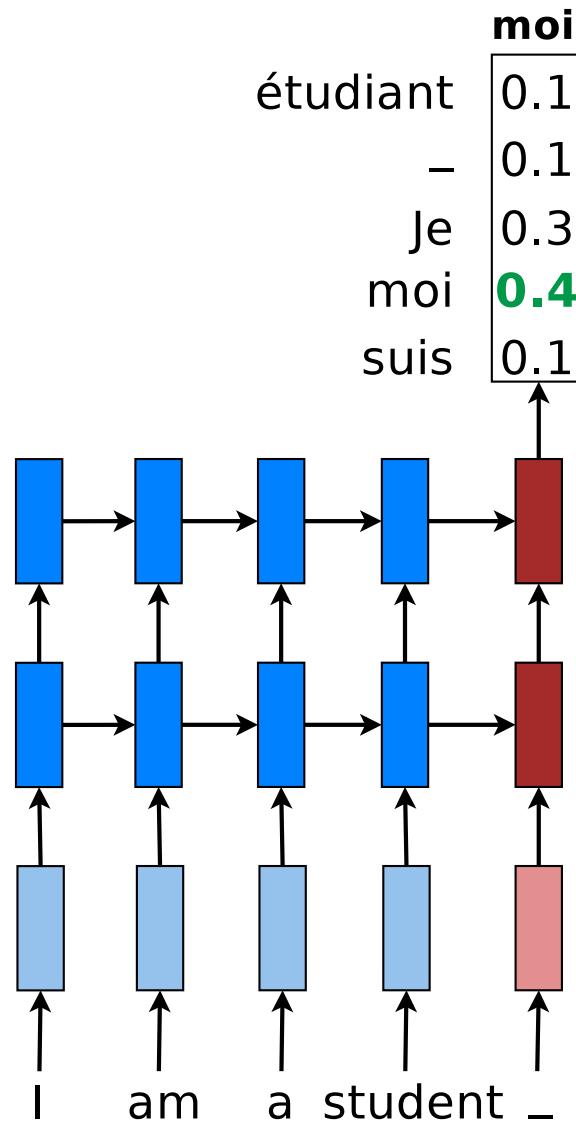
RNN gradients are accumulated.

Training vs. Testing

- *Training*
 - Correct translations are available.
- *Testing*
 - Only source sentences are given.

