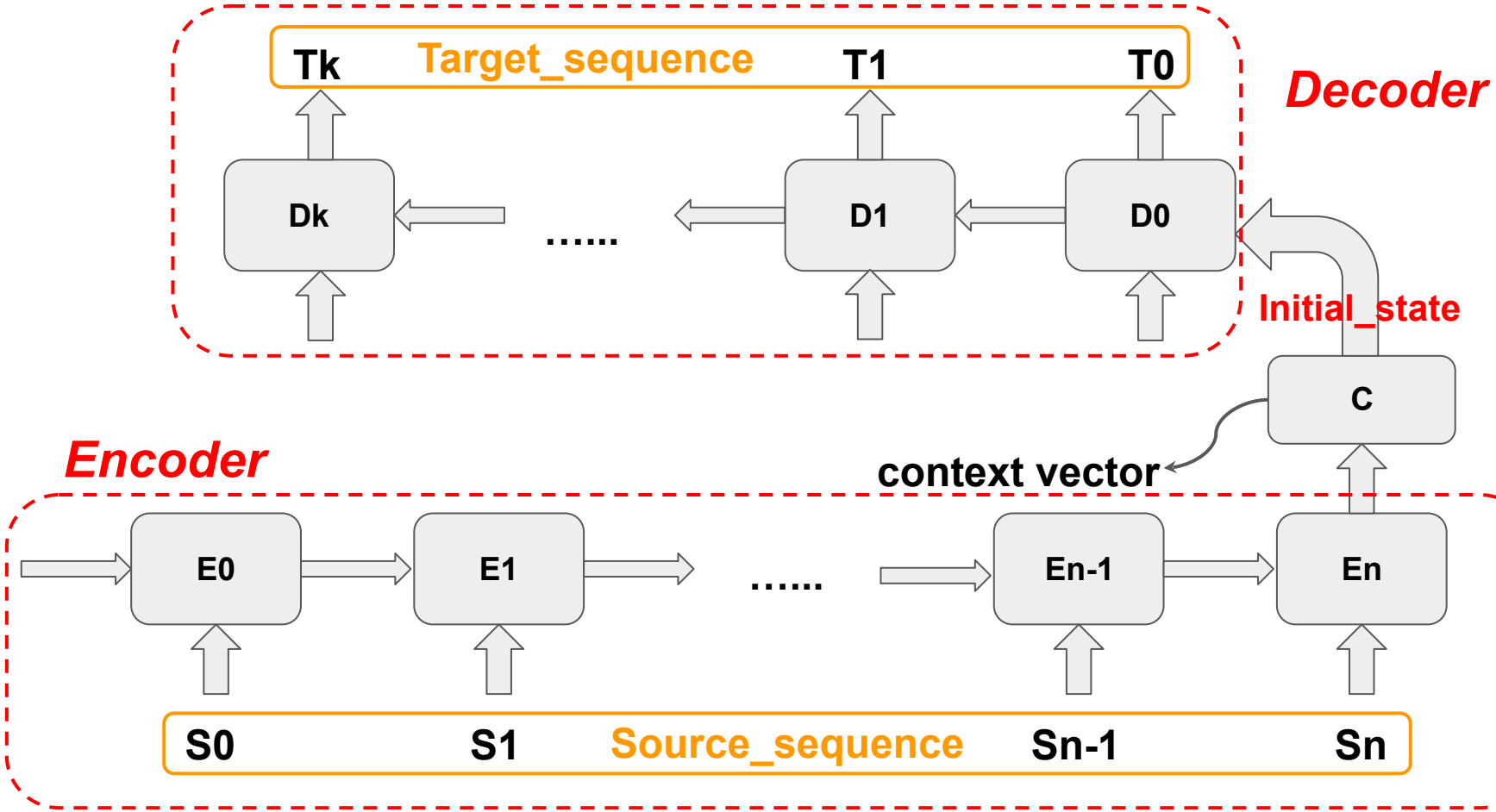
Seq2seq

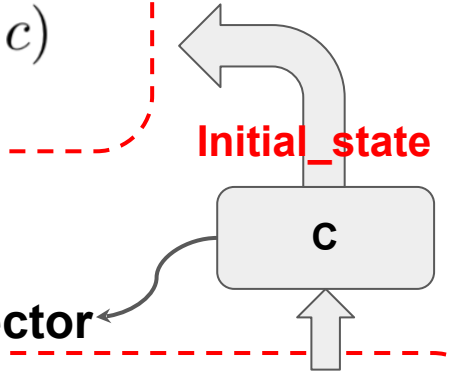

Tk
Target_sequence
T1
T0
Decoder
Dk
.....
D1
D0
Initial_state
C
context vector
Encoder
E0
E1
.....
En-1
En
S0
S1
Source_sequence
Sn-1
Sn

Seq2seq

Decoder

- **Same weight for each target sentence**

