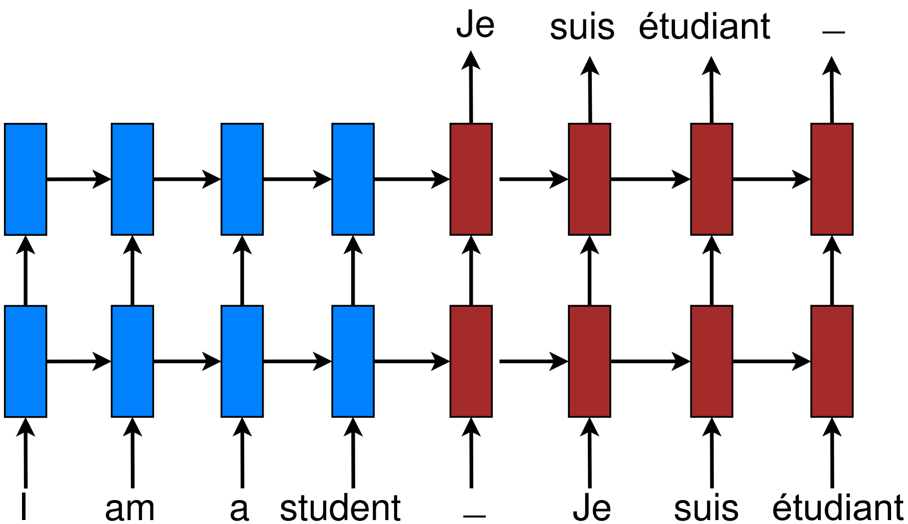
Effective Approaches to Attention-based Neural Machine Translation

Thang Luong Hieu Pham and Chris Manning
EMNLP 2015
Presented by: Yunan Zhang

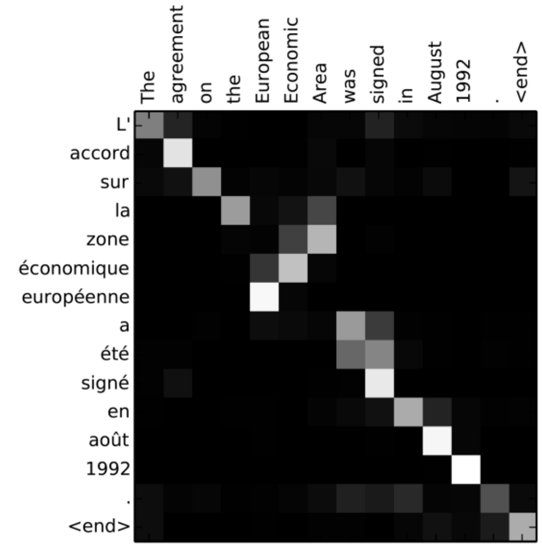
Neural Machine Translation

(Sutskever et al., 2014)



Attention Mechanism

(Bahdanau et al., 2015)



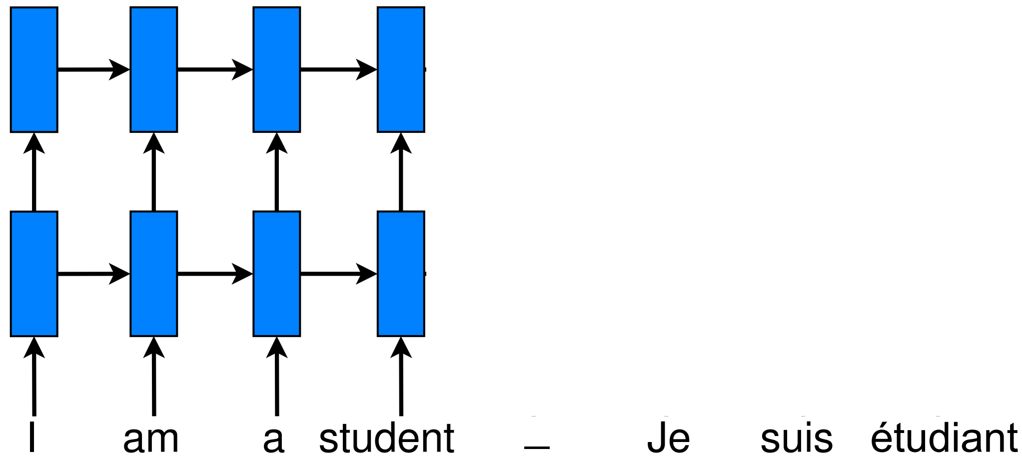
Recent innovation in deep learning:

- Control problem (Mnih et al., 14)
- Speech recognition (Chorowski et al., 14)
- Image captioning (Xu et al., 15)