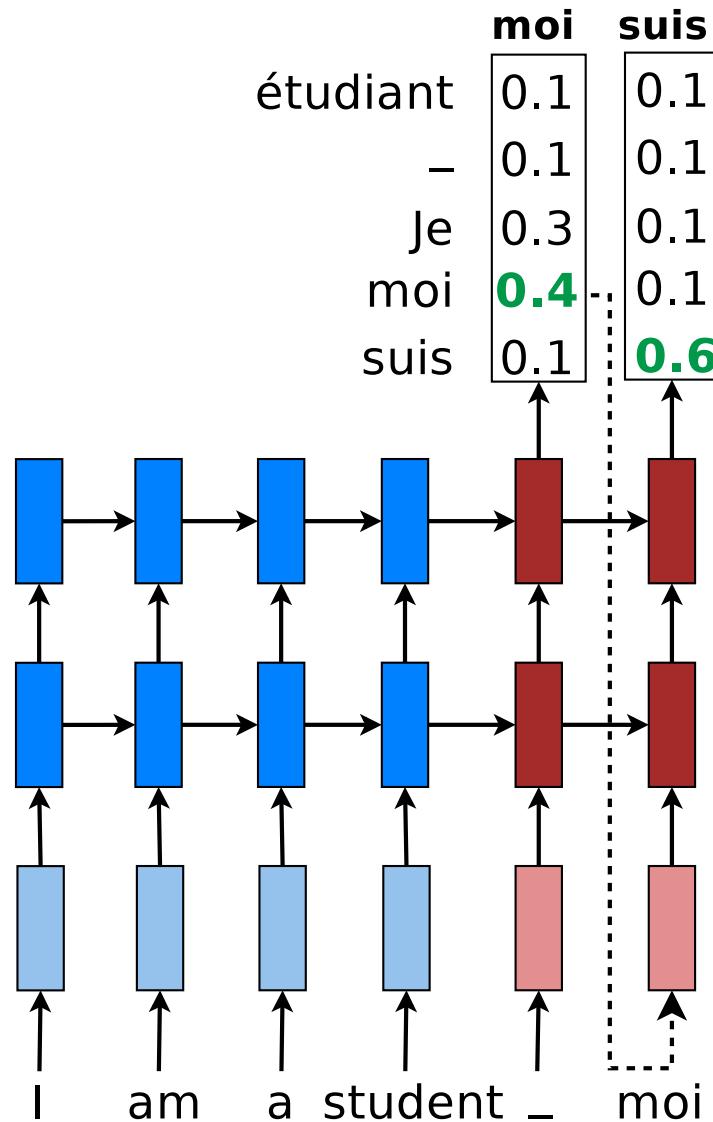


Testing



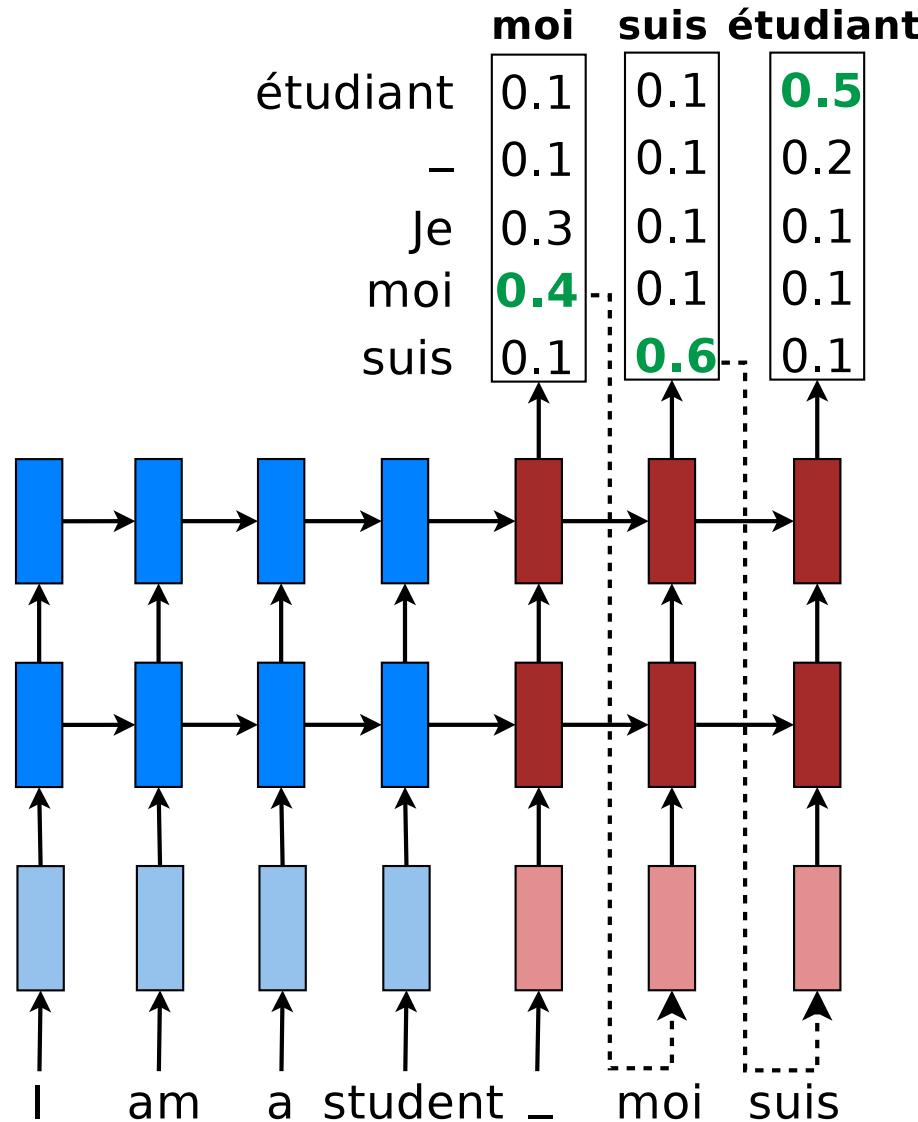
- Feed the **most likely** word

Testing



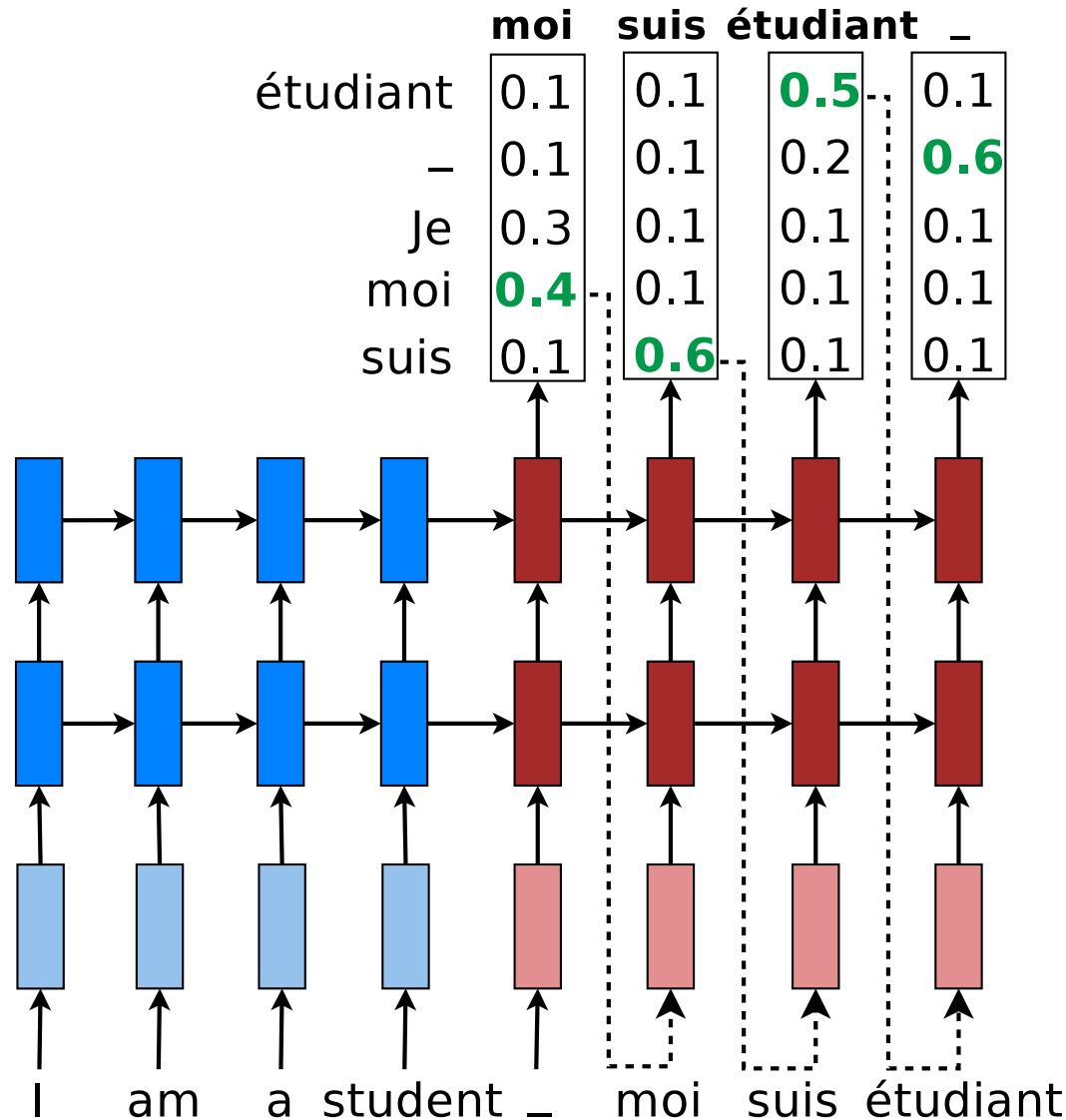
- Feed the **most likely** word

Testing



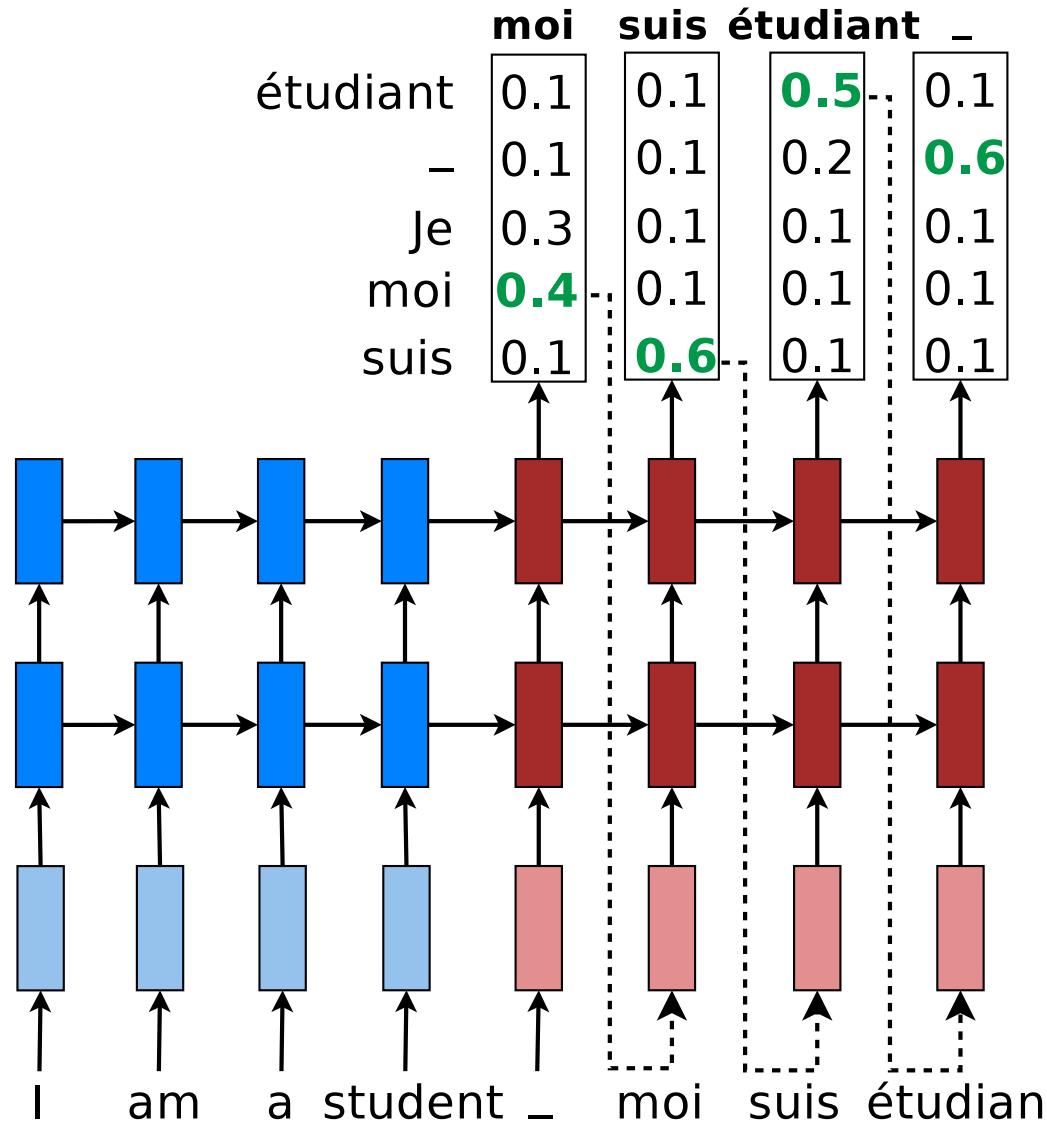
- Feed the **most likely** word

Testing



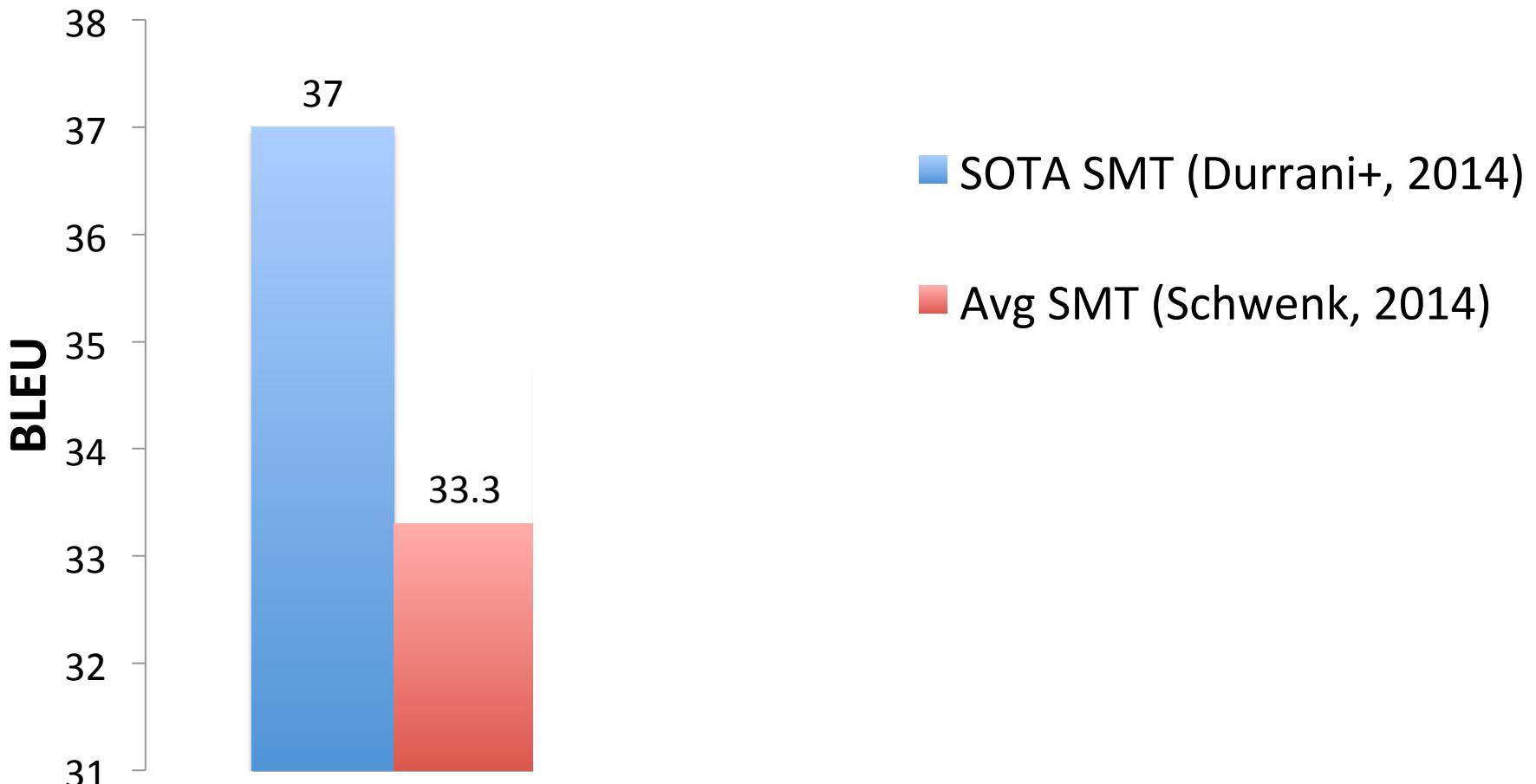
- Feed the **most likely word**

Testing

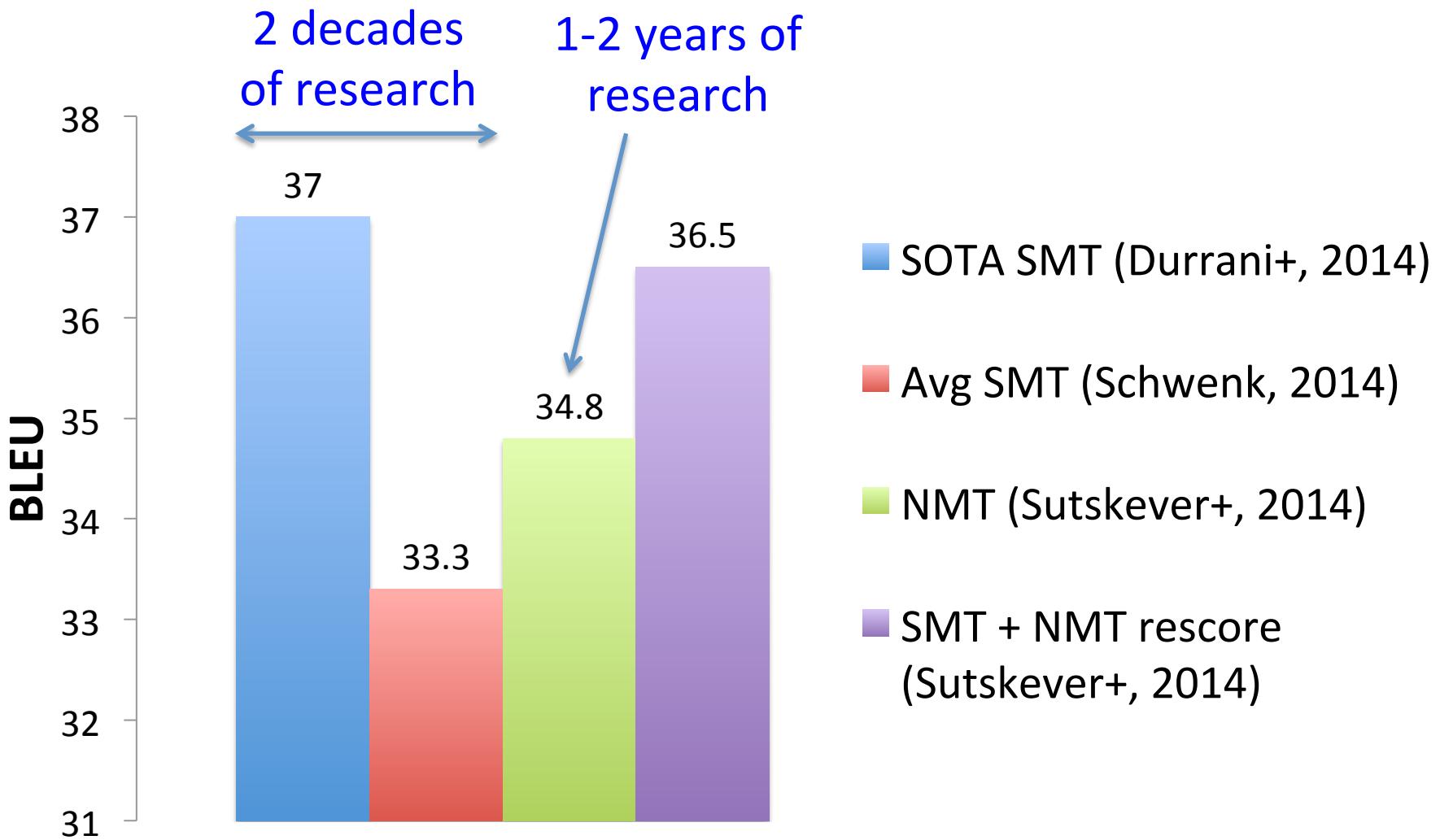


Simple beam-search decoders!

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 ...



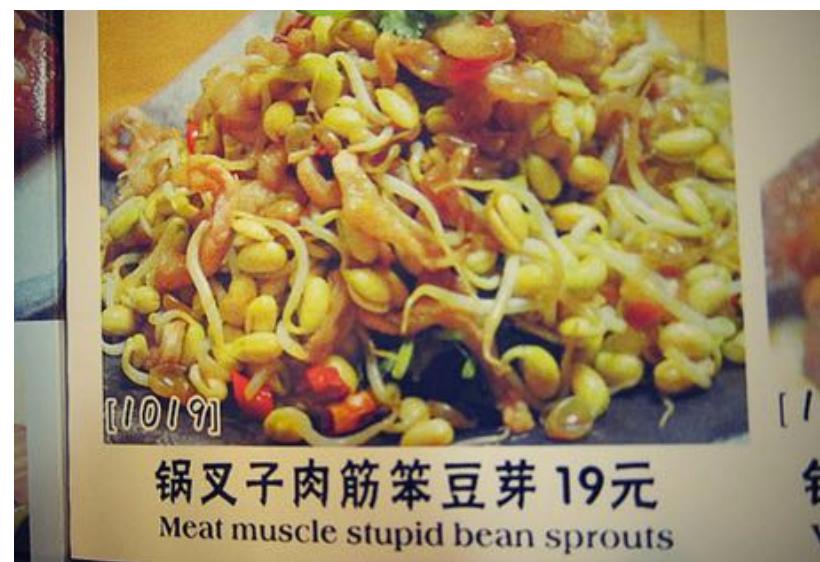
Sale of chicken murder



Go back toward your behind



5/19/16 Deep fried baby



Thang Luong - Neural Machine Translation
54
Meat muscle stupid bean sprouts

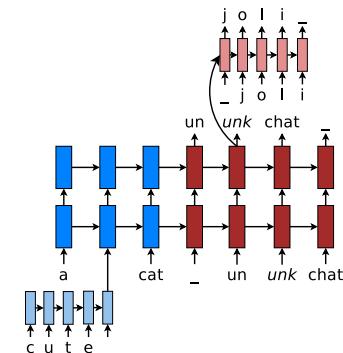
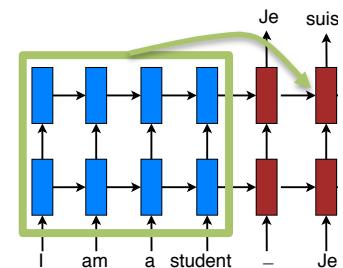
Limitations

- #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

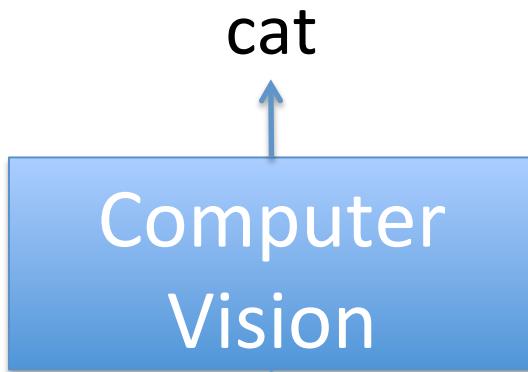
- #1: the *vocabulary size* problem
 - Sol: “copy” mechanism.
- #2: the *sentence length* problem
 - Sol: attention mechanism.
- #3: the *language complexity* problem
 - Sol: character-level translation.

The ~~ecotax~~ <unk> portico in ~~Pont-de-Buis~~
Le <unk> <unk> de <unk>
~~portique~~ ~~écotaxe~~ ~~Pont-de-Buis~~

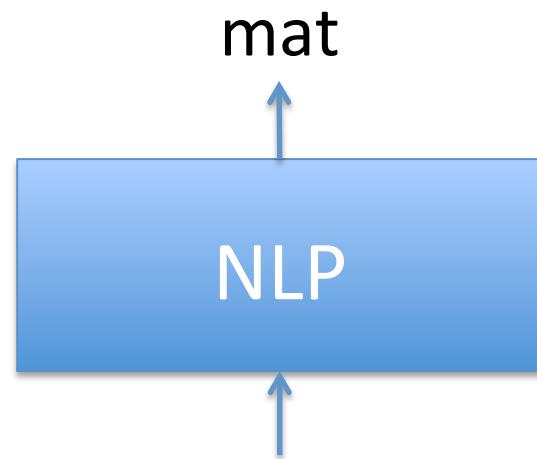


CV vs. NLP

1K categories



1M categories



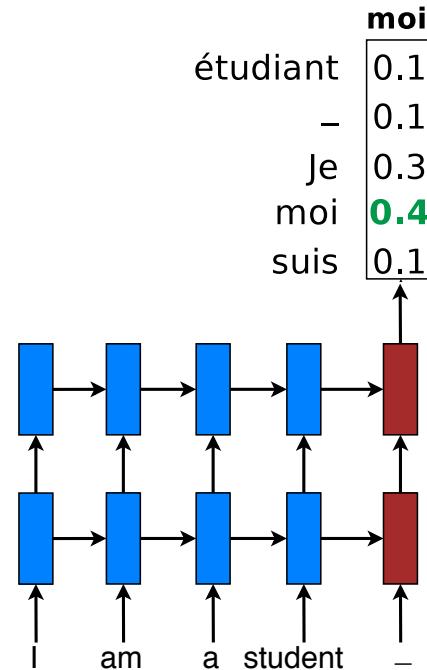
#1 The Vocabulary Size Problem

- 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.

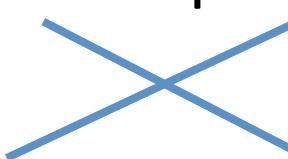
SOTA for English-French translation.

*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.
- Add relative positions.

The ecotax portico in Pont-de-Buis
Le portique écotaxe de Pont-de-Buis



The <unk> portico in <unk>
Le unk₁ unk₋₁ de unk₀



“Attention” for rare words

Our approach – *post-process*

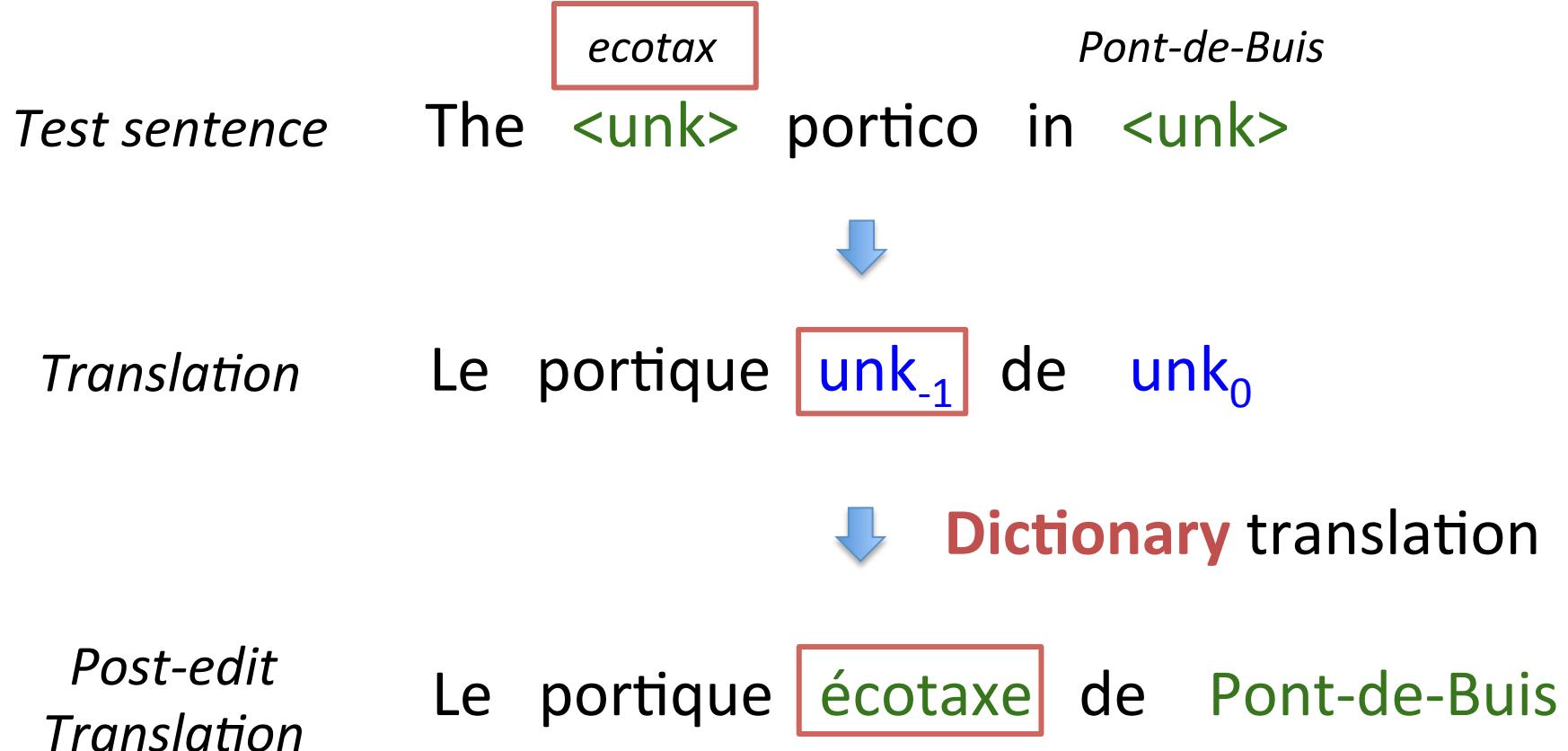
Test sentence The <unk> portico in <unk>

ecotax *Pont-de-Buis*

↓

Translation Le portique unk₁ de unk₀

Our approach – *post-process*



Our approach – *post-process*

Test sentence The <unk> portico in <unk>

ecotax

Pont-de-Buis



Translation Le portique unk₁ de unk₀

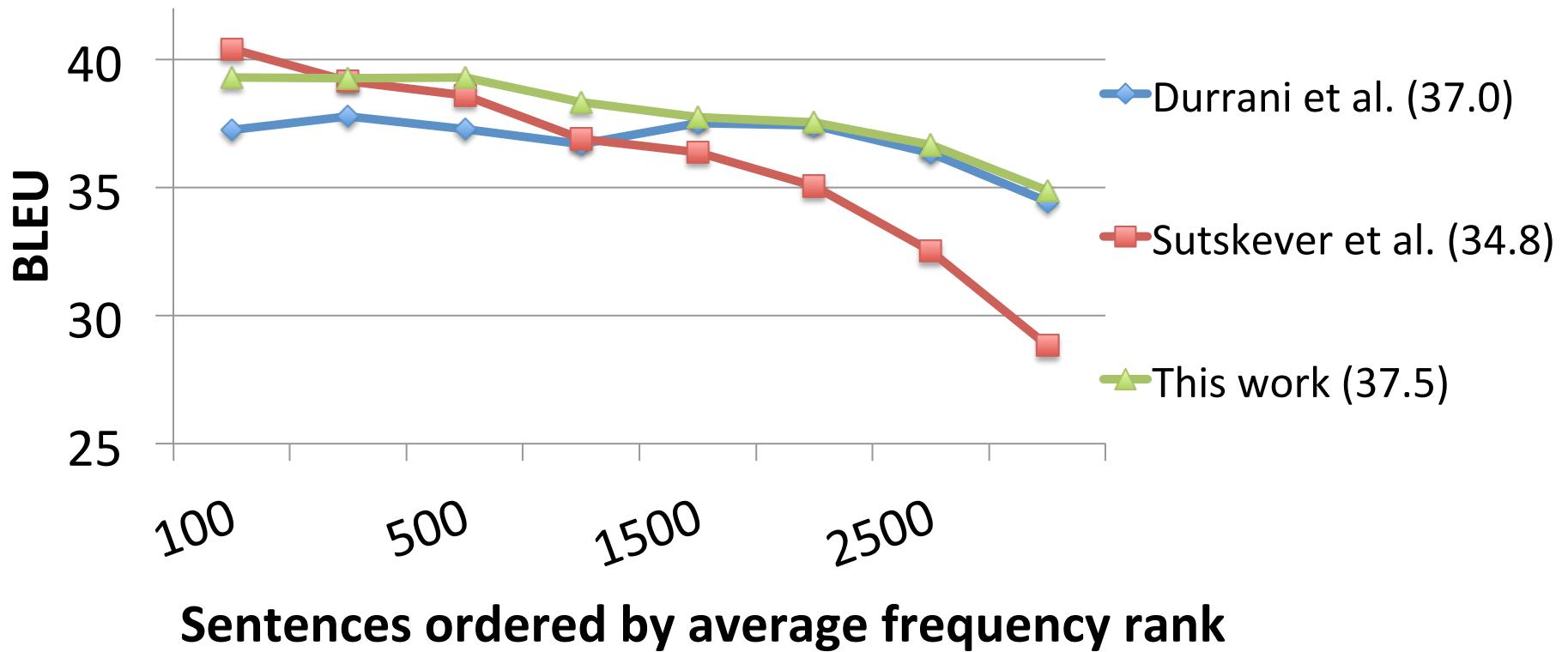


Identity copy

*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 cataract surgery will help clear a backlog . |
| human | 2600 opérations supplémentaires , notamment dans le domaine de la chirurgie orthopédique et de la cataracte , aideront à rattraper le retard . |
| trans | En outre , unk₁ opérations supplémentaires , dont la chirurgie unk₅ et la unk₆ , permettront de résorber l' arriéré . |
| trans +unk | En outre , 2600 opérations supplémentaires , dont la chirurgie orthopédiques et la 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 unk₀ , Richard unk₀ , a quitté unk₁ 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 unk₅ . |
| 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

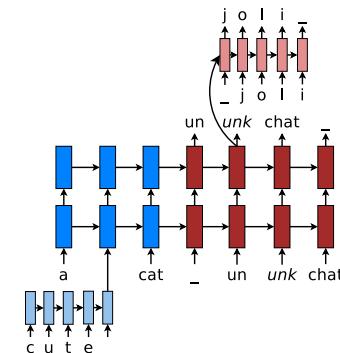
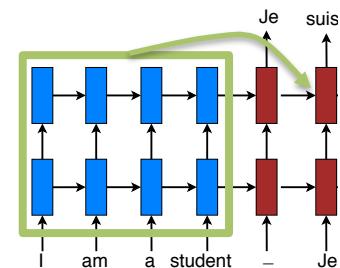
| | |
|------------|---|
| source | But concerns have grown after Mr Mazanga was quoted as saying Renamo was abandoning the 1992 peace accord . |
| human | Mais l' inquiétude a grandi après que M. Mazanga a déclaré que la Renamo abandonnait l' accord de paix de 1992 . |
| trans | Mais les inquiétudes se sont accrues après que M. unkpos₃ a déclaré que la unk₃ unk₃ 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**.