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$



Encoder

- **Encode each source sentence into a fix-length context vector**

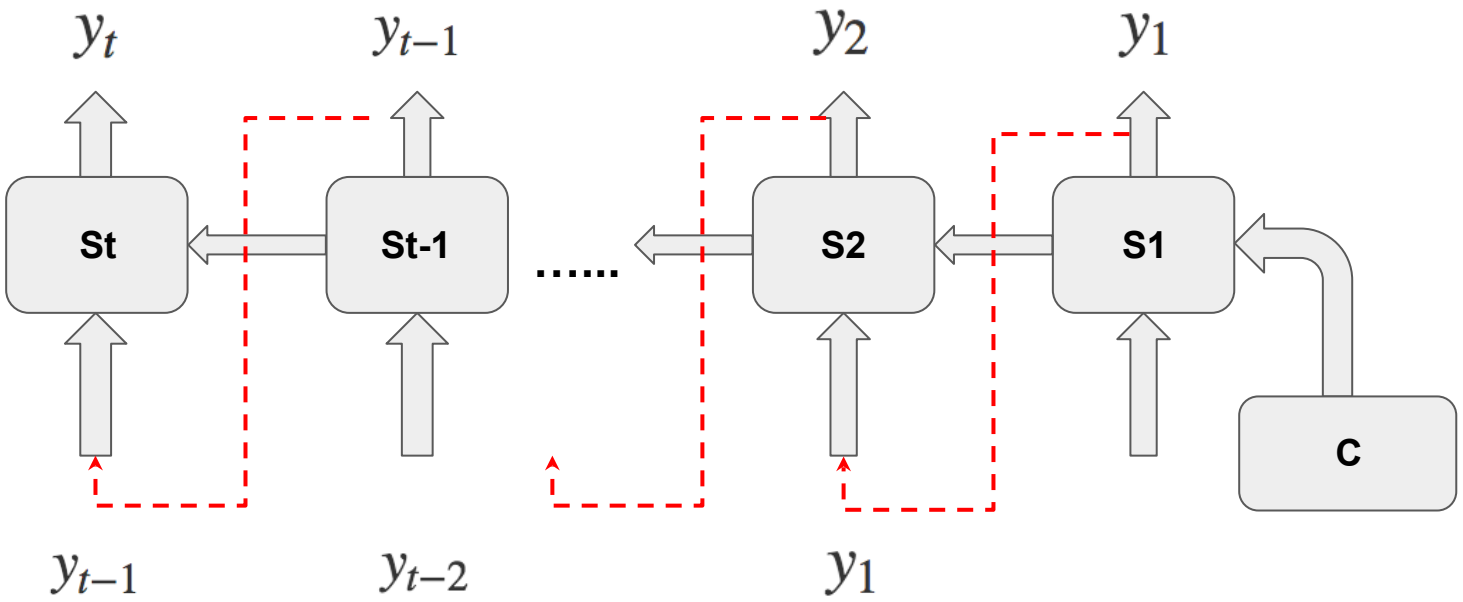
$$h_t = f\left(x_t, h_{t-1}\right)$$
$$c = q\left(\{h_1, \cdots, h_{T_x}\}\right)$$