New approach: recent SOTA results

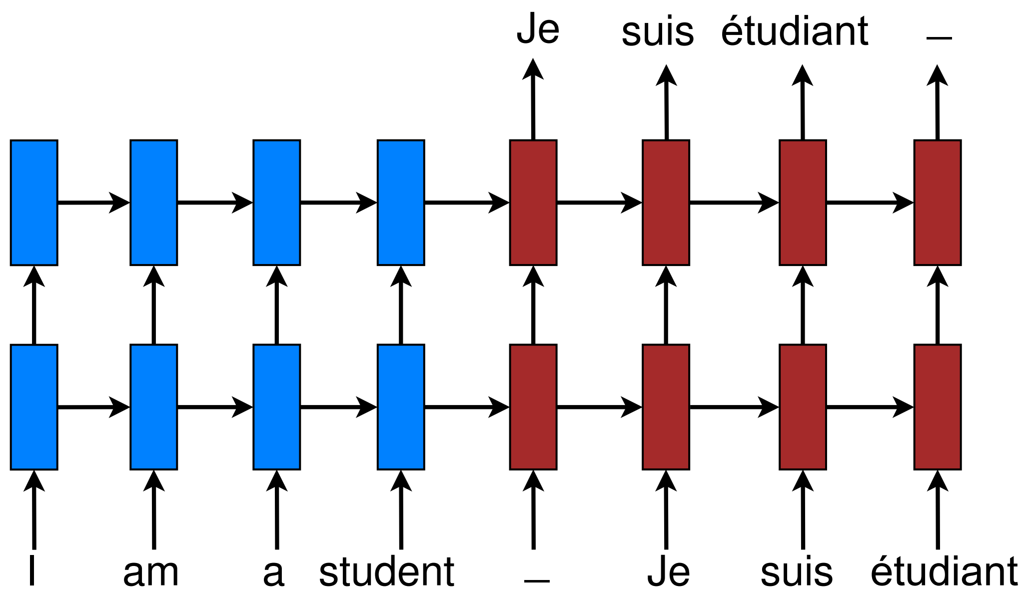
- Propose a new and better attention mechanism.
- Examine other variants of attention models.
- Achieve new SOTA results WMT English-German.

Neural Machine Translation (NMT)



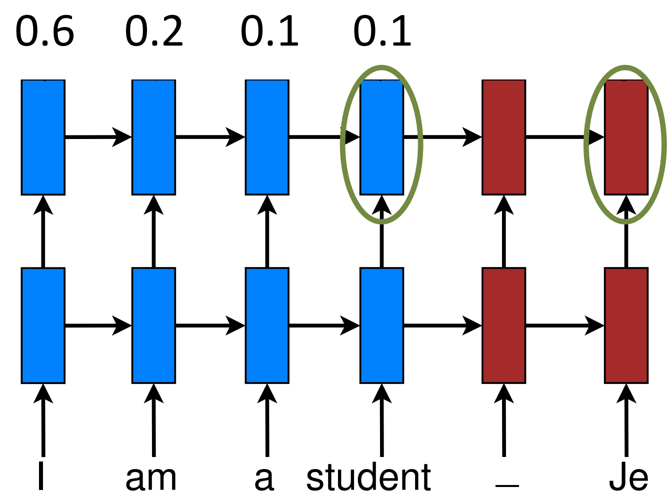
- Big RNNs trained end-to-end.

Neural Machine Translation (NMT)



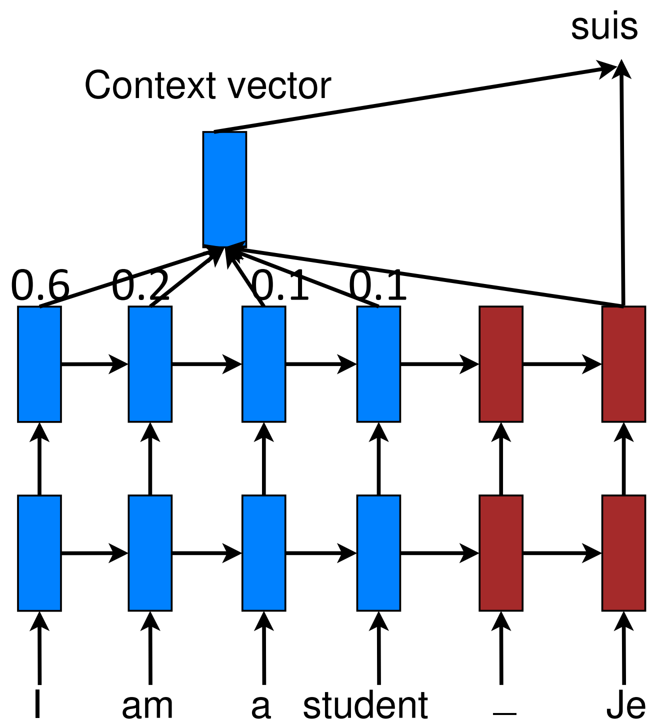
- Big RNNs trained end-to-end: encoder-decoder.
  - Generalize well to long sequences.
  - Small memory footprint.
  - Simple decoder.

Attention Mechanism



- Maintain a memory of source hidden states
- Memory here means a weighted average of the hidden states
- The weight is determined by comparing the