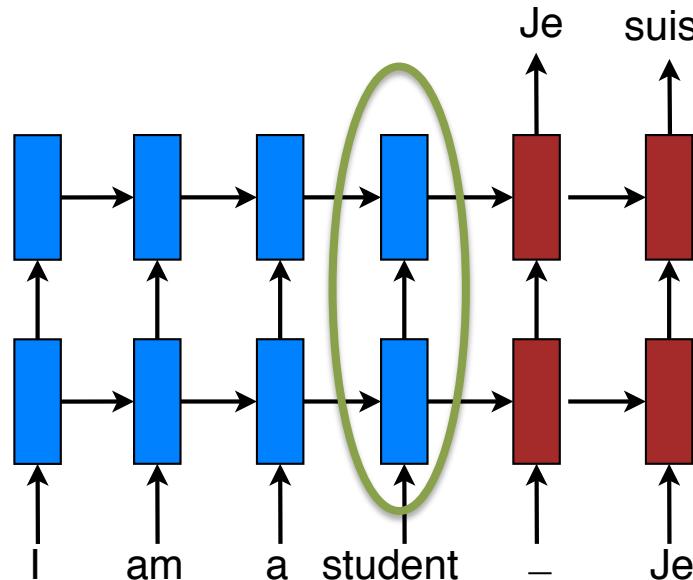
Advancing NMT

- #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.

The ecotax <unk> portico in Pont-de-Buis
Le <unk> <unk> de <unk>
portique écotaxe Pont-de-Buis



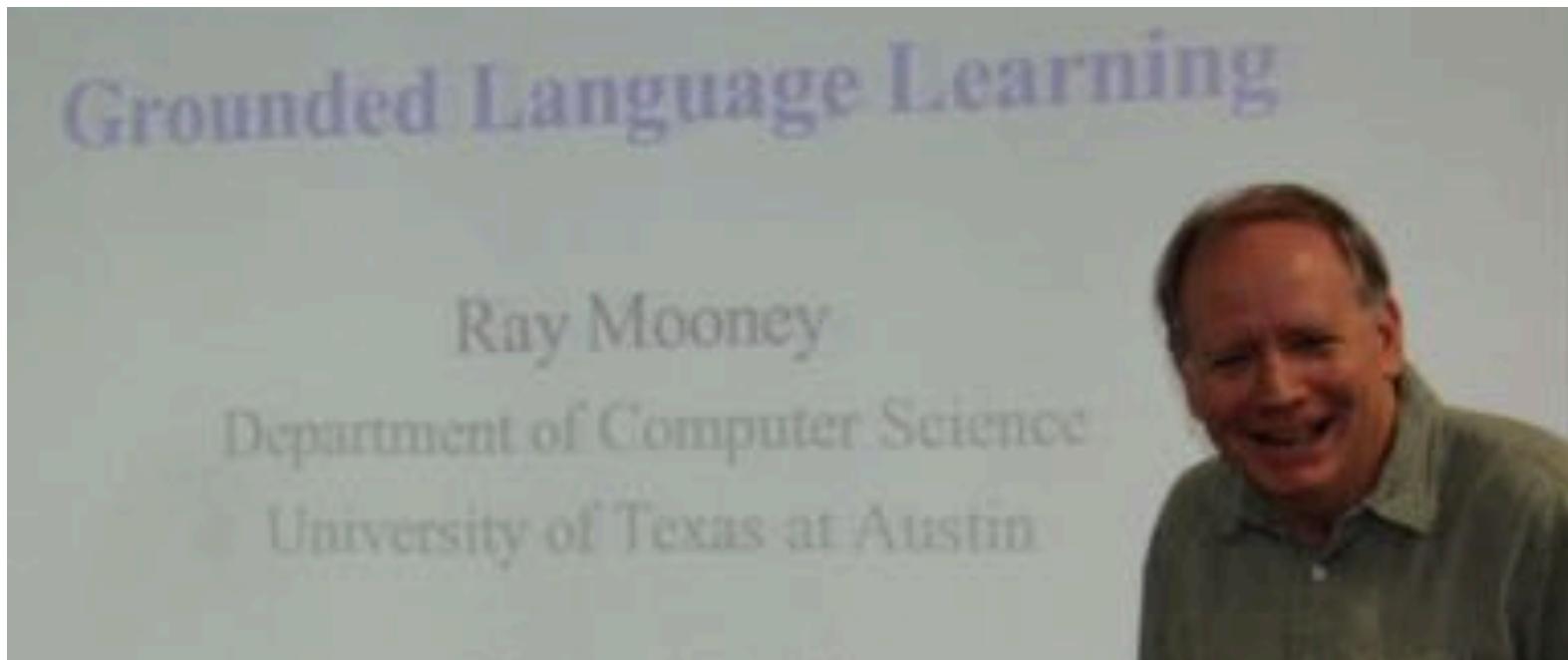
#2 The Sentence Length Problem



- Translation quality degrades with long sentences.

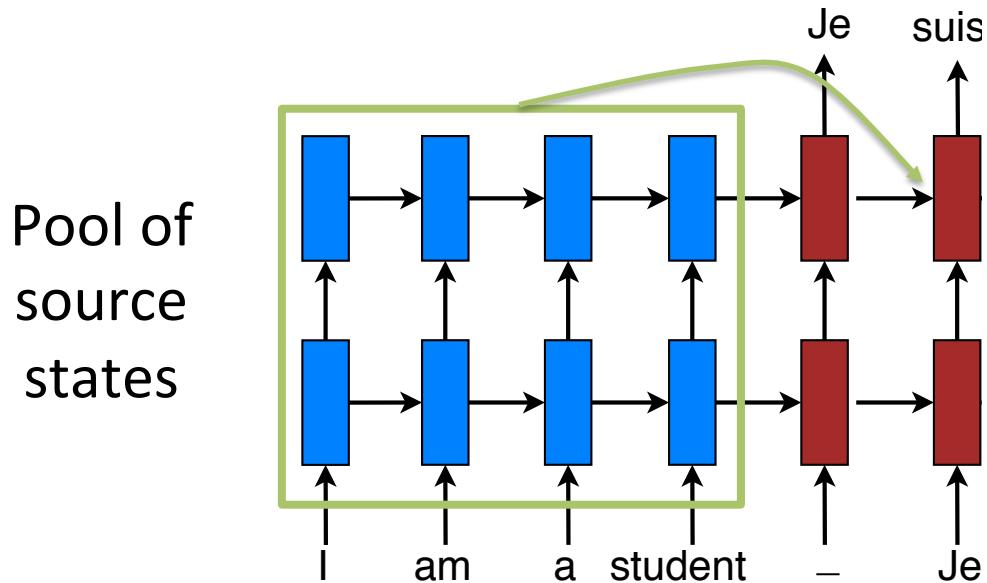
Problem: sentence meaning is represented by a fixed-dimensional vector.

You can't cram the meaning of a whole
sentence into a single vector!

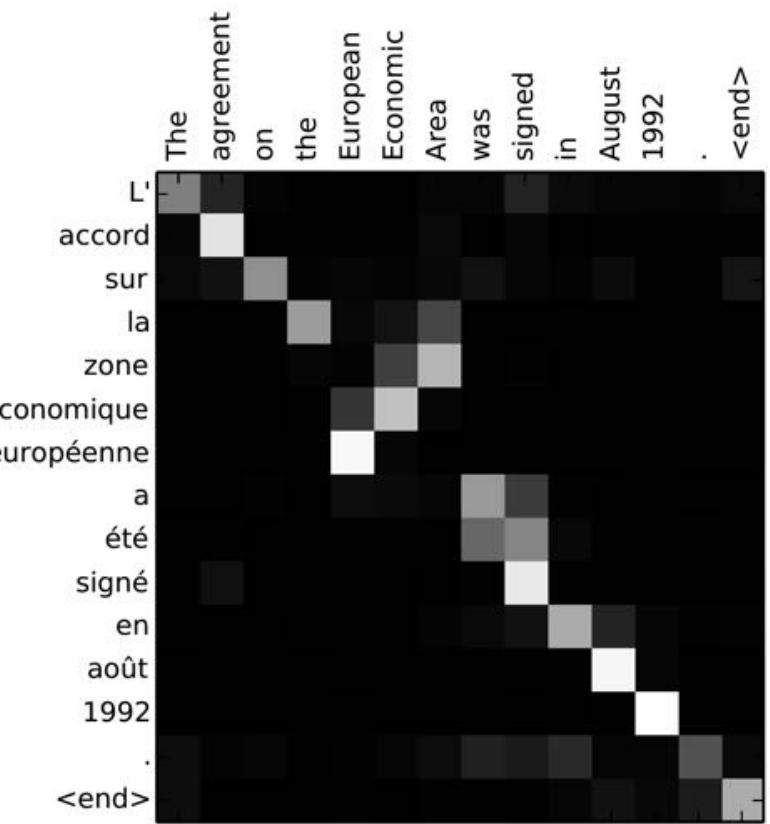
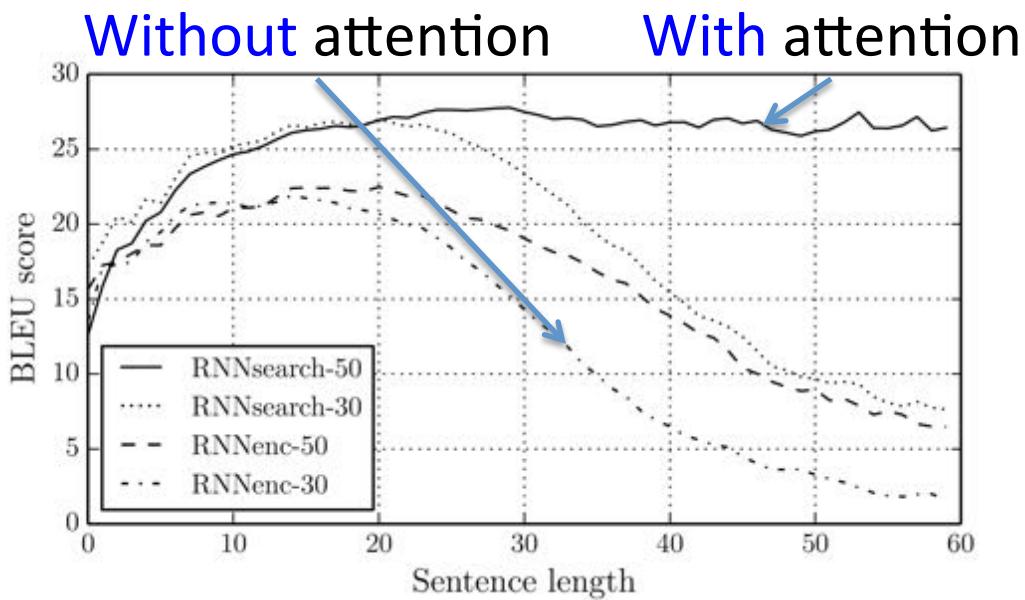


(Adapted from KyungHuyn Cho' talk)

Attention Mechanism

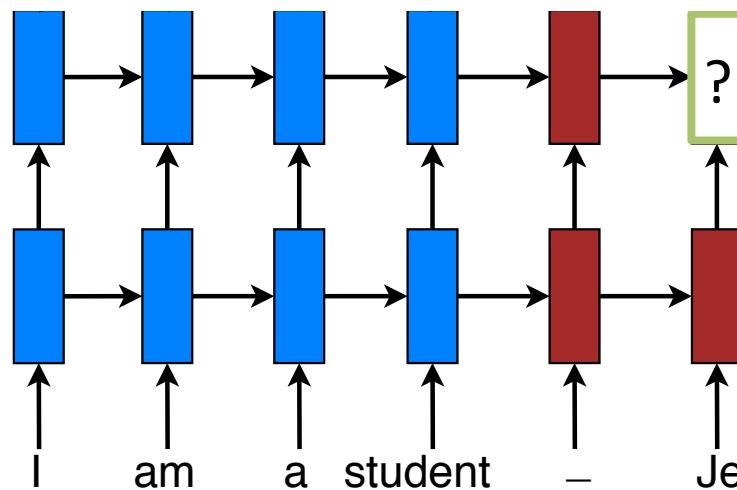


- **Solution:** random access memory
 - Retrieve as needed.



Dzmitry Bahdanau, Kyung-Hyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Translate and Align. ICLR 2015.

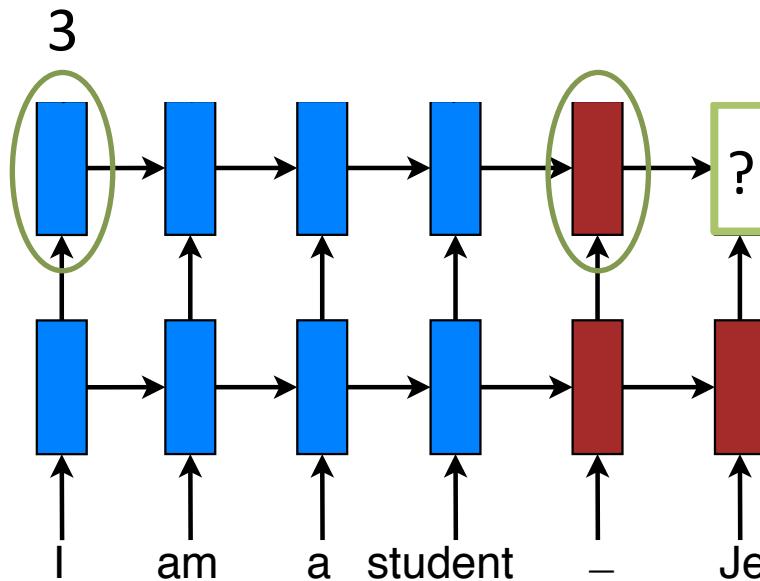
Attention Mechanism



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

Attention Mechanism – *Scoring*

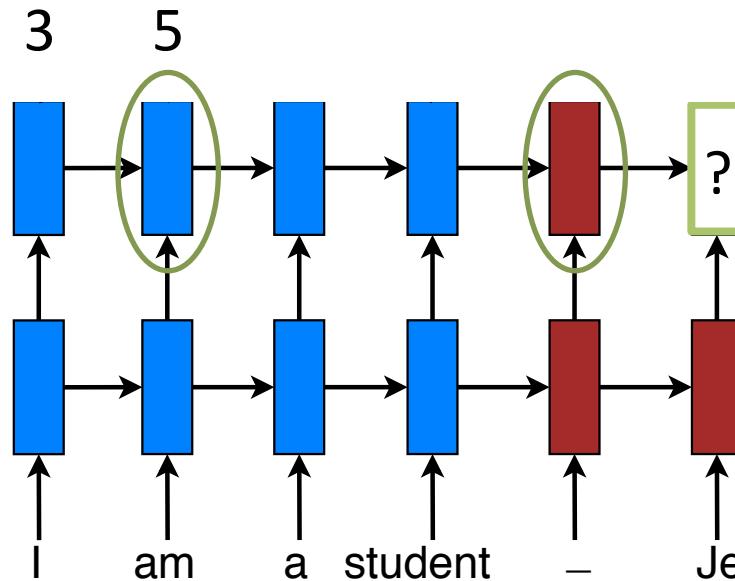
$$\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

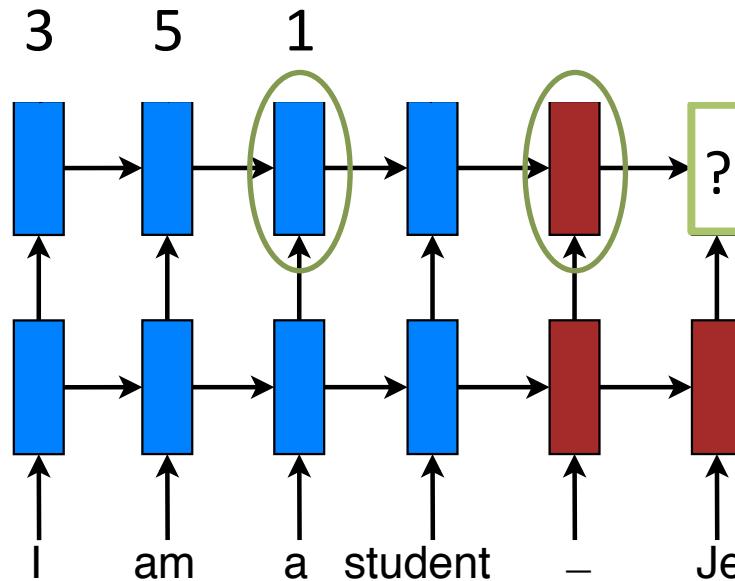
$$\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s)$$



- Compare target and source hidden states.