Seq2seq

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, c) = g(y_{t-1}, s_t, c)$$

$$s_t = \tanh(Uy_{t-1} + Ws_{t-1} + b)$$

$$y_t = \text{sigmoid}(Vs_t)$$



NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE

Dzmitry Bahdanau
Jacobs University Bremen, Germany

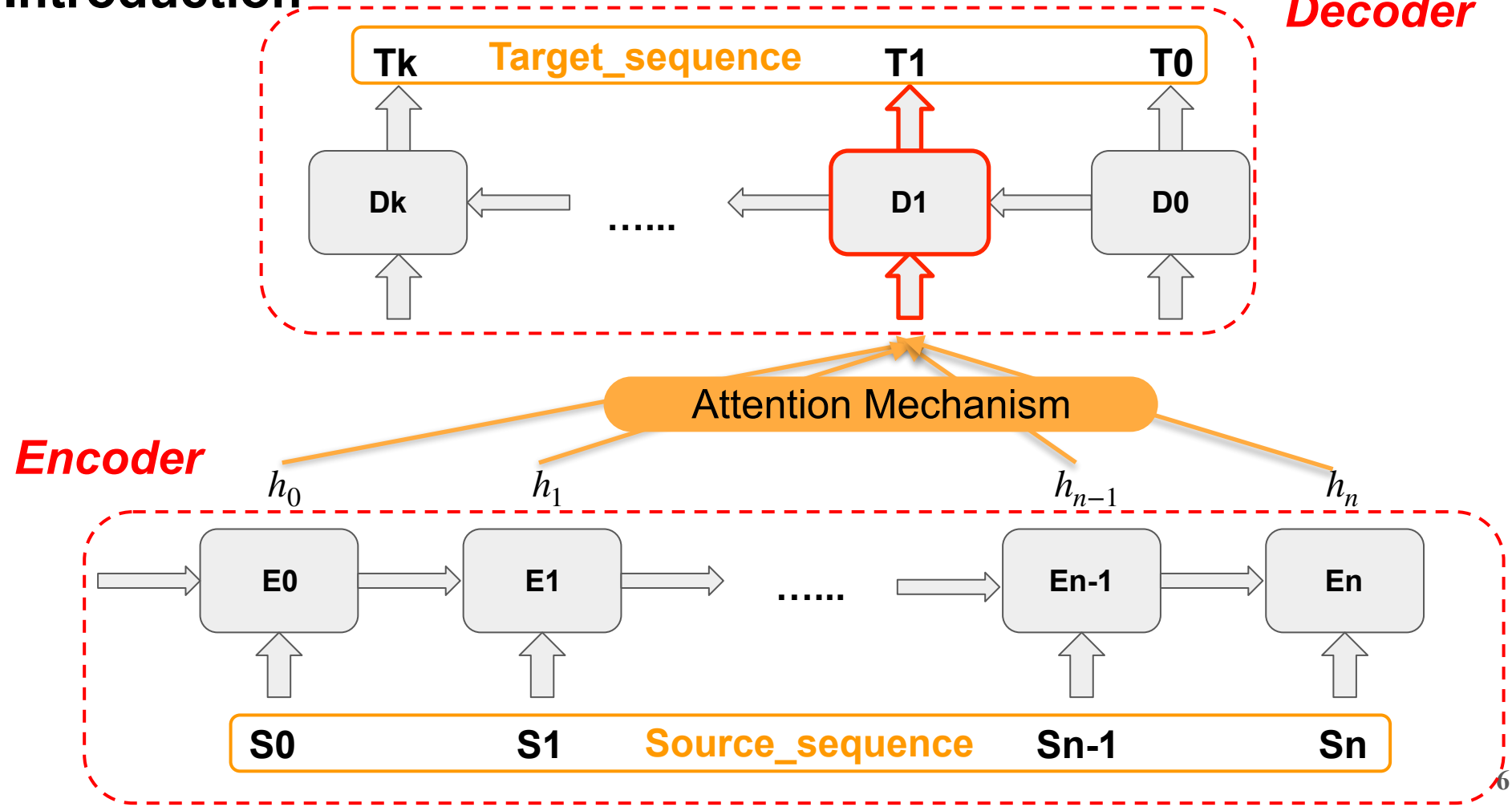
KyungHyun Cho **Yoshua Bengio***
Université de Montréal