Attention Mechanism

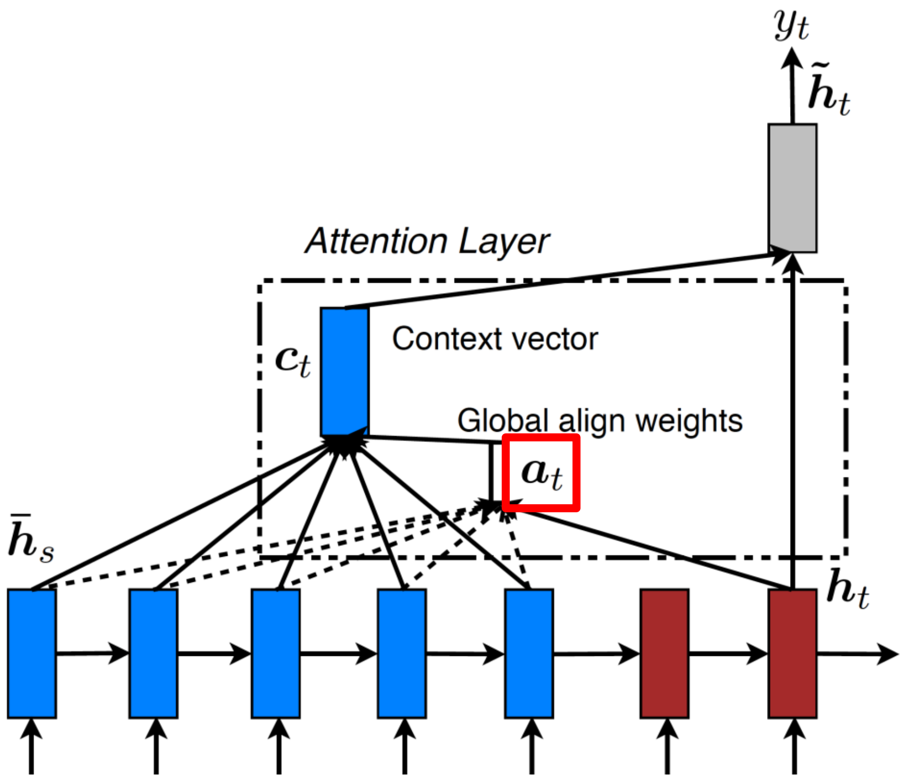


- Maintain a memory of source hidden states
  - Able to translate long sentences.

Motivation

- A new attention mechanism: local attention
  - Use a subset of source states each time.
  - Better results with focused attention!
- Global attention: use all source states
  - Other variants of (Bahdanau et al., 15)

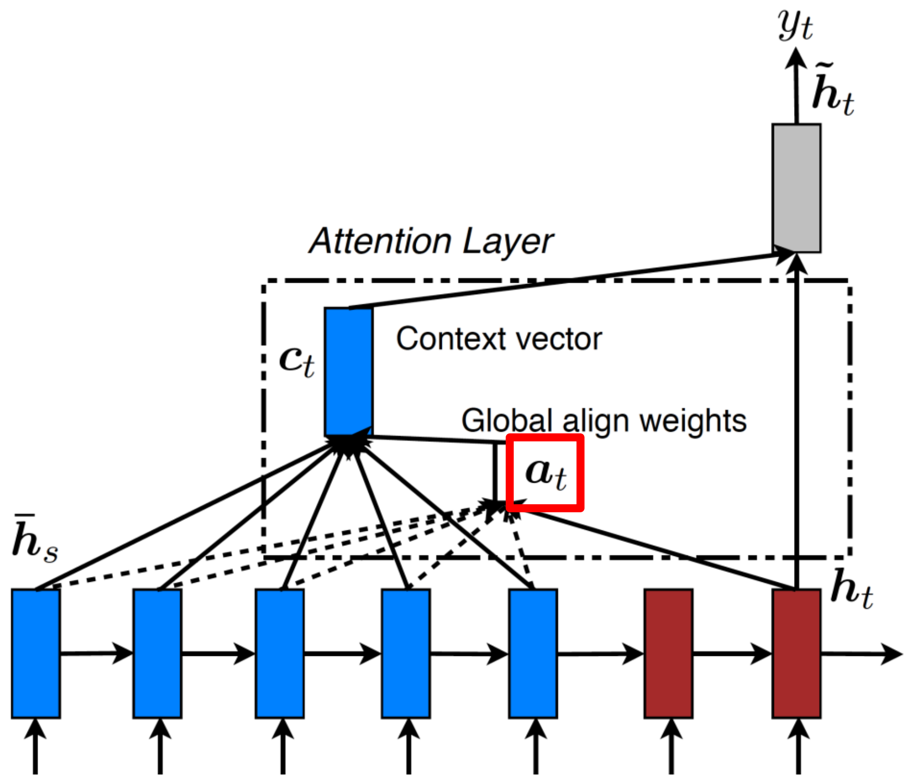
Global Attention



- Alignment weight vector:

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

Global Attention



$$\boldsymbol{a}_t(s) = \operatorname{align}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \frac{\exp\left(\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\operatorname{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)}$$

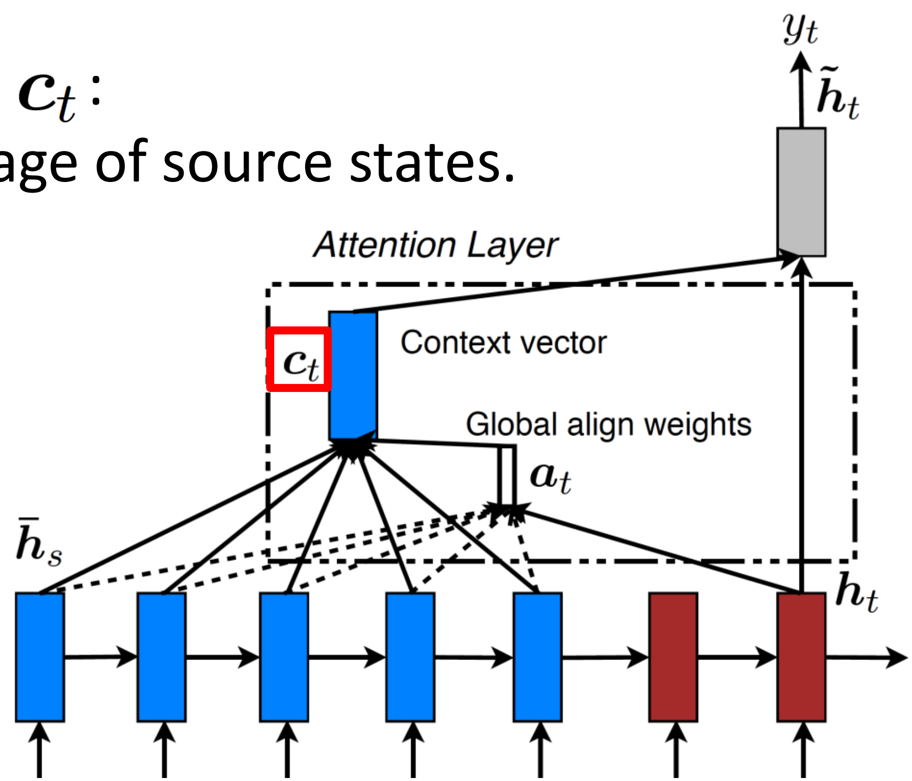
- Alignment weight vector:

