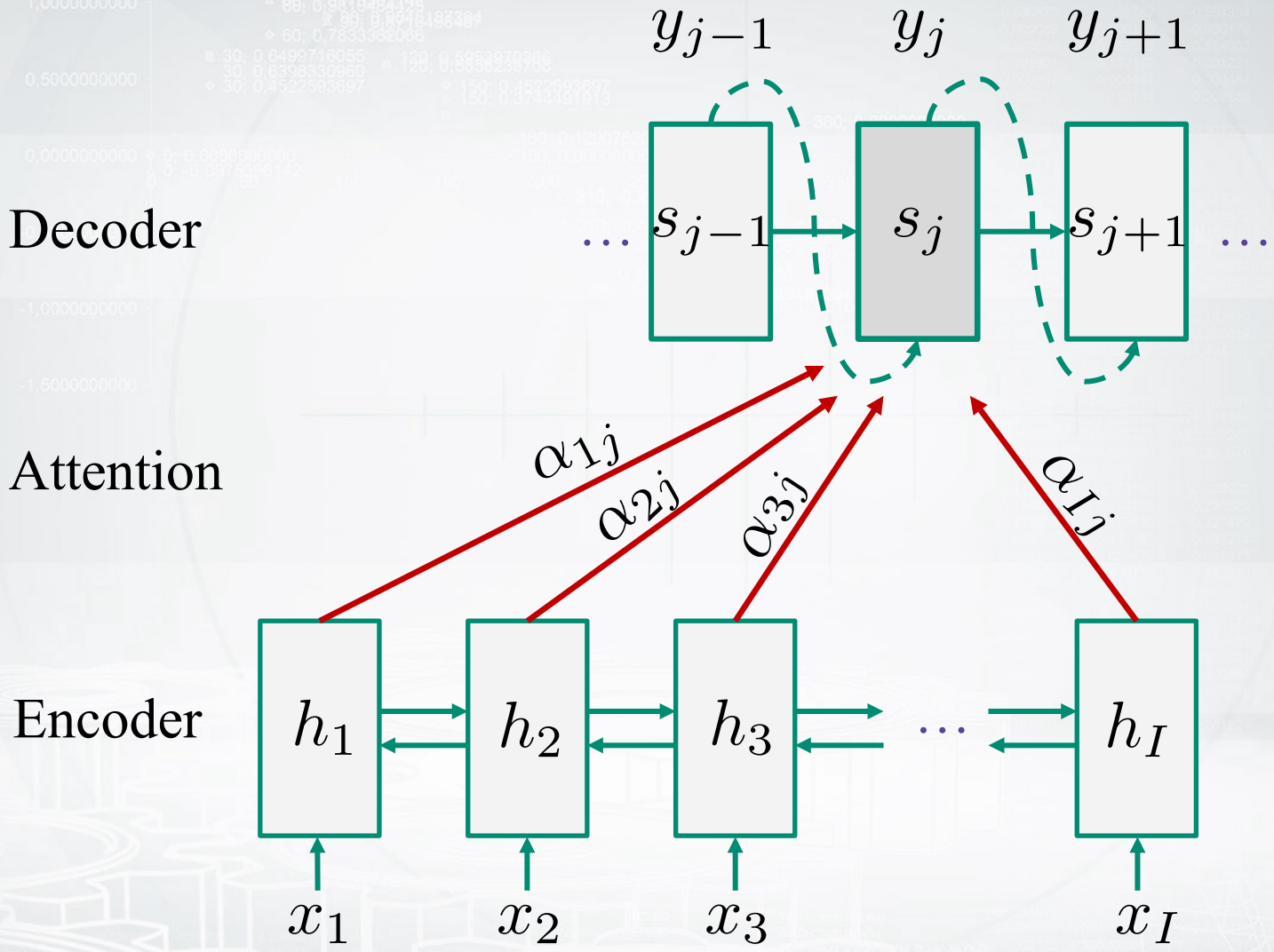
Attention mechanism

Attention mechanism



Bahdanau et. al - Neural Machine Translation by jointly learning to align and translate, 2015.