https://arxiv.org/abs/1409.0473

Kang Wen Wei

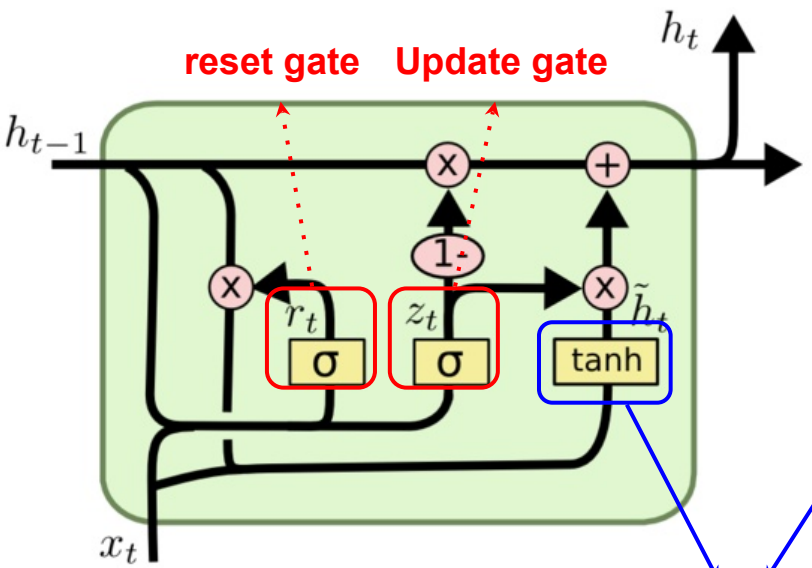
- Introduction
- Related work
- Attention mechanism
- Datasets
- Comparison
- Evaluation

Introduction



Related work

GRU - Gated Recurrent Unit



Reset gate Update gate

$$z_t = \sigma\left(x_t U^z + h_{t-1} W^z\right)$$

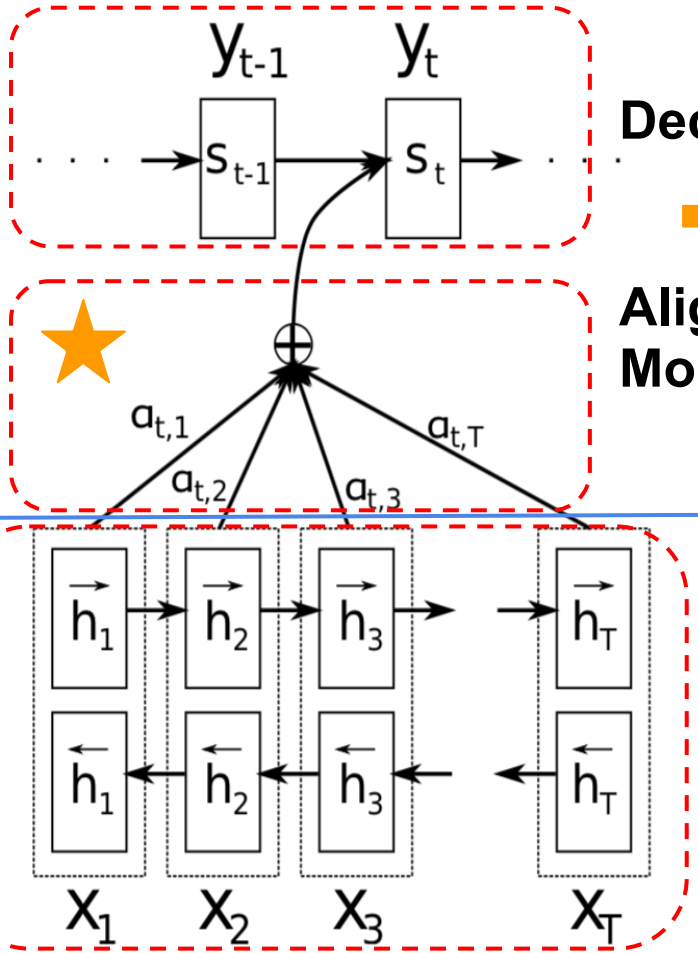
$$r_t = \sigma\left(x_t U^r + h_{t-1} W^r\right)$$