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

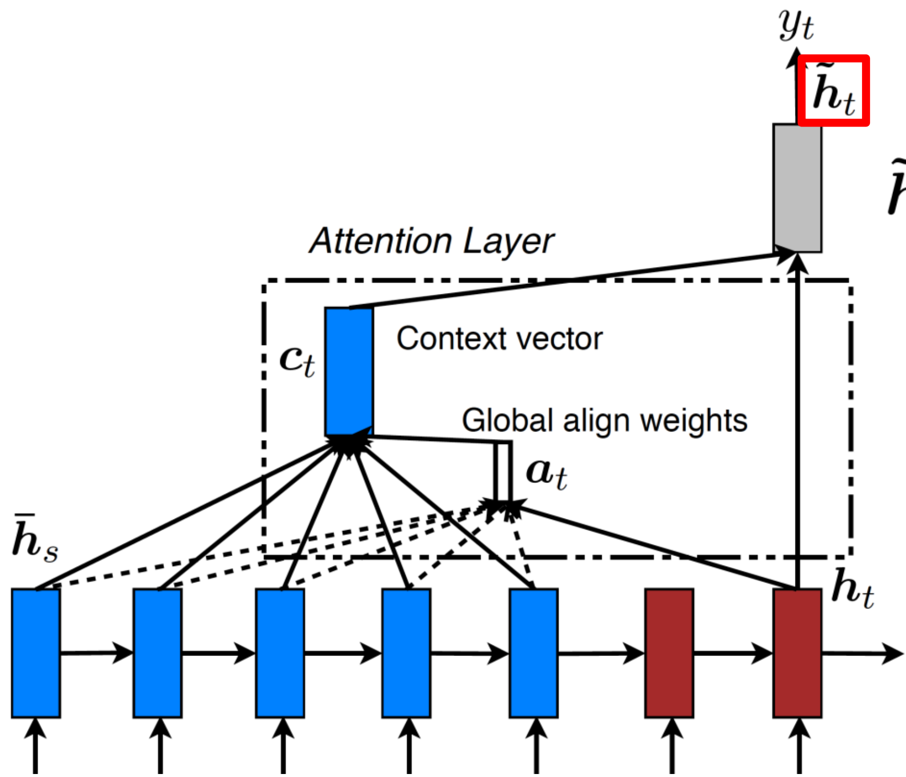
(Bahdanau et al., 15)

Global Attention

Context vector $\boldsymbol{c}_t$:
weighted average of source states.



Global Attention

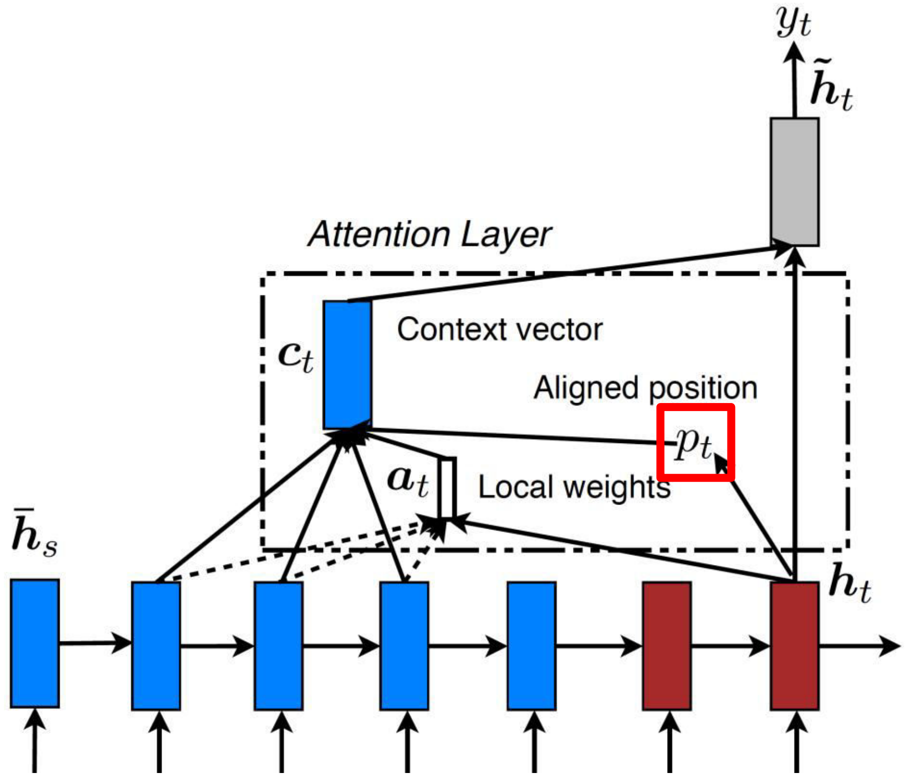


Attentional vector

$$\tilde{\boldsymbol{h}}_t = \tanh(\boldsymbol{W_c}[\boldsymbol{c}_t; \boldsymbol{h}_t])$$

$$p(y_t | y_{<t}, x) = \mathrm{softmax}(\boldsymbol{W_s} \tilde{\boldsymbol{h}}_t)$$

Local Attention

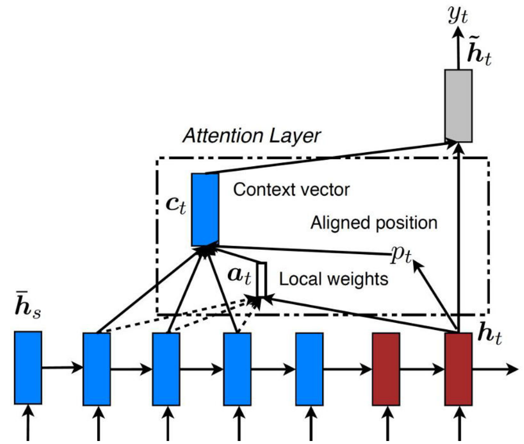


aligned positions?