Attention mechanism

- Encoder states are weighted to obtain the representation relevant to the decoder state:

$$v_j = \sum_{i=1}^{I} \alpha_{ij} h_i$$

- The weights are learnt and should find the most relevant encoder positions:

$$\alpha_{ij} = \frac{\exp(sim(h_i, s_{j-1}))}{\sum_{i'=1}^{I} \exp(sim(h_{i'}, s_{j-1}))}$$

How to compute attention weights?

- **Additive attention:**

$$sim(h_i, s_j) = w^T \tanh(W_h h_i + W_s s_j)$$

- **Multiplicative attention:**

$$sim(h_i, s_j) = h_i^T W s_j$$

- **Dot product also works:**

$$sim(h_i, s_j) = h_i^T s_j$$

Put all together

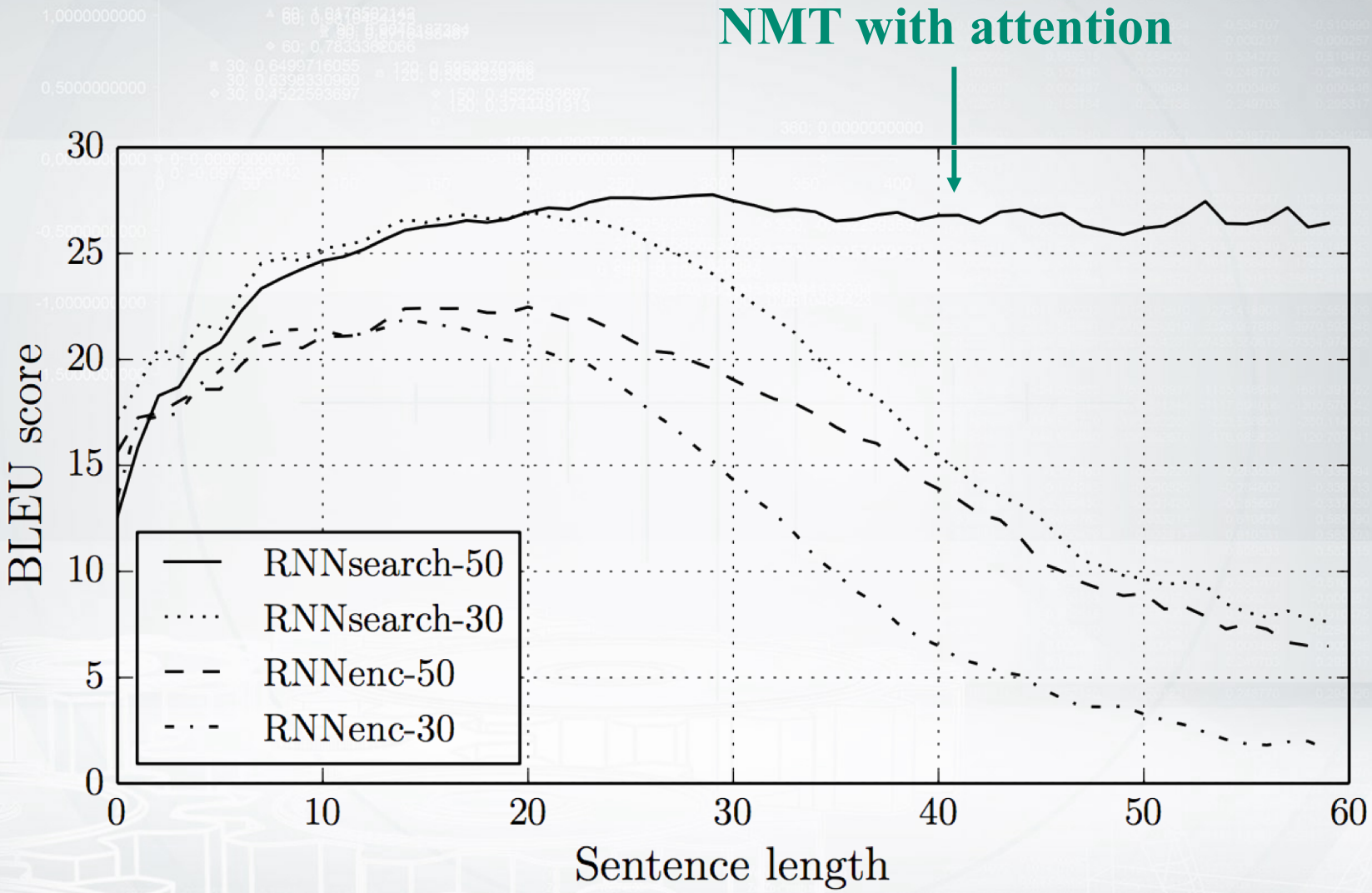
$$p(y_1, \ldots y_J | x_1, \ldots x_I) = \prod_{j=1}^{J} p(y_j | v_j, y_1, \ldots y_{j-1})$$