Attention Mechanism – *Scoring*

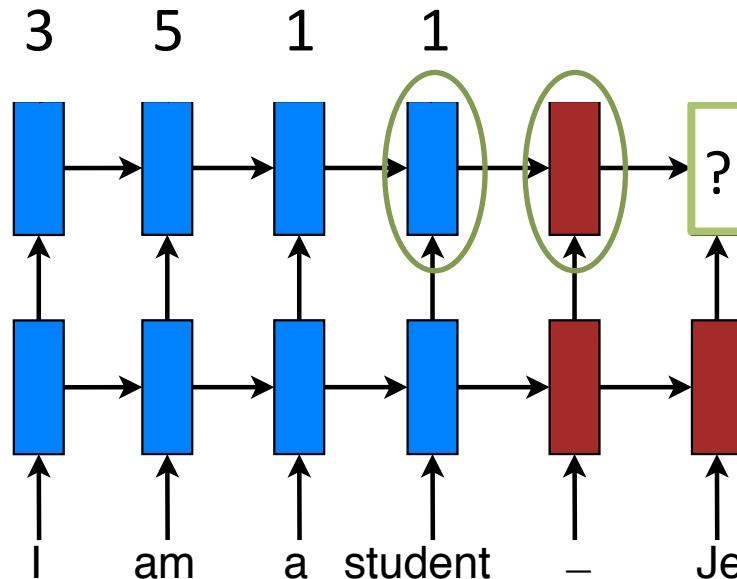
$$\text{score}(\mathbf{h}_t, \bar{\mathbf{h}}_s)$$



- Compare target and source hidden states.

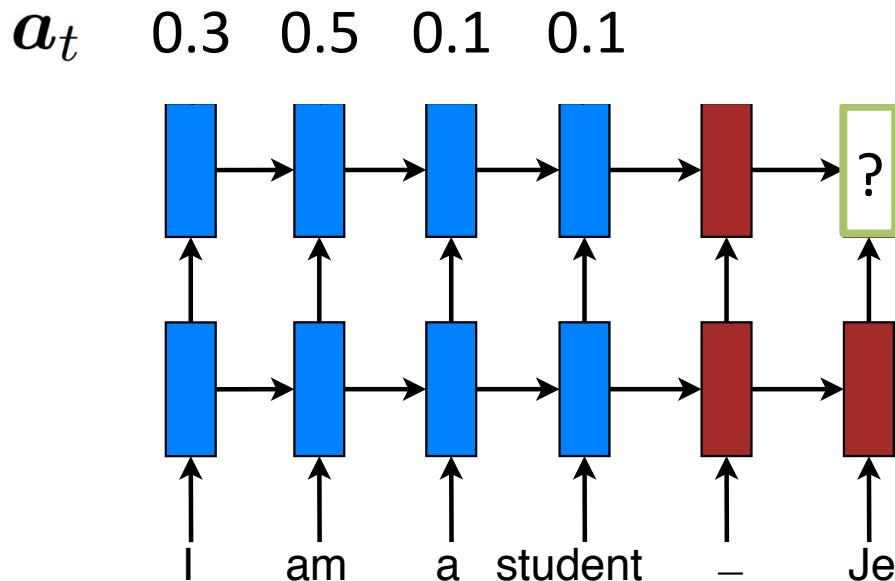
Attention Mechanism – *Scoring*

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



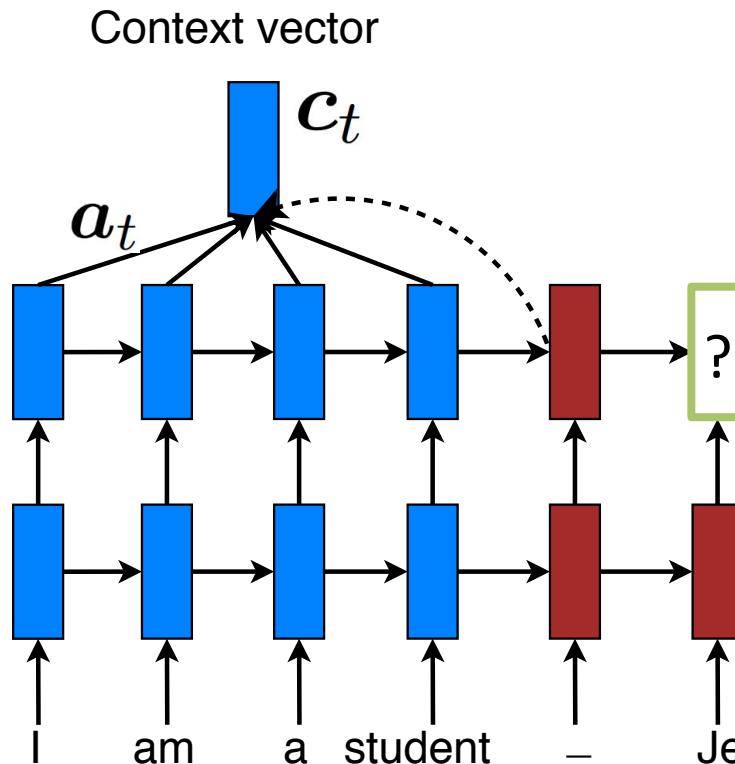
- Compare target and source hidden states.

Attention Mechanism – Normalization



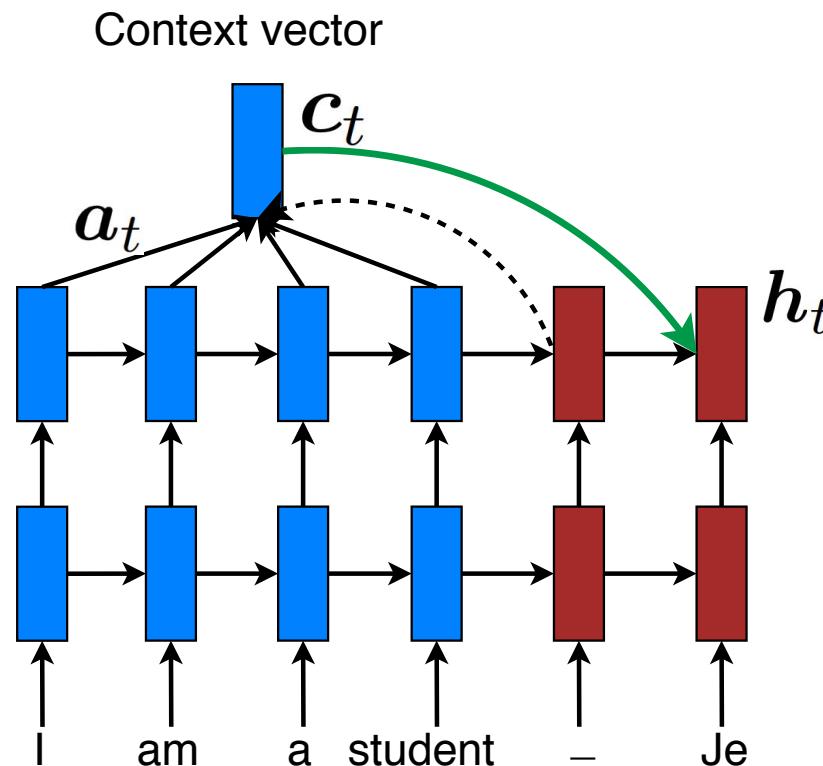
- Convert into alignment weights.

Attention Mechanism – *Context vector*



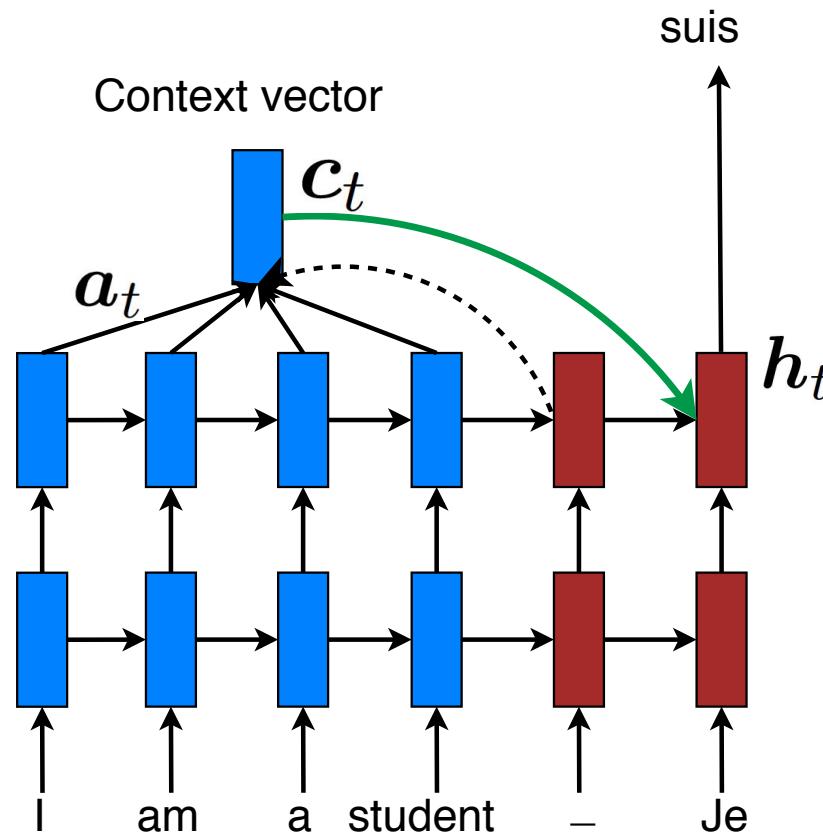
- Build **context** vector: weighted average.

Attention Mechanism – *Hidden state*



- Compute the next hidden state.

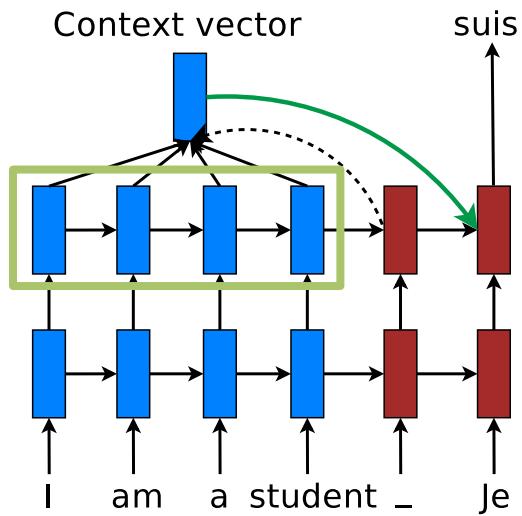
Attention Mechanism – Predict



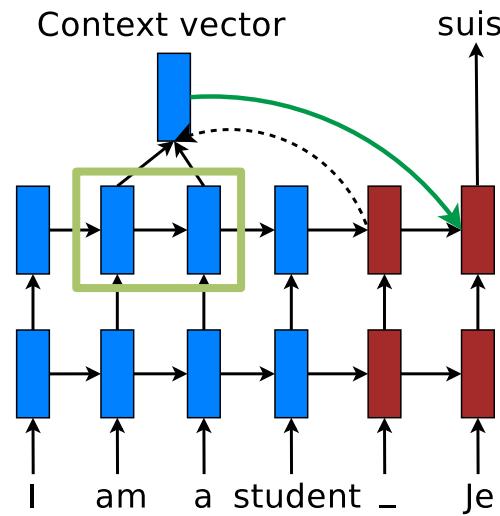
- Predict the **next word**.



- Examine various attention mechanisms:



Global: *all* source states.

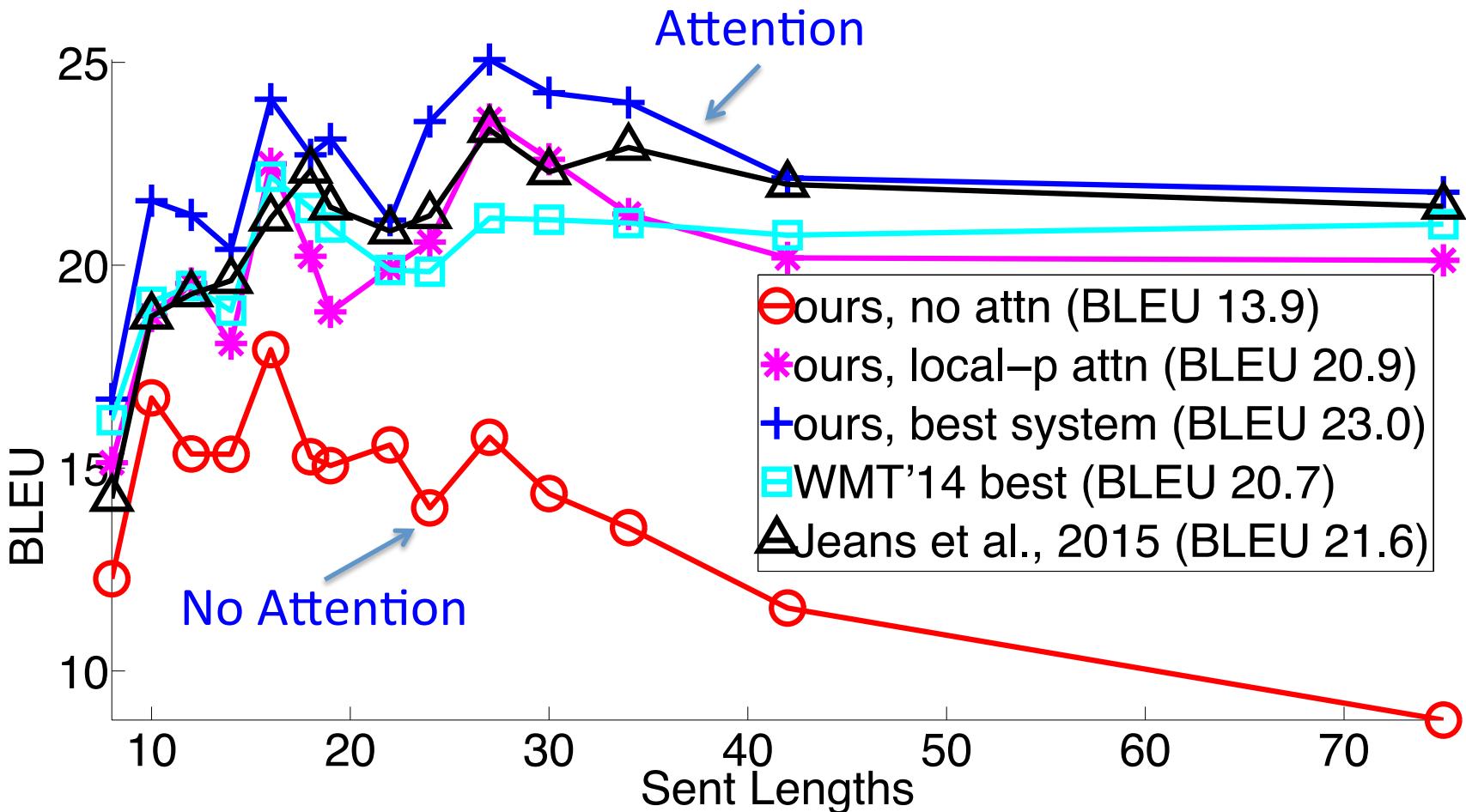


Local: *subset* of source states.

SOTA for English-German translation.

Thang Luong, 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 <i>Miranda Kerr</i> 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 über die <unk> , dass ein <unk> <unk> mit Sicherheit nicht vereinbar ist mit Sicherheit und Sicherheit , sagte Roger Cameron , CEO der US - <unk> . |

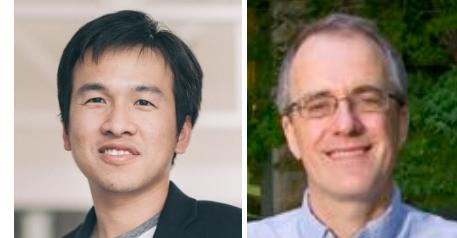
- Translate a **doubly-negated phrase** correctly

Sample English-German translations

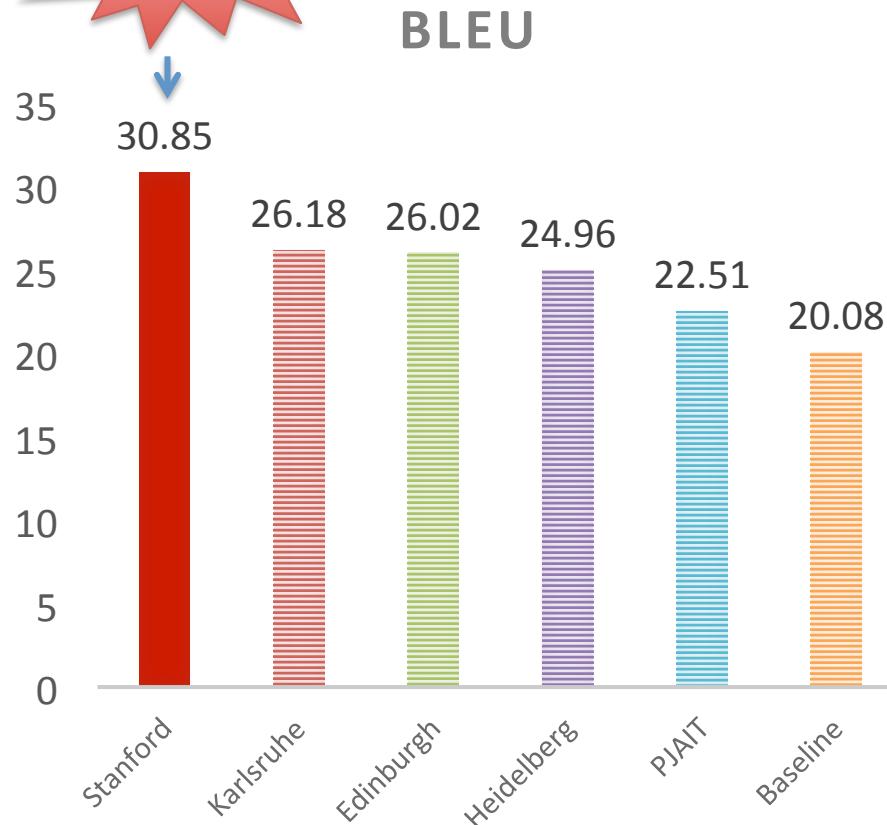
| | |
|-------------|--|
| source | We're pleased the FAA recognizes that an enjoyable passenger experience is not incompatible with safety and security , said Roger Dow , CEO of the U.S. Travel Association . |
| human | Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht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