- p_t defines a focused win $[p_t - D, p_t + D]$.

- A blend between soft & hard attention (Xu et

Local Attention (2)



- Predict aligned positions:

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

Real value in [0, S]

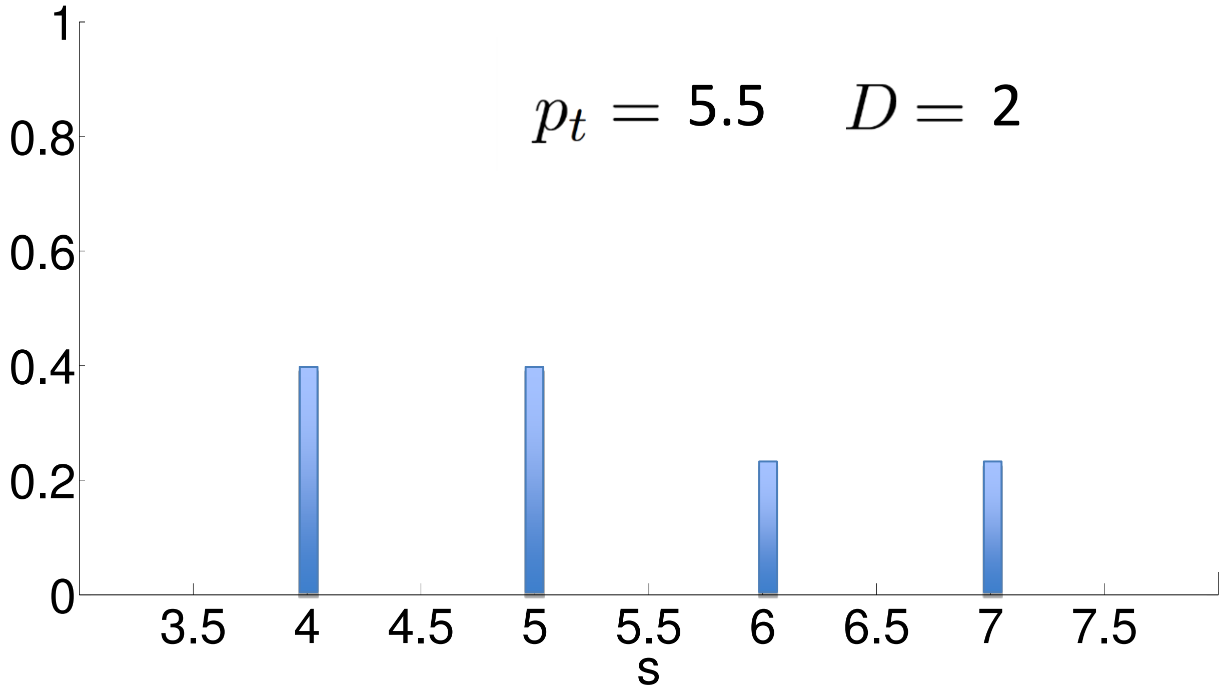
Source sentence

How do we learn to the position parameters?

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

Local Attention (3)

Alignment weights $\boldsymbol{a}_t(s)$



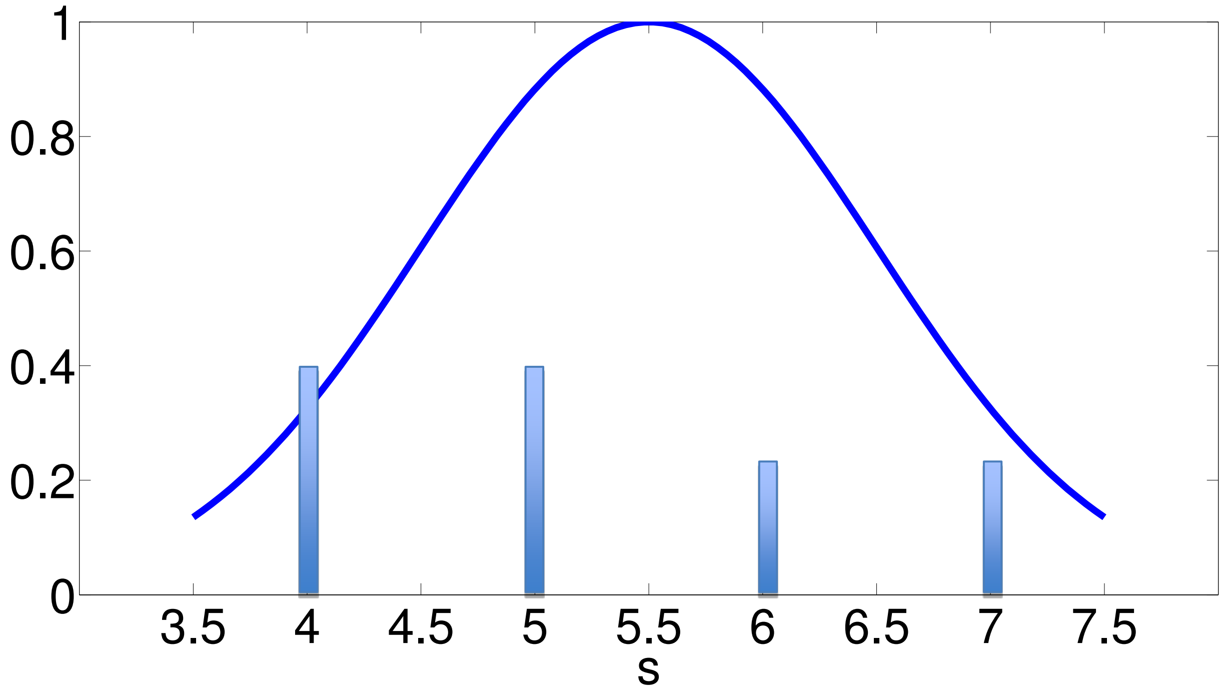
- Like global model: for integer s in $[p_t - D, p_t + D]$
  - Compute $\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)$