$$\tilde{h}_t = \tanh\left(x_t U^h + (r_t * h_{t-1}) W^h\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

State candidate Current State

Attention Mechanism



Decoder

Alignment Model

$$p(y_i|y_1,\ldots,y_{i-1},\mathbf{x}) = g(y_{i-1}, s_i, c_i)$$

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$$\mathbf{x} = (x_1,\ldots,x_{T_x}),\ x_i \in \mathbb{R}^{K_x}$$

1. $e_{ij} = a(s_{i-1}, h_j)$

2. $\alpha_{ij} = \dfrac{\exp(e_{ij})}{\sum_{k=1}^{T_x}\exp(e_{ik})}$

3. $c_i = \sum_{j=1}^{T_x}\alpha_{ij}h_j$

Encoder

$$\overrightarrow{h}_i = \begin{cases} (1-\overrightarrow{z}_i)\circ\overrightarrow{h}_{i-1} + \overrightarrow{z}_i \circ \underline{\overrightarrow{h}}_i & , \text{if } i > 0 \\ 0 & , \text{if } i = 0 \end{cases}$$

$\underline{\overrightarrow{h}}_i$: State candidate

$\overrightarrow{z}_i$: Update gate

$\overrightarrow{h}_i$: Current state

Attention Mechanism

Bi-directional GRU (Forward)

$$\overrightarrow{h}_i = \begin{cases} (1 - \overrightarrow{z}_i) \circ \overrightarrow{h}_{i-1} + \overrightarrow{z}_i \circ \underline{\overrightarrow{h}}_i & , \text{if } i > 0 \\ 0 & , \text{if } i = 0 \end{cases}$$

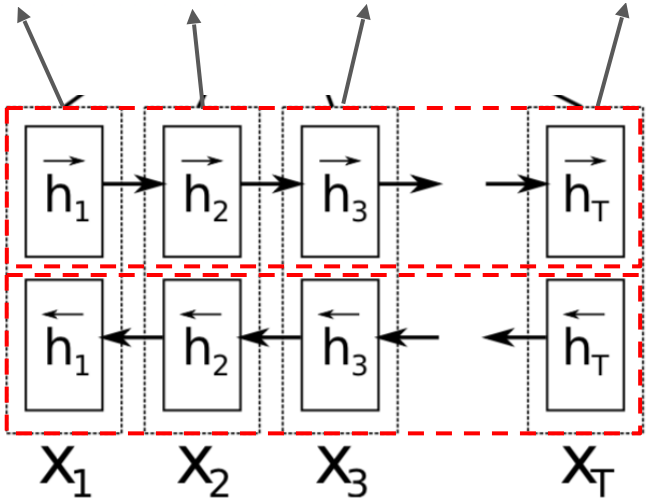
$\circ$: **Element-wise Multiplication**

$$h_1 = \begin{bmatrix} \overrightarrow{h}_1 \\ \overleftarrow{h}_1 \end{bmatrix} \quad h_2 = \begin{bmatrix} \overrightarrow{h}_2 \\ \overleftarrow{h}_2 \end{bmatrix} \quad h_3 = \begin{bmatrix} \overrightarrow{h}_3 \\ \overleftarrow{h}_3 \end{bmatrix} \quad h_T = \begin{bmatrix} \overrightarrow{h}_T \\ \overleftarrow{h}_T \end{bmatrix}$$