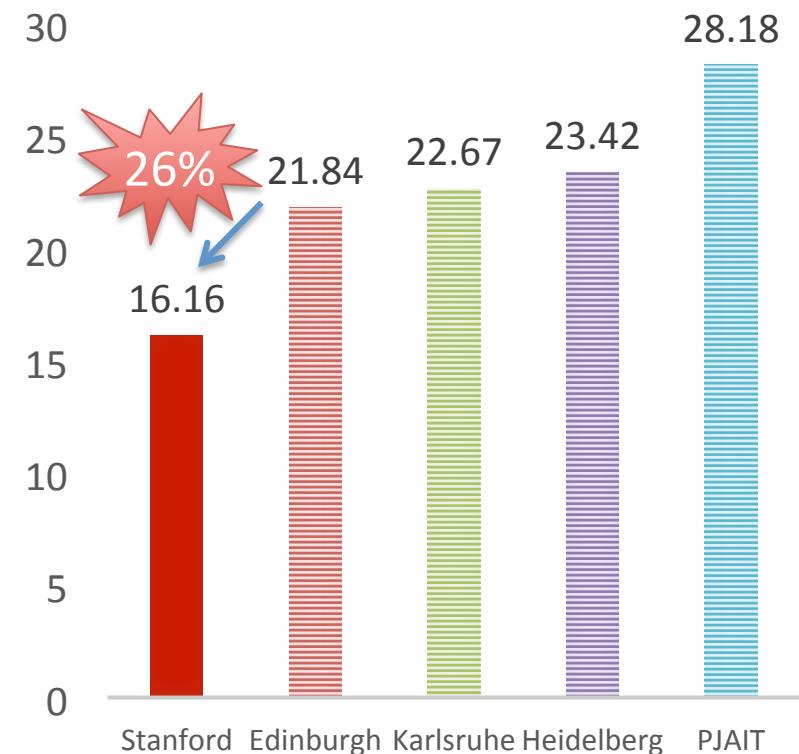
TED talk, English-German



Winning



HUMAN TER

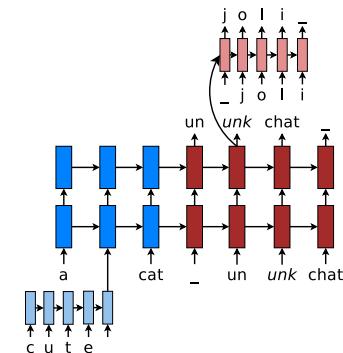
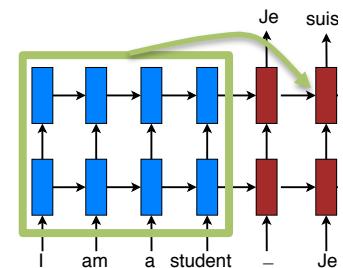


Thang Luong and Chris Manning. Stanford Neural Machine Translation Systems for Spoken Language Domain. IWSLT 2015.

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.

The ~~ecotax~~ <unk> portico in ~~Pont-de-Buis~~
Le <unk> <unk> de <unk>
~~portique~~ ~~écotaxe~~ ~~Pont-de-Buis~~



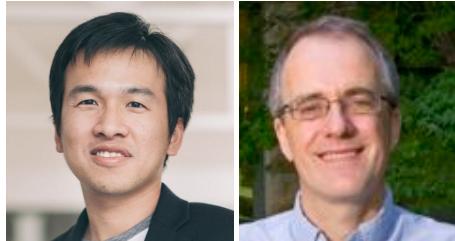
#3 The rare word problem

- “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*: *gooooooood 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).



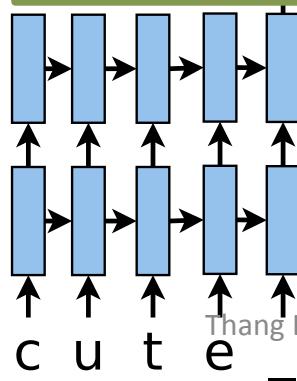
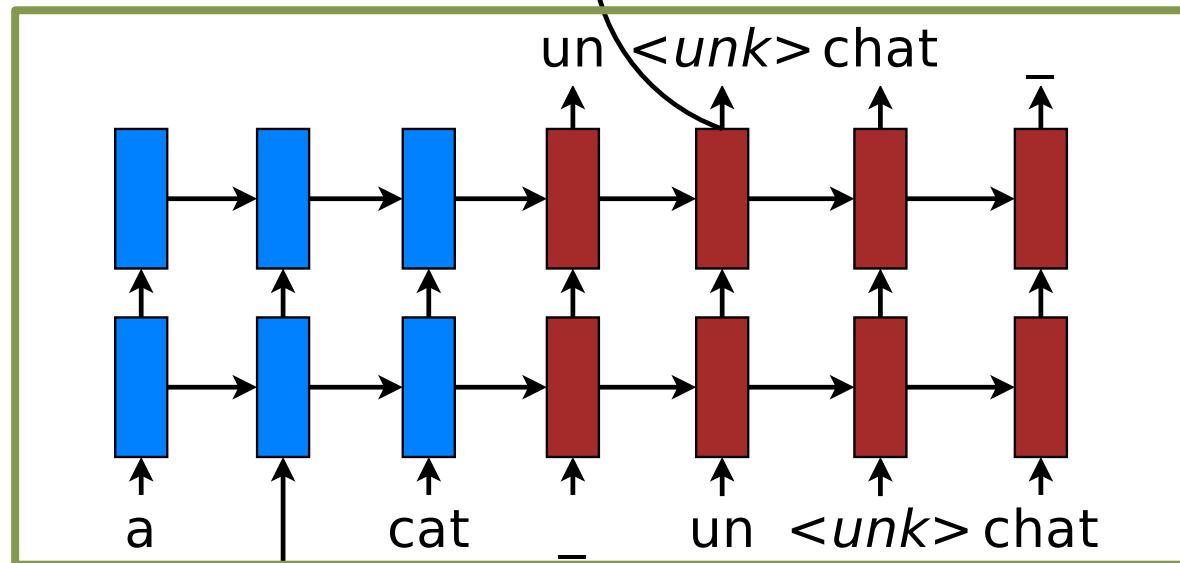
- A *best-of-both-worlds* architecture:
 - Translate mostly at the **word** level
 - Only go the **character** level when needed.
- Additional $+2.1 \mapsto +11.4$ BLEU improvement.

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.

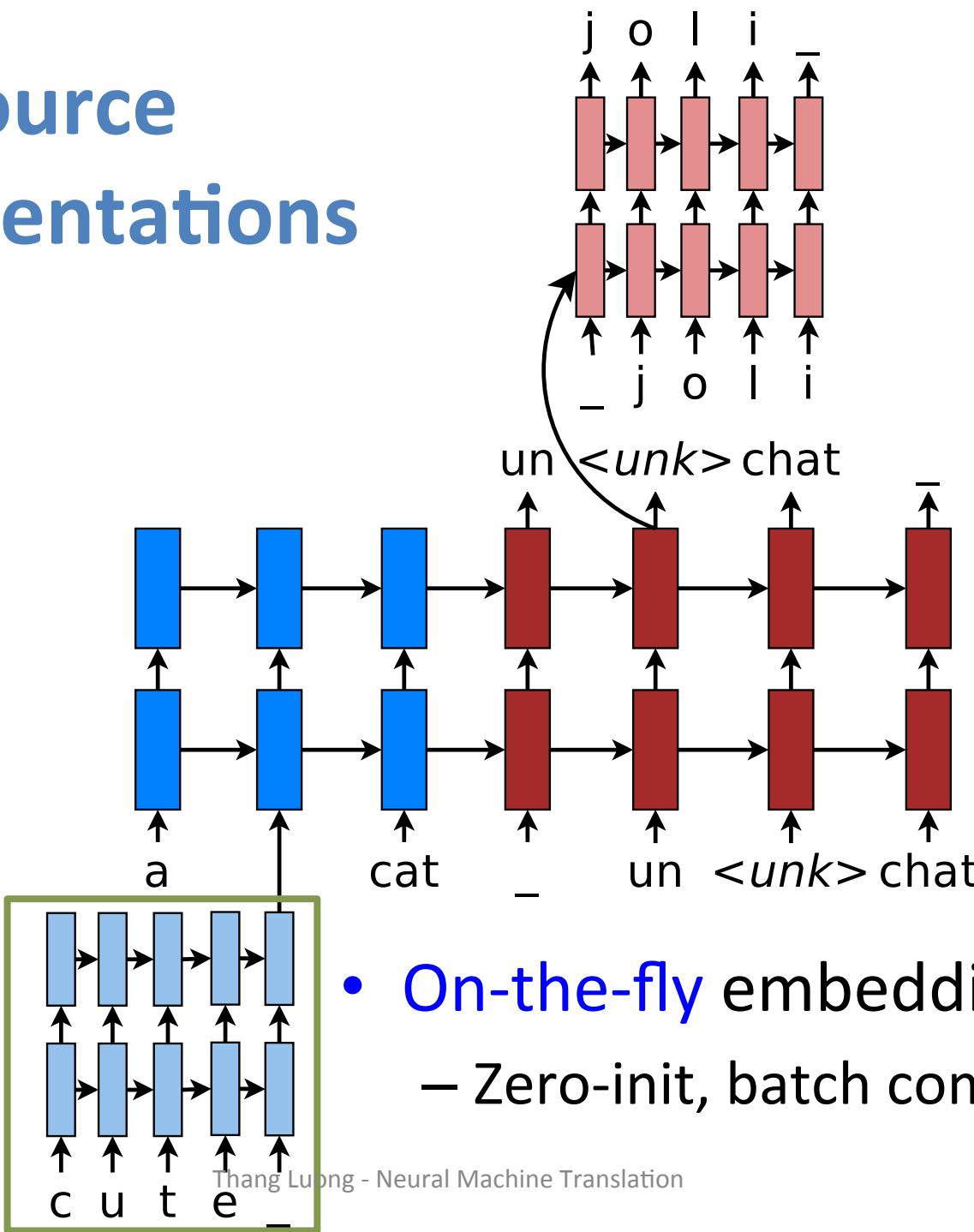
Hybrid NMT

Word-level
(4 layers)

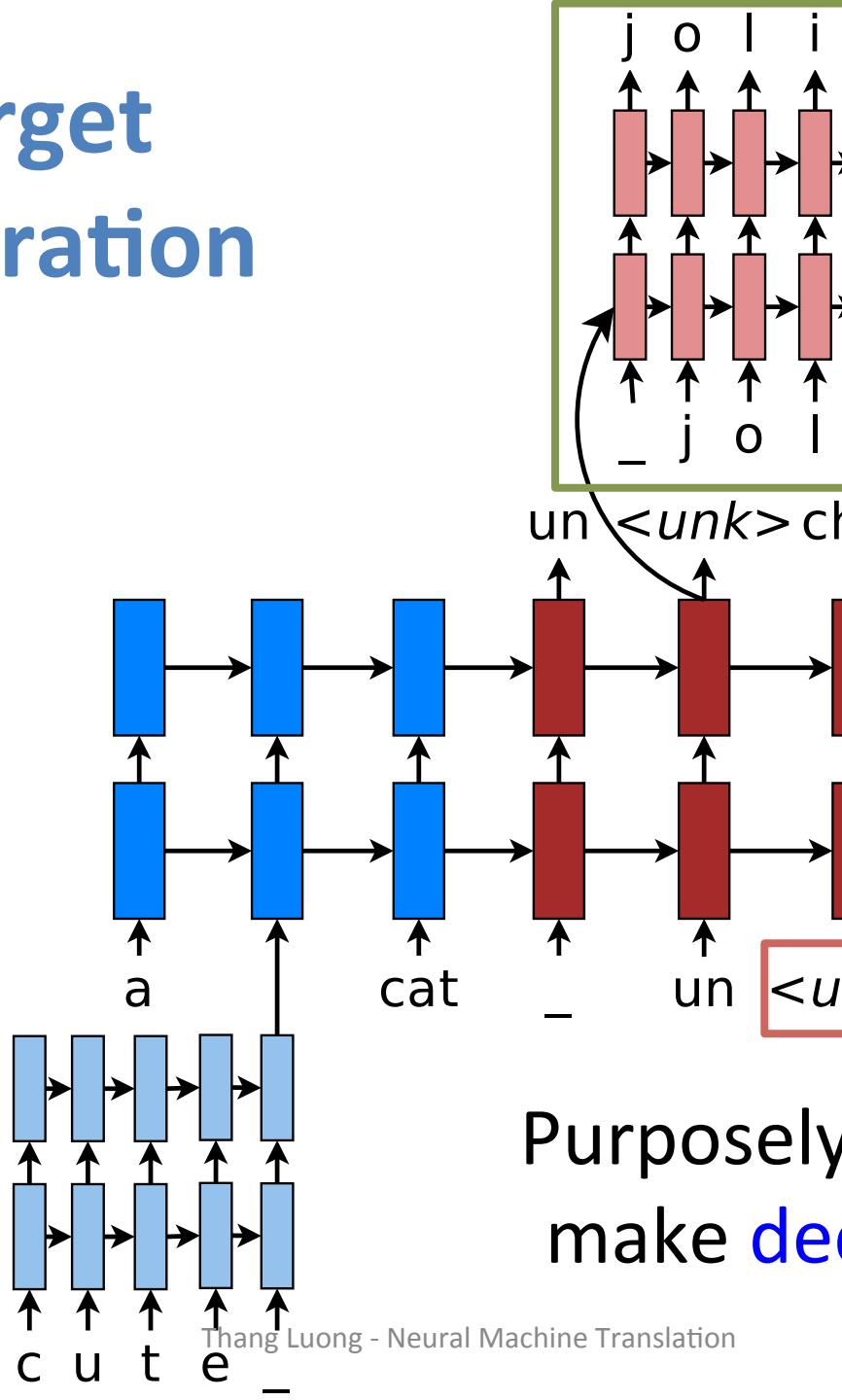


End-to-end training
8-stacking LSTM layers.

Source Representations



Target Generation



Init with word hidden states.

Purposely, use <unk> to make decoding easier.

English-Czech WMT'15 Results

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

} 30x data
3 systems

} Large vocab
+ unk replace

English-Czech WMT'15 Results

| Systems | BLEU |
|--|-------------|
| <i>Winning entry</i> (Bojar & Tamchyna, 2015) | 18.8 |
| <i>Existing word-level NMT</i> (Jean et al., 2015) | |
| <i>Single model</i> | 15.7 |
| <i>Ensemble</i> 4 models | 18.3 |
| <i>Our character-based NMT</i> | |
| <i>Single model</i> (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



English-Czech WMT'15 Results

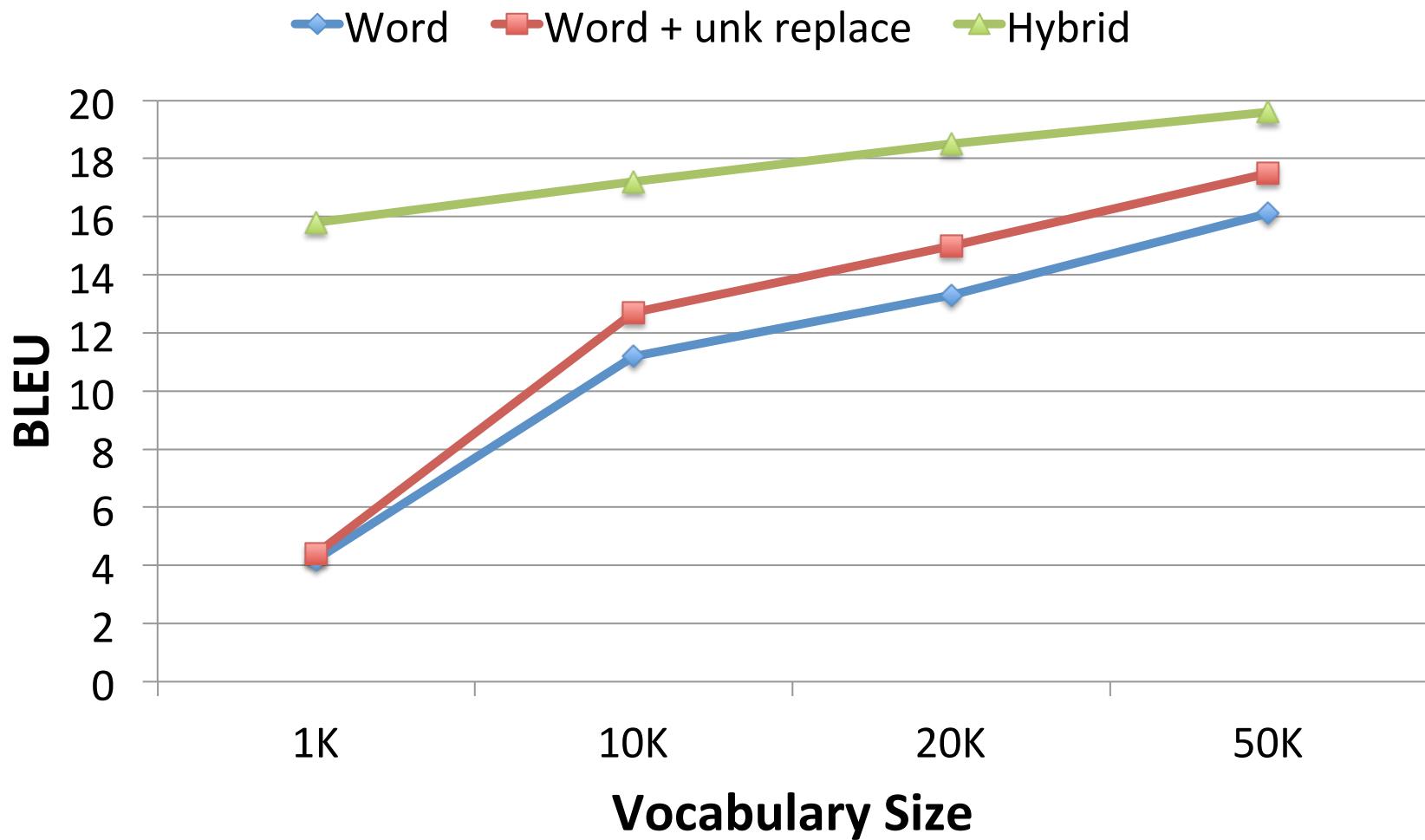
| Systems | BLEU |
|--|-------------|
| Winning entry (Bojar & Tamchyna, 2015) | 18.8 |
| Existing word-level NMT (Jean et al., 2015) | |
| Single model | 15.7 |
| Ensemble 4 models | 18.3 |
| Our character-based NMT | |
| Single model (600-step backprop) | 15.9 |
| Our hybrid NMT | |
| Single model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