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

Local Attention (3)

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

Truncated Gaussian

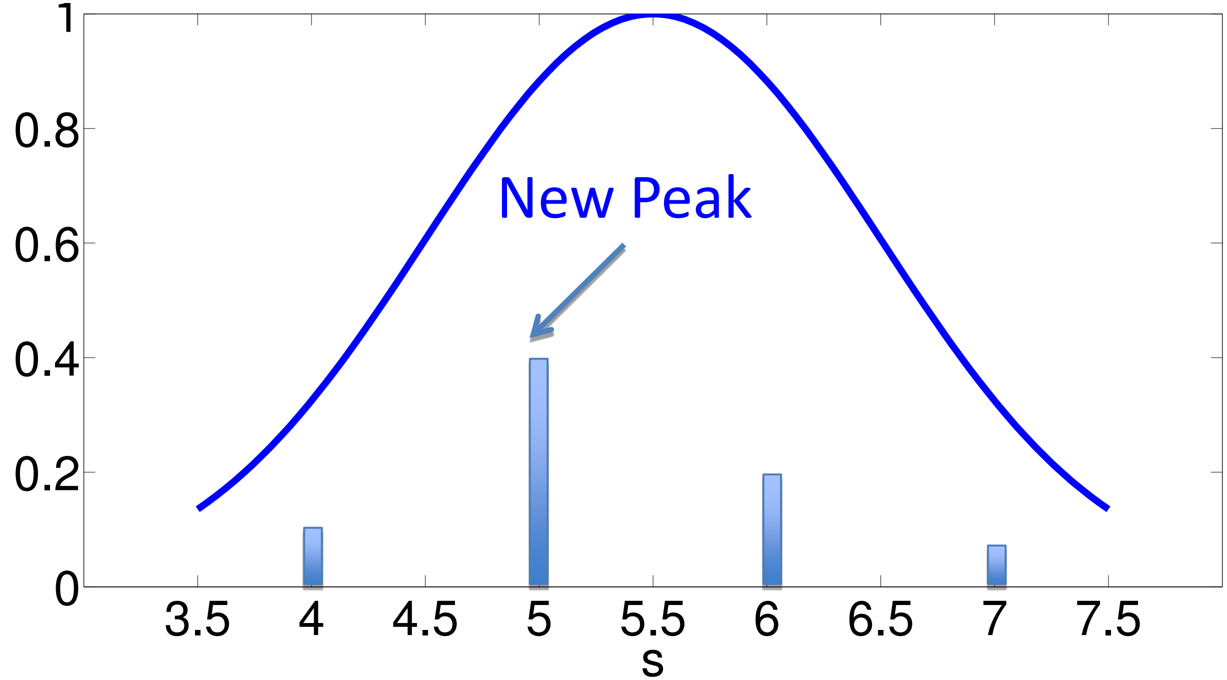


- Favor points close to the center.

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

Local Attention (3)

$$\frac{\exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s)\right)}{\sum_{s'} \exp\left(\mathrm{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_{s'})\right)} \quad \exp\left(-\frac{(s-p_t)^2}{2\sigma^2}\right)$$



Experiments

- WMT English ⇄ German (4.5M sentence pairs).

- Setup: (Sutskever et al., 14, Luong et al., 15)
  - 4-layer stacking LSTMs: 1000-dim cells/embeddings.
  - 50K most frequent English & German words

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning system – *phrase-based + large LM* (Buck et al.)		**20.7**
Our NMT systems		
Base	10.6	11.3

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning system – *phrase-based + large LM* (Buck et al.)		**20.7**
Our NMT systems		
Base	10.6	11.3
Base + **reverse**	9.9	12.6 (+1.3)

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning system – *phrase-based + large LM* (Buck et al.)		**20.7**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+1.3)
Base + reverse + **dropout**	8.1	14.0 (+1.4)

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning system – *phrase-based + large LM* (Buck et al.)		**20.7**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+1.3)
Base + reverse + dropout	8.1	14.0 (+1.4)
Base + reverse + dropout + **global attn**	7.3	**16.8 (+2.8)**

- Large progressive gains:
  - Attention: +2.8 BLEU

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning system – *phrase-based + large LM* (Buck et al.)		**20.7**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (+1.3)
Base + reverse + dropout	8.1	14.0 (+1.4)
Base + reverse + dropout + global attn	7.3	**16.8 (+2.8)**
Base + reverse + dropout + global attn + **feed input**	6.4	**18.1 (+1.3)**

- Large progressive gains:
  - Attention: +2.8 BLEU Feed input: +1.3 BLEU
- BLEU & perplexity correlation (Luong et al.,

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning sys – *phrase-based + large LM* (Buck et al., 2014)		**20.7**
Existing NMT systems (Jean et al., 2015)		
RNNsearch		16.5
RNNsearch + unk repl. + large vocab + *ensemble* 8 models		**21.6**
Our NMT systems		
Global attention	7.3	16.8 (+2.8)
Global attention + feed input	6.4	18.1 (+1.3)

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning sys – *phrase-based + large LM* (Buck et al., 2014)		**20.7**
Existing NMT systems (Jean et al., 2015)		
RNNsearch		16.5
RNNsearch + unk repl. + large vocab + *ensemble* 8 models		**21.6**
Our NMT systems		
Global attention	7.3	16.8 (+2.8)
Global attention + feed input	6.4	18.1 (+1.3)
Local attention + feed input	5.9	**19.0 (+0.9)**

- Local-predictive attention: +0.9 BLEU gain.