- Still encoder-decoder architecture with RNNs:

$$h_i = f(h_{i-1}, x_i) \qquad s_j = g(s_{j-1}, [y_{j-1}, v_j])$$

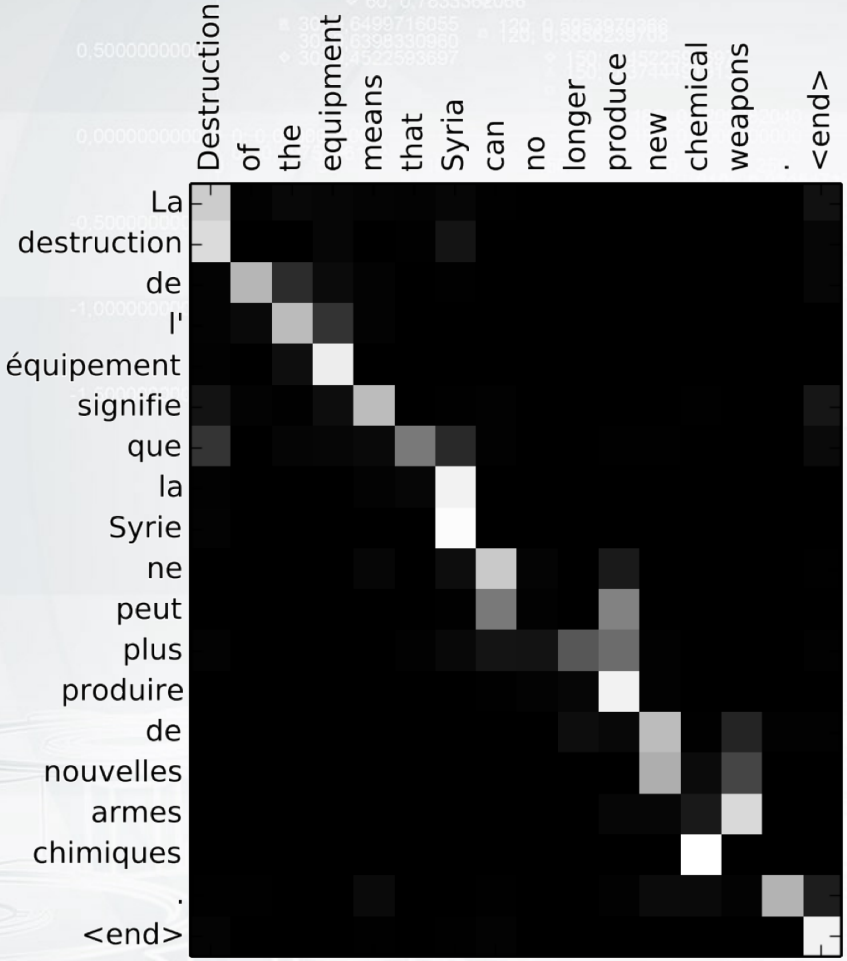
- But the source representations differ for each position j of the decoder.