} 30x data
3 systems

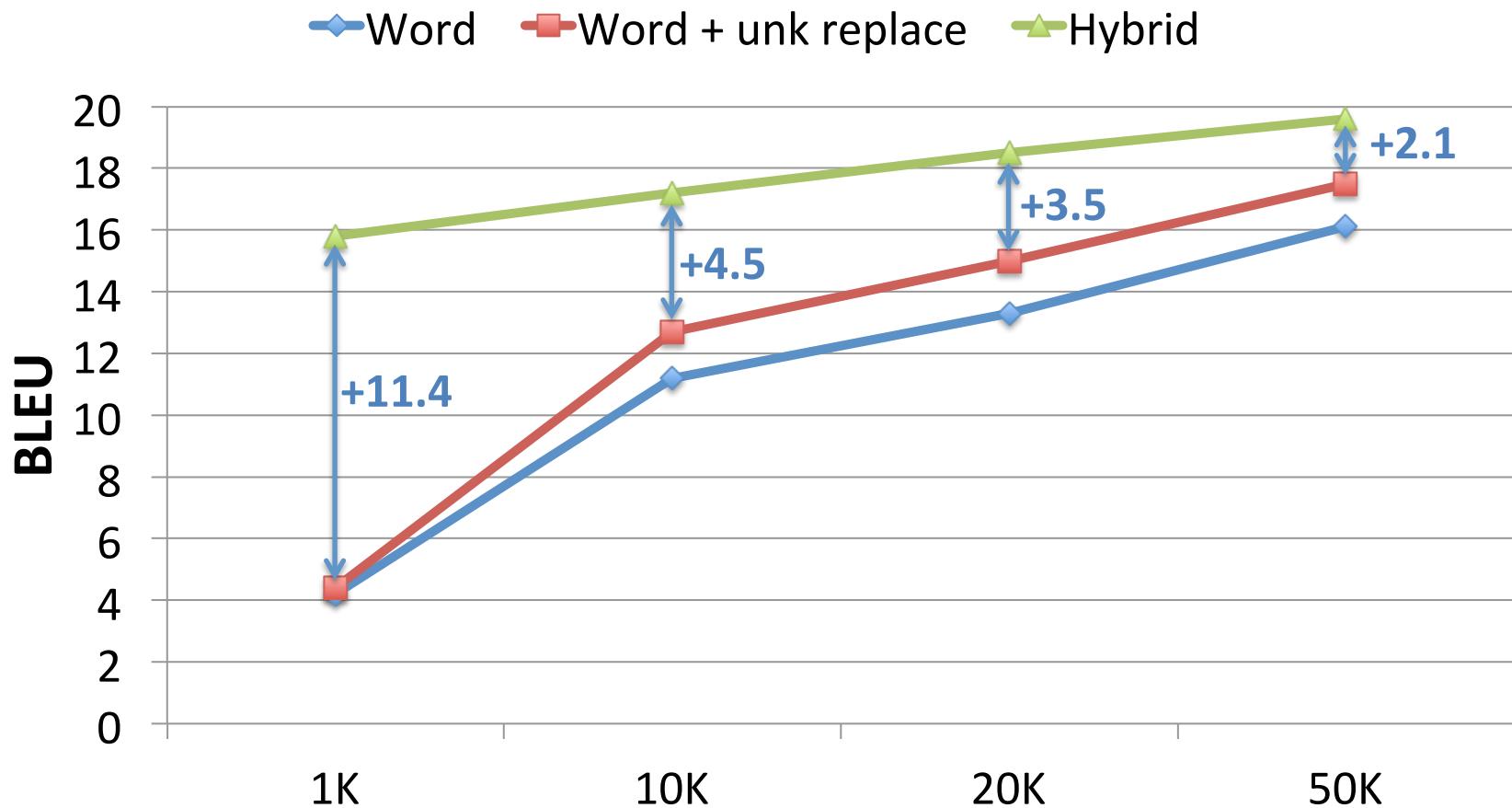
} Large vocab
+ unk replace



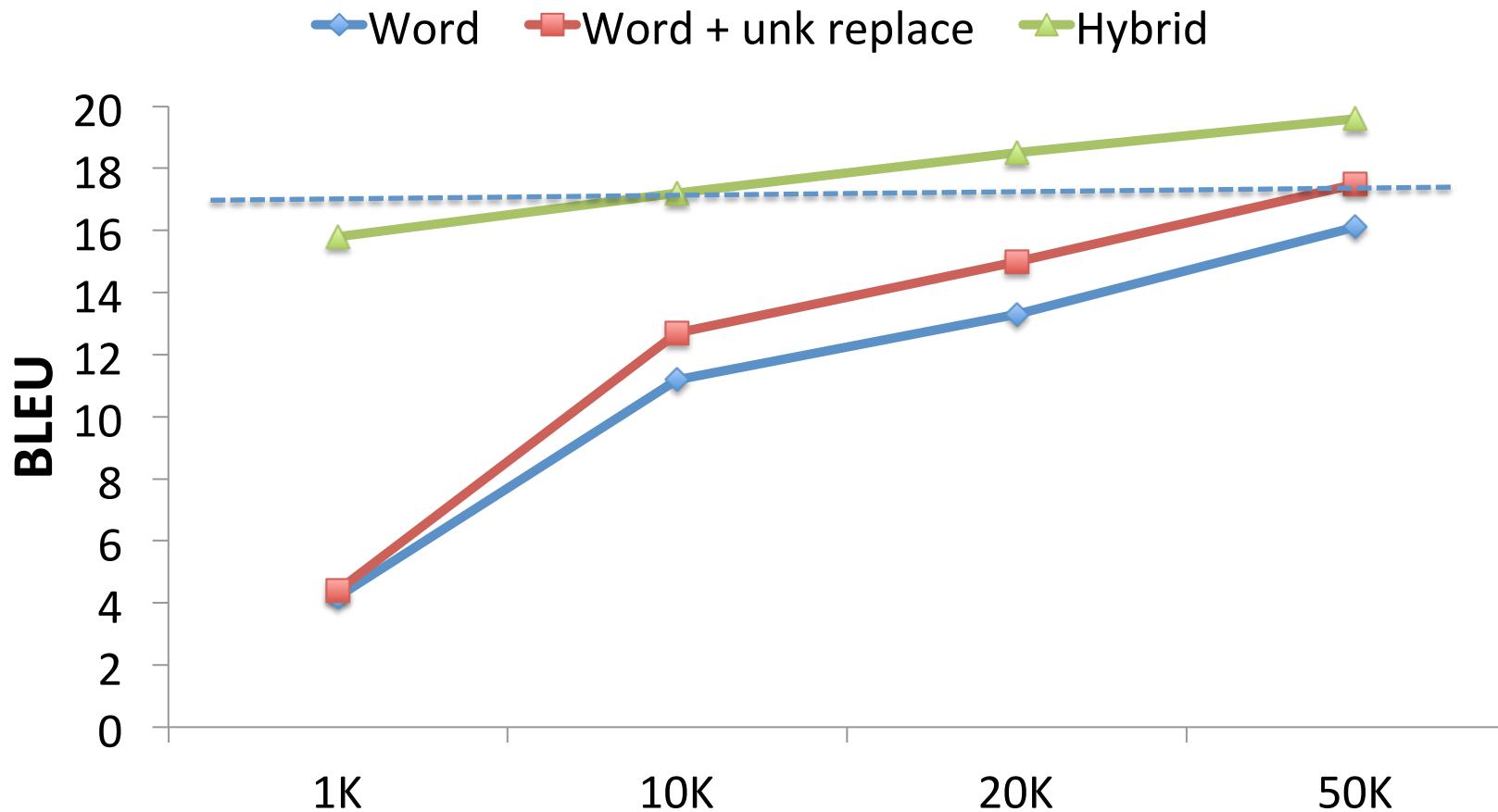
Effects of Vocabulary Sizes



Effects of Vocabulary Sizes

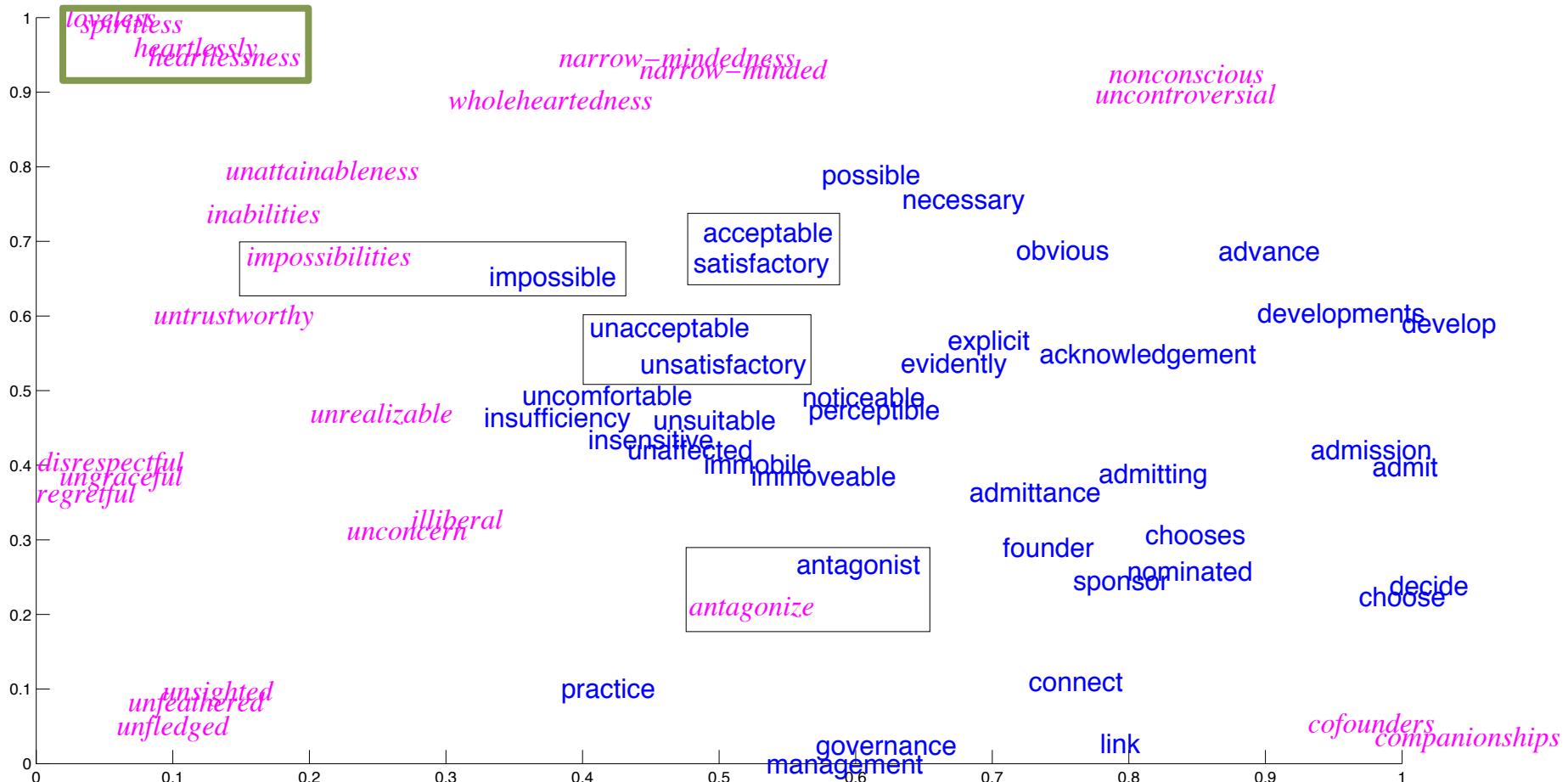


Effects of Vocabulary Sizes



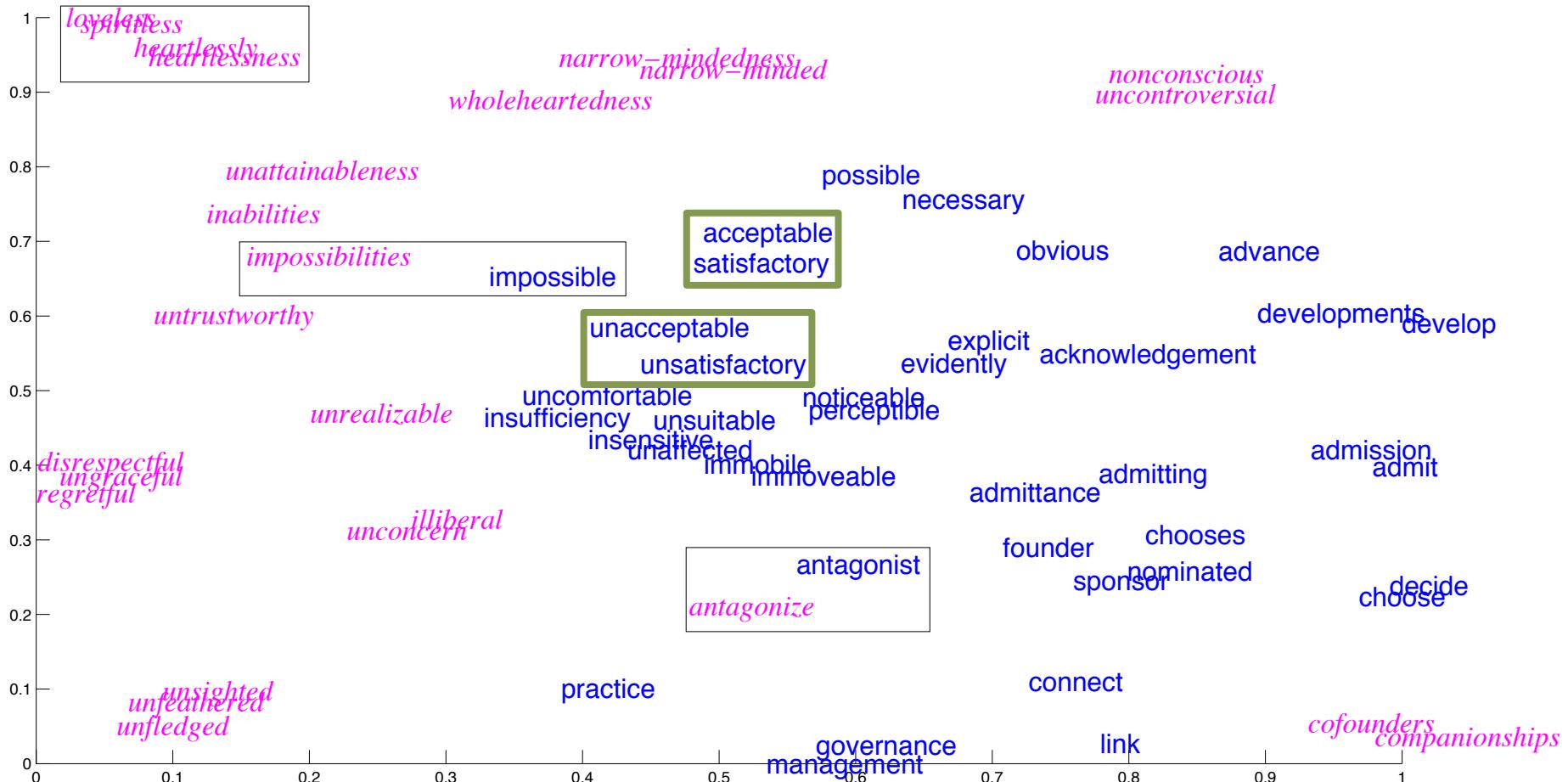
Small-vocab hybrid = Large-vocab word

Rare Word Embeddings



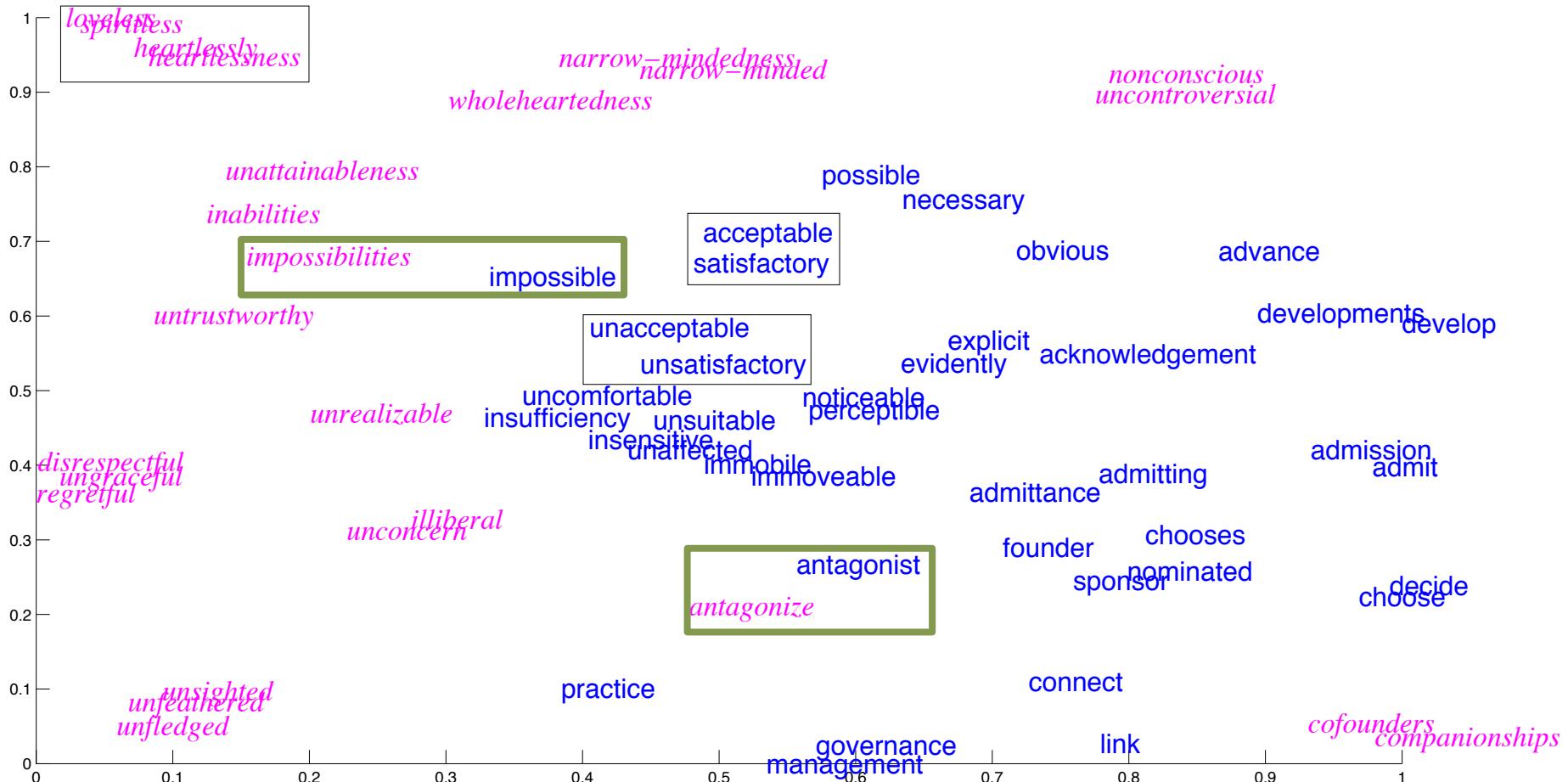
- Word & character-based embeddings.