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning sys – *phrase-based + large LM* (Buck et al., 2014)		**20.7**
Existing NMT systems (Jean et al., 2015)		
RNNsearch		16.5
RNNsearch + unk repl. + large vocab + *ensemble* 8 models		**21.6**
Our NMT systems		
Global attention	7.3	16.8 (+2.8)
Global attention + feed input	6.4	18.1 (+1.3)
Local attention + feed input	5.9	19.0 (+0.9)
Local attention + feed input + **unk replace**	5.9	**20.9 (+1.9)**

- Unknown replacement: +1.9 BLEU
  - (Luong et al., '15), (Jean et al., '15).

English-German WMT'14 Results

Systems	Ppl	BLEU
Winning sys – *phrase-based + large LM* (Buck et al., 2014)		**20.7**
Existing NMT systems (Jean et al., 2015)		
RNNsearch		16.5
RNNsearch + unk repl. + large vocab + *ensemble* 8 models		**21.6**
Our NMT systems		
Global attention	7.3	16.8 (+2.8)
Global attention + feed input	6.4	18.1 (+1.3)
Local attention + feed input	5.9	19.0 (+0.9)
Local attention + feed input + unk replace	5.9	20.9 (+1.9)
Ensemble 8 models + unk replace		**23.0 (+2.1)**

New SOTA!

WMT'15 English-Results

English-German Systems	BLEU
Winning system – *NMT + 5-gram LM reranker* (Montreal)	24.9
Our ensemble 8 models + unk replace	**25.9**

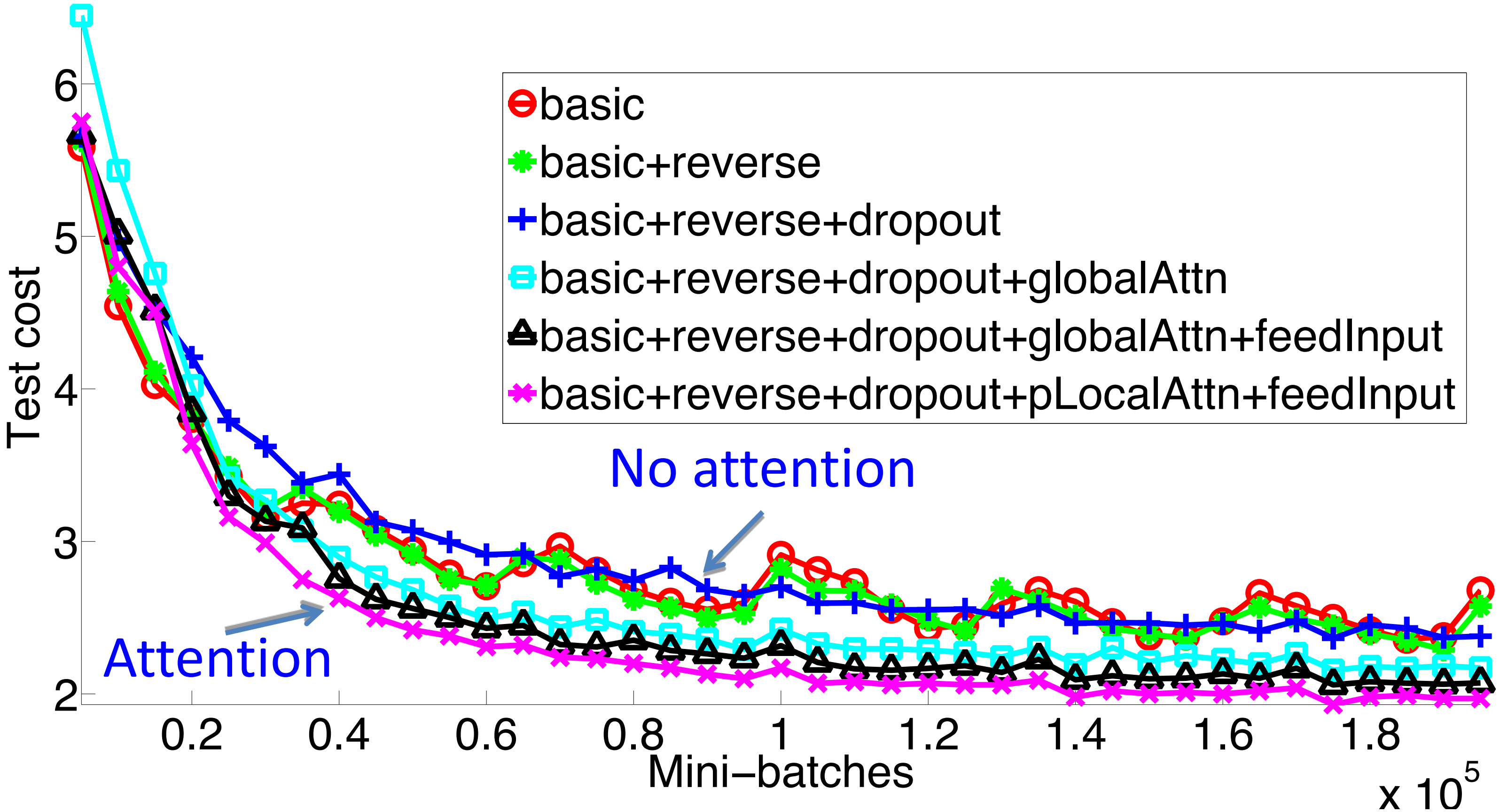
New SOTA!