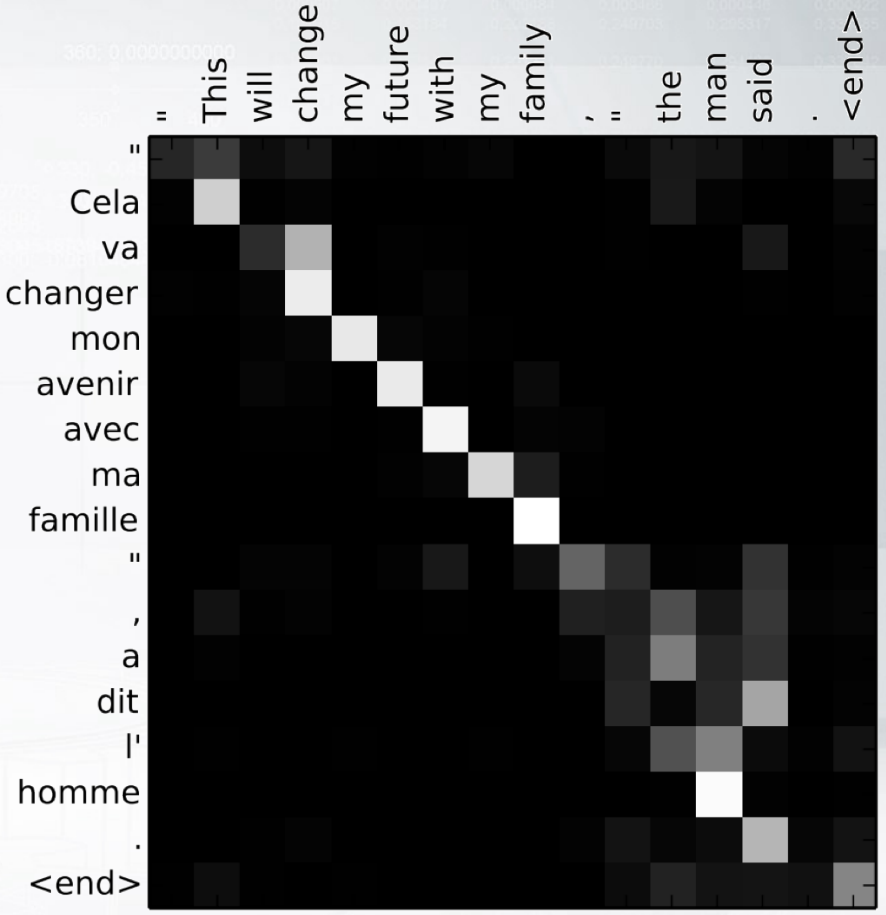
Helps for long sentences



Bahdanau et. al. Neural Machine Translation by jointly learning to align and translate, 2015.

Example: attention (alignments)





Bahdanau et. al. Neural Machine Translation by jointly learning to align and translate, 2015.

Is the attention similar to what humans do?

- *For humans:* **saves time**

Attention saves time when reading (i.e. we look only to the relevant parts of the sentence).

- *For machines:* **wastes time**

To compute the attention weights, the model carefully examines ALL the positions, thus wastes even more time.

Local attention

1. Find the most relevant position a_j in the source

- Monotonic alignments: $a_j = j$
- Predictive alignments: $a_j = I \cdot \sigma(b^T \tanh(W s_j))$

2. Attend only positions within a window $[a_j - h; a_j + h]$

- Compute scores as usual
- Probably multiply by a Gaussian centered in a_j

Luong et. al. Effective Approaches to Attention-based Neural Machine Translation, 2015.

Global vs local attention

System	Perplexity	BLEU
global (location)	6.4	19.3
global (dot)	6.1	20.5
global (mult)	6.1	19.5
local-m (dot)	>7.0	x
local-m (mult)	6.2	20.4
local-p (dot)	6.6	19.6
local-p (mult)	**5.9**	**20.9**

Luong et. al. Effective Approaches to Attention-based Neural Machine Translation, 2015.

Global vs local attention

	System	Perplexity	BLEU
$W s_j \rightarrow$	global (location)	6.4	19.3
$h_i^T s_j \rightarrow$	global (dot)	6.1	20.5
$h_i^T W s_j \rightarrow$	global (mult)	6.1	19.5
	local-m (dot)	>7.0	x
	local-m (mult)	6.2	20.4
	local-p (dot)	6.6	19.6
	local-p (mult)	**5.9**	**20.9**

Luong et. al. Effective Approaches to Attention-based Neural Machine Translation, 2015.