Rare Word Embeddings



- Word & character-based embeddings.

Rare Word Embeddings



- Word & character-based embeddings.

Sample English-Czech translations

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



Sample English-Czech translations

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

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

Sample English-Czech translations

| | |
|--------|---|
| source | As the Reverend Martin Luther King Jr. <i>said fifty years ago</i> : |
| 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. : |



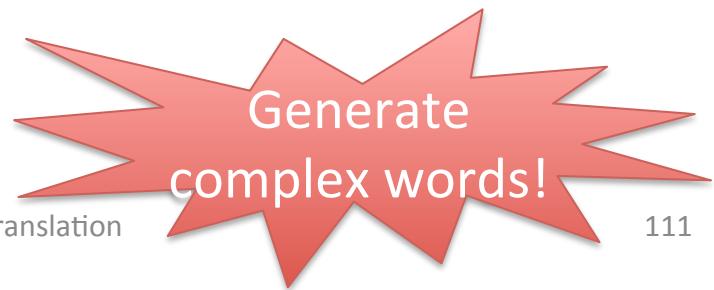
Sample English-Czech translations

| | |
|--------|---|
| source | As the Reverend Martin Luther King Jr. <i>said fifty years ago</i> : |
| human | Jak před padesáti lety řekl reverend Martin Luther King Jr. : |
| char | Jako reverend Martin Luther král <i>říkal před padesáti lety</i> : |
| word | Jak řekl reverend Martin <unk> King <unk> před padesáti lety : Jak řekl reverend Martin Luther King řekl <i>před padesáti lety</i> : |
| 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 |
| hybrid | Její 11-year-old dcera Shani , řekla , že je to trochu divné |
| | Její <unk> dcera , <unk> <unk> , řekla , že je to <unk> <unk> |
| | Její jedenáctiletá dcera , Graham Bart , řekla , že cítí trochu divný |



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 |
| hybrid | Její 11-year-old dcera Shani , řekla , že je to trochu divné |
| | 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

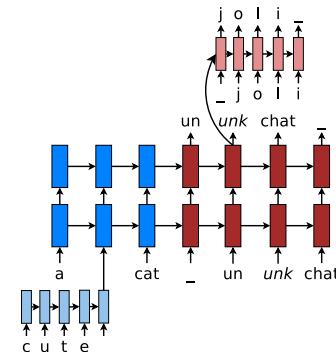
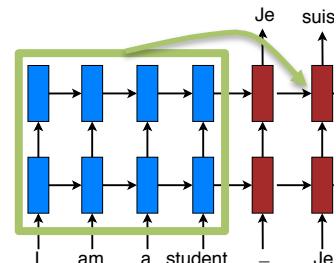
| | |
|--------|---|
| source | Her 11-year-old daughter , Shani Bart , said it felt a little bit weird |
| human | Její jedenáctiletá dcera Shani Bartová prozradila , že je to trochu zvláštní |
| char | Její jedenáctiletá dcera , Shani Bartová , říkala , že cítí trochu divně |
| word | Její <unk> dcera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

The ecotax <unk> portico in Pont-de-Buis
Le <unk> <unk> de <unk>
portique écotaxe Pont-de-Buis



NMT & beyond

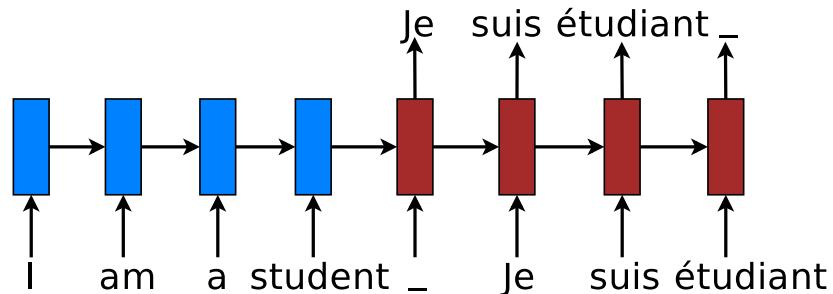


- **Unsupervised learning** for NMT
 - Utilize monolingual data.
- **Long-context** NMT
 - Translating an article / a book.
 - Smarter attention, longer sequences.
- **Multi-modal** Language Understanding System
 - Multi-lingual translation + speech recognition + more
 - Multi-task learning

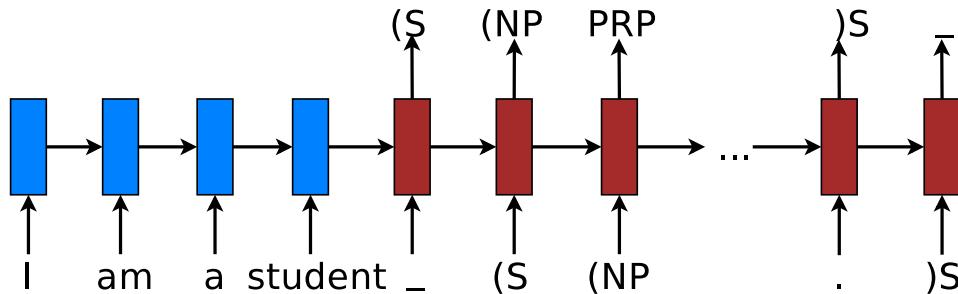
I'm done.
But if you are curious, read on!

#4 For the future of NMT

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

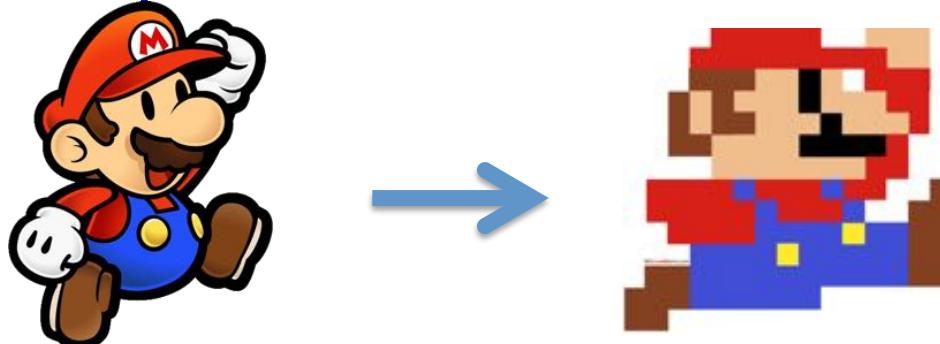


Machine translation



Constituent parsing

- Can we compress NMT for mobile devices?



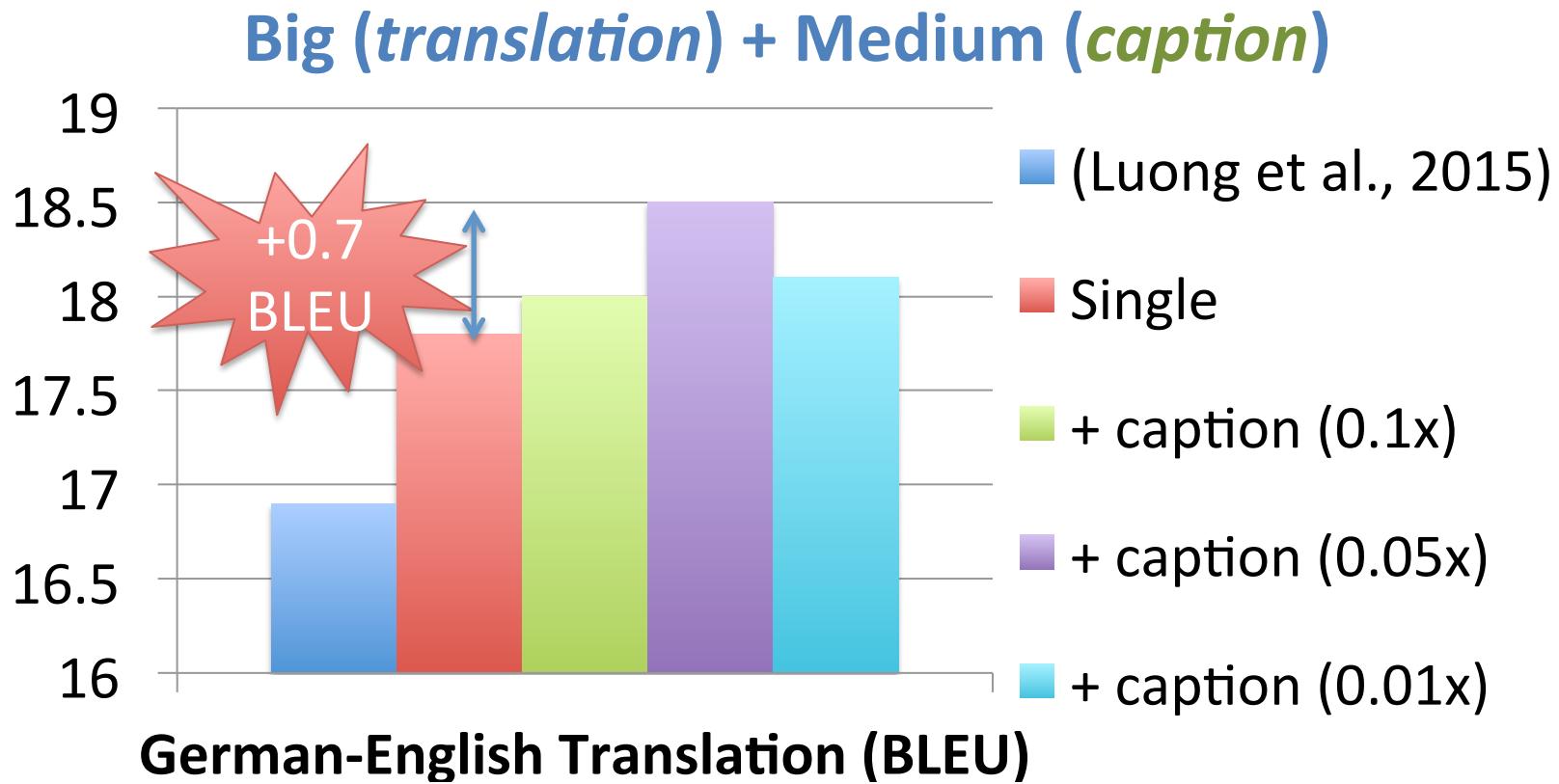
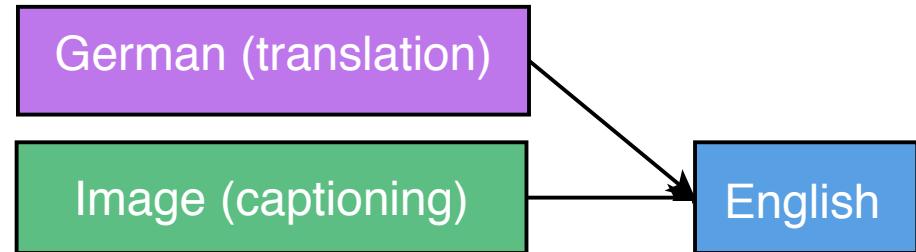
Our work



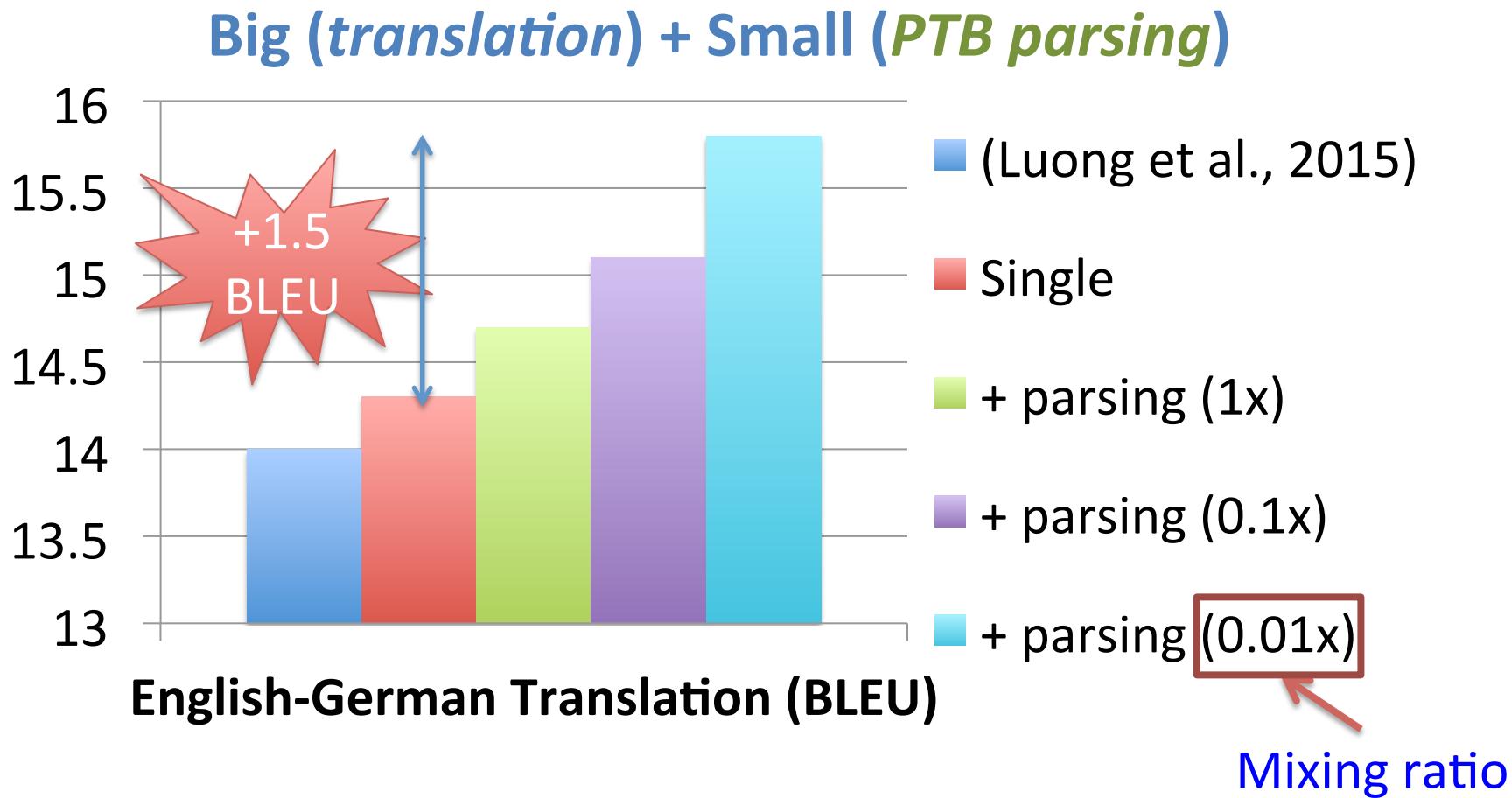
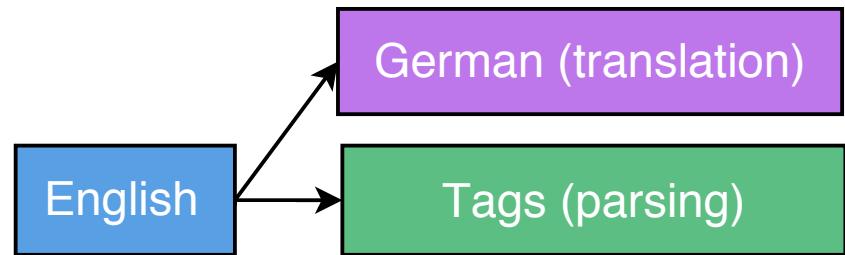
- Multi-task learning:
 - Machine translation
 - Image caption generation
 - Constituent parsing
 - Unsupervised learning
- Translation improvement: up to +1.5 BLEU.
- State-of-the-art in constituent parsing.