- WMT'15 *German-English*: similar gains
  - Attention: +2.7 BLEU Feed input: +1.0 BLEU

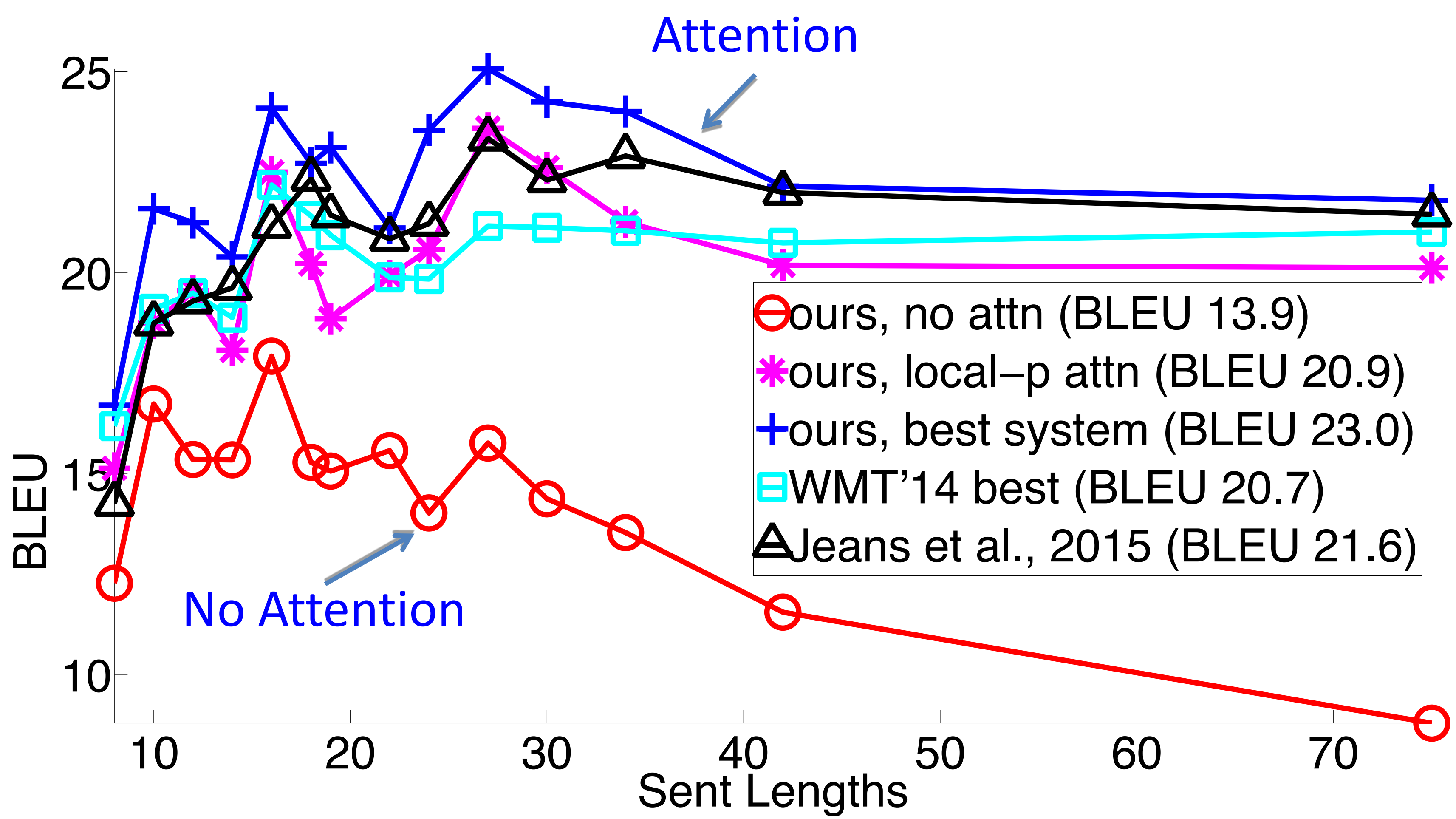
Analysis

- Learning curves
- Long sentences
- Alignment quality
- Sample translations

Learning Curves



Translate Long Sentences



Alignment Quality

Models	AER
Berkeley aligner	**0.32**
Our NMT systems	
Global attention	0.39
Local attention	0.36
Ensemble	0.34

- RWTH gold alignment data
  - 508 English-German Europarl sentences.
- Force decode our models

Competitive AERs!

Sample English-German translations

src	″ We ′ re pleased the FAA recognizes that an enjoyable passenger experience is *not incompatible* with safety and security , ″ said Roger Dow , CEO of the U.S. Travel Association .
ref	“ Wir freuen uns , dass die FAA erkennt , dass ein angenehmes Passagiererlebnis nicht **im Wider- spruch zur Sicherheit steht** ” , sagte Roger Dow , CEO der U.S. Travel Association .
be st	″ 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 .
ba se	″ 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 German-English translations

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

- Translate well long sentences.

Conclusion

- Two effective attentional mechanisms:
  - Global and **local** attention
  - State-of-the-art results in WMT English-German.

- Detailed analysis:
  - Better in translating names.
  - Handle well long sentences.
  - Achieve competitive AERs.

- Thank you!