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

Many-to-one: shared decoder



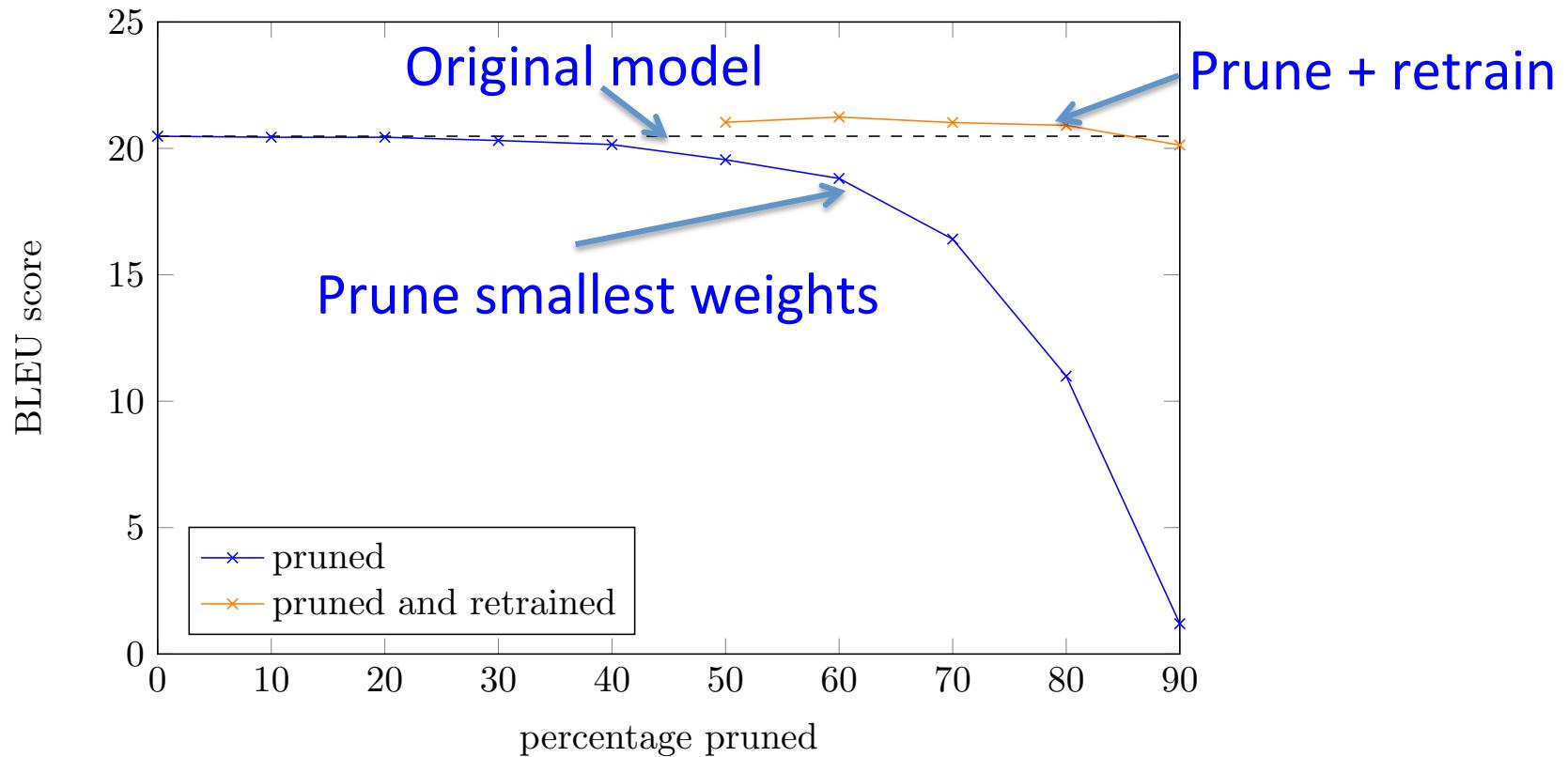
One-to-many: shared encoder





Our work

- Compress NMT via pruning & retraining:

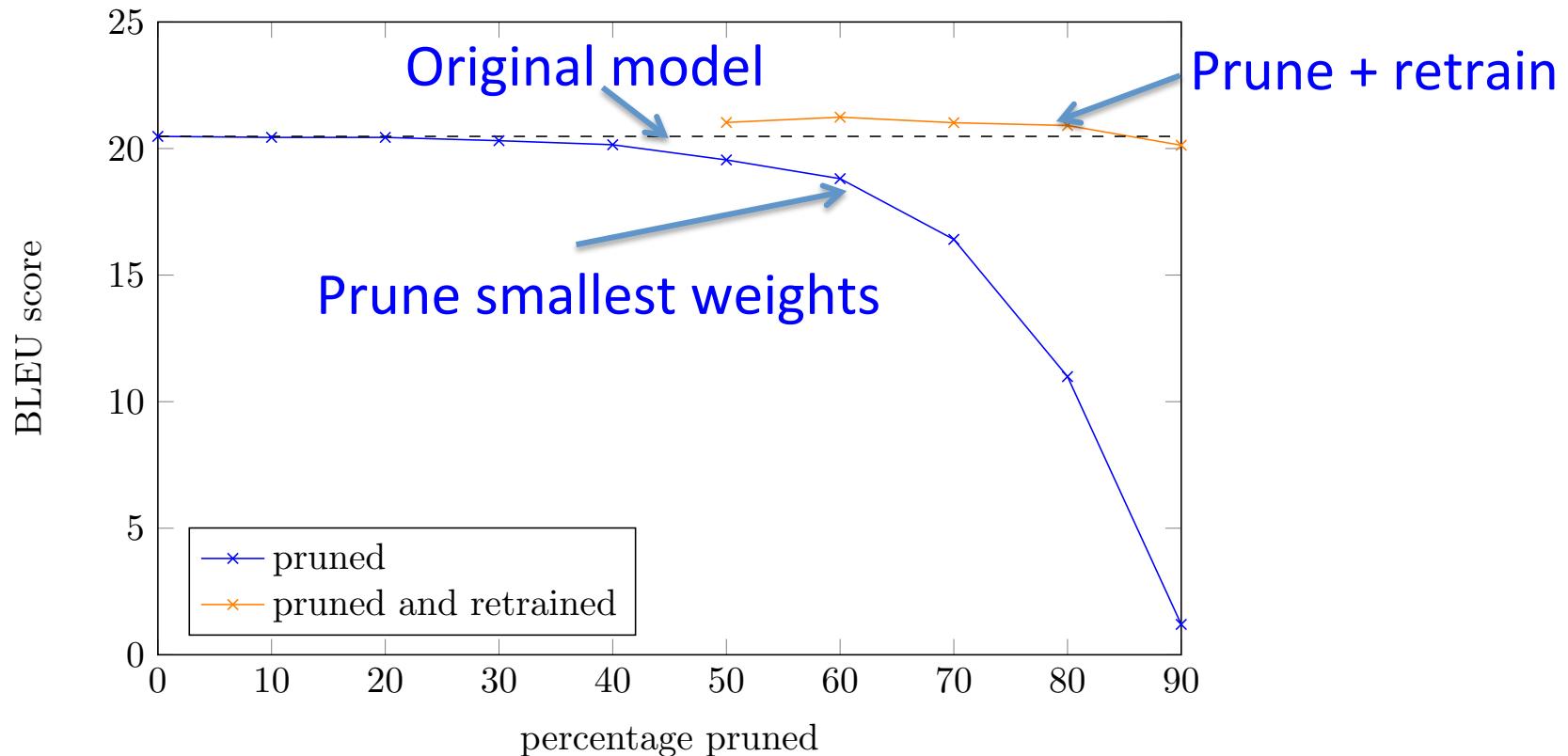


Abigail See, Thang Luong*, and Chris Manning. **Compression of Neural Machine Translation Models via Pruning.** In submission.*

Our work



- Compress NMT via pruning & retraining:



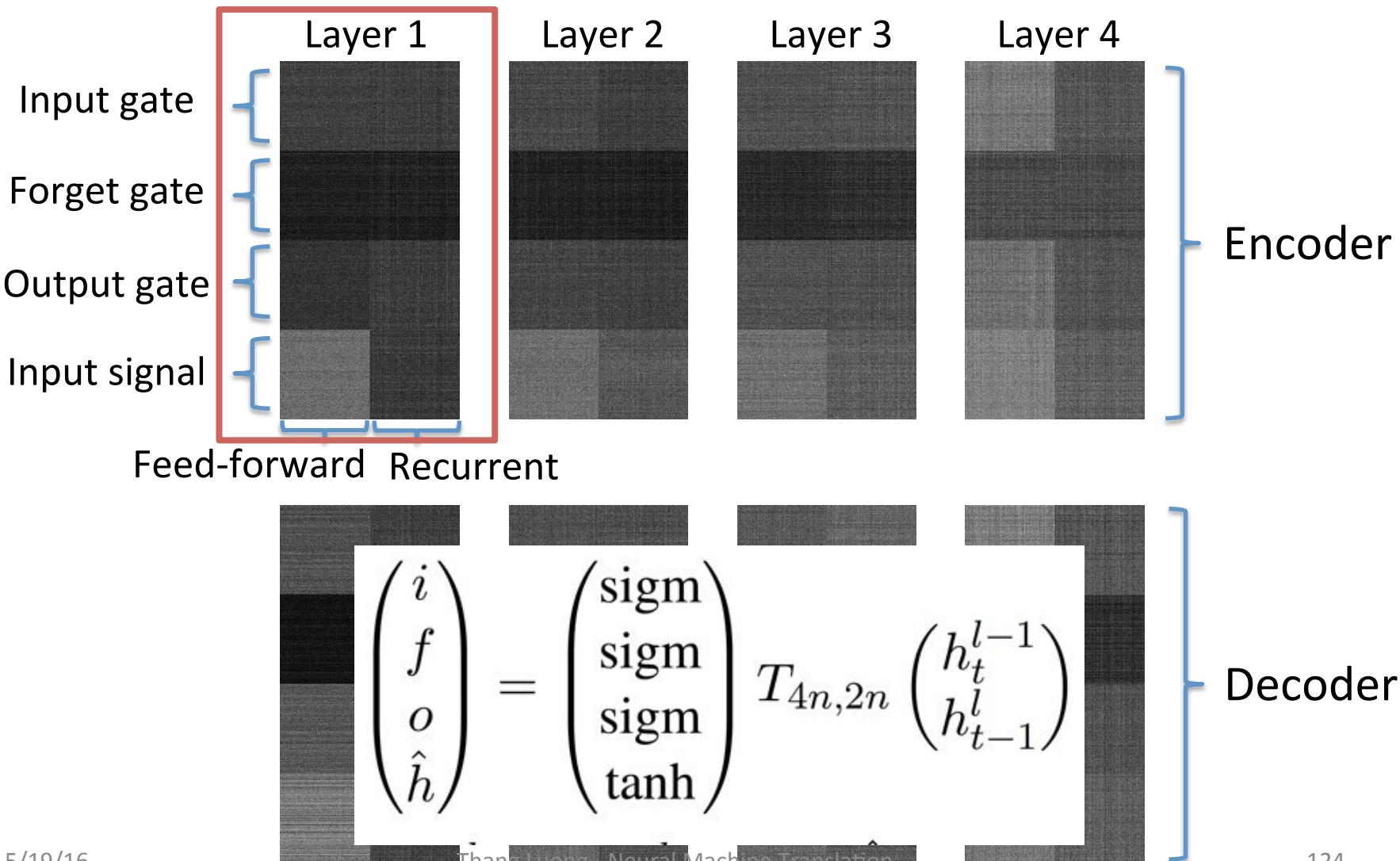
Prune 80% without loss of performance.

NMT Redundancy – *Embeddings*

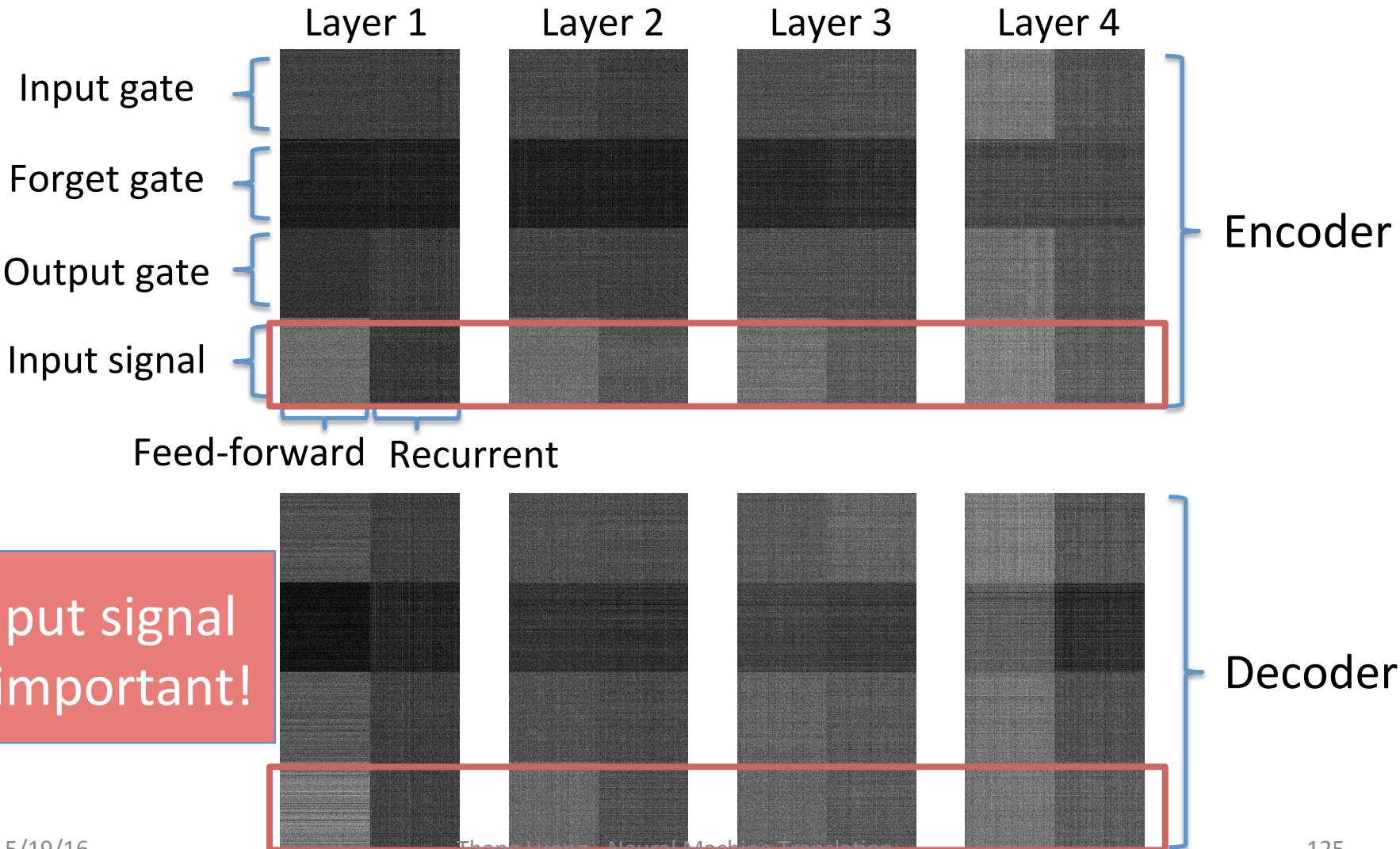
- Frequent words have larger weights
 - white: large.
 - black: small.



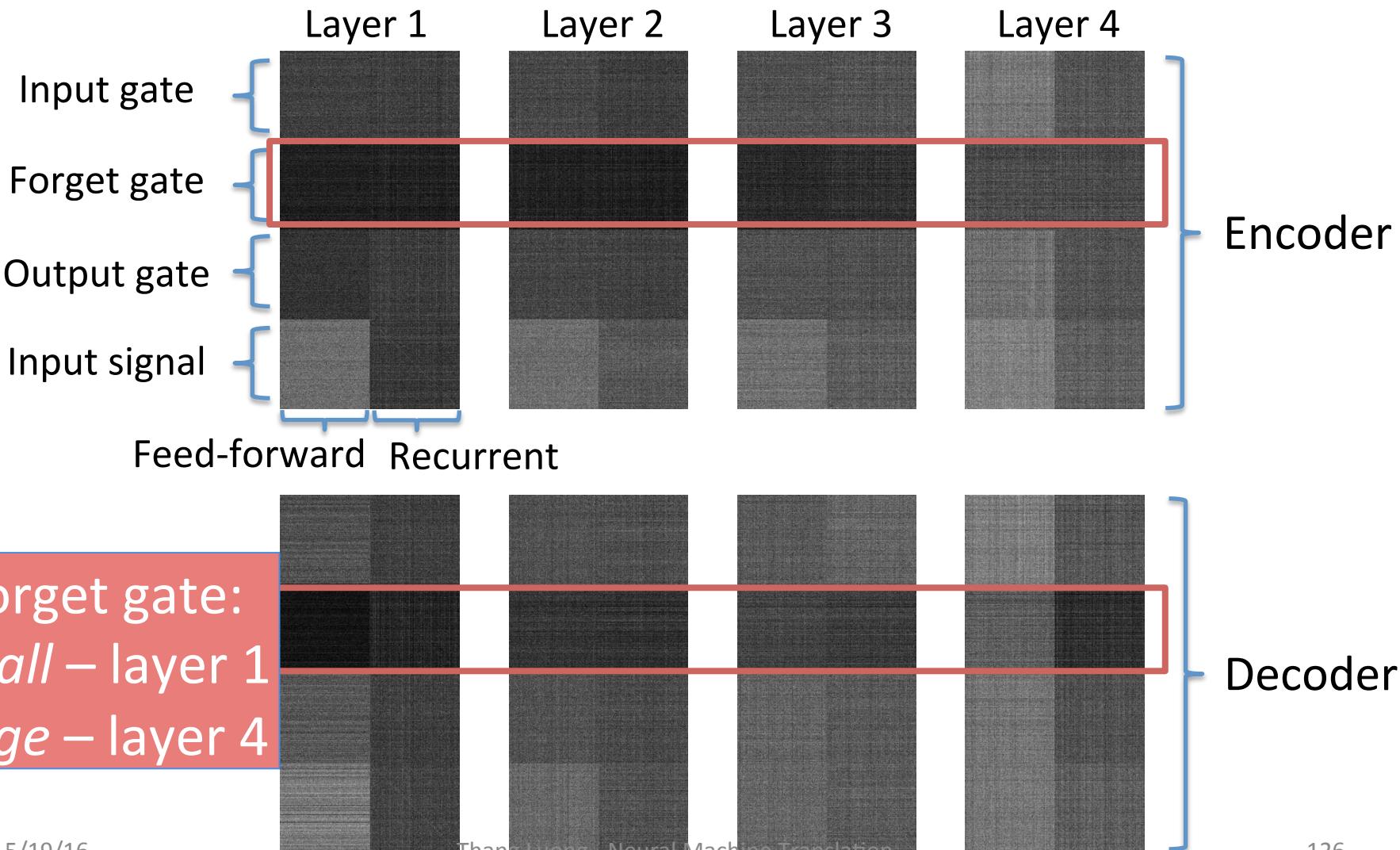
NMT Redundancy – LSTM



NMT Redundancy – LSTM



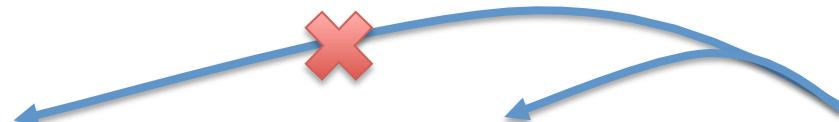
NMT Redundancy – LSTM



Future Challenges



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.