$$\begin{cases} \overrightarrow{h}_i, \overleftarrow{h}_i \in \mathbb{R}^{n \times Kx} \\ h_i \in \mathbb{R}^{2n \times Kx} \end{cases}$$

n : Number of hidden units

Kx : Vocabulary sizes of source language



Forward $\{x_1, x_2, \ldots, x_{Tx-1}, x_{Tx}\}$

Backward $\{x_{Tx}, x_{Tx-1}, \ldots, x_2, x_1\}$

Attention Mechanism

$$\overrightarrow{h}_i = \begin{cases} (1 - \overrightarrow{z}_i) \circ \overrightarrow{h}_{i-1} + \overrightarrow{z}_i \circ \underline{\overrightarrow{h}}_i & , \text{if } i > 0 \\ 0 & , \text{if } i = 0 \end{cases}$$

$$\overrightarrow{r}_i = \sigma\left(\overrightarrow{W}_r \overline{E} x_i + \overrightarrow{U}_r \overrightarrow{h}_{i-1}\right)$$

$$\overrightarrow{z}_i = \sigma\left(\overrightarrow{W}_z \overline{E} x_i + \overrightarrow{U}_z \overrightarrow{h}_{i-1}\right)$$

$$\underline{\overrightarrow{h}}_i = \tanh\left(\overrightarrow{W} \overline{E} x_i + \overrightarrow{U}\left[\overrightarrow{r}_i \circ \overrightarrow{h}_{i-1}\right]\right)$$

$\overrightarrow{z}_i$: **Update gate**

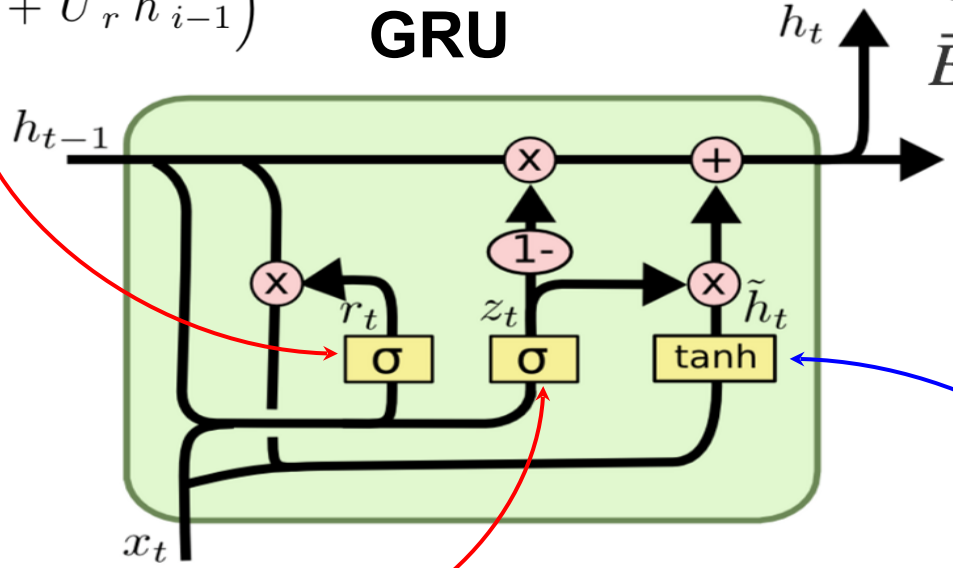
$\overrightarrow{r}_i$: **Reset gate**

σ : **Sigmoid**

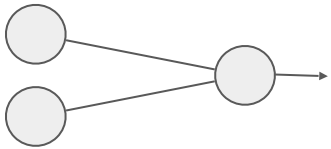
$\underline{\overrightarrow{h}}_i$: **State candidate**

$\bar{E}x_i$: **Word Embedding matrix**
Weight sharing

GRU



Attention Mechanism



1. $e_{ij} = a(s_{i-1}, h_j)$ **Soft-Alignment**

$$e_{ij} = v_a^\top \tanh\left(W_a s_{i-1} + U_a h_j\right)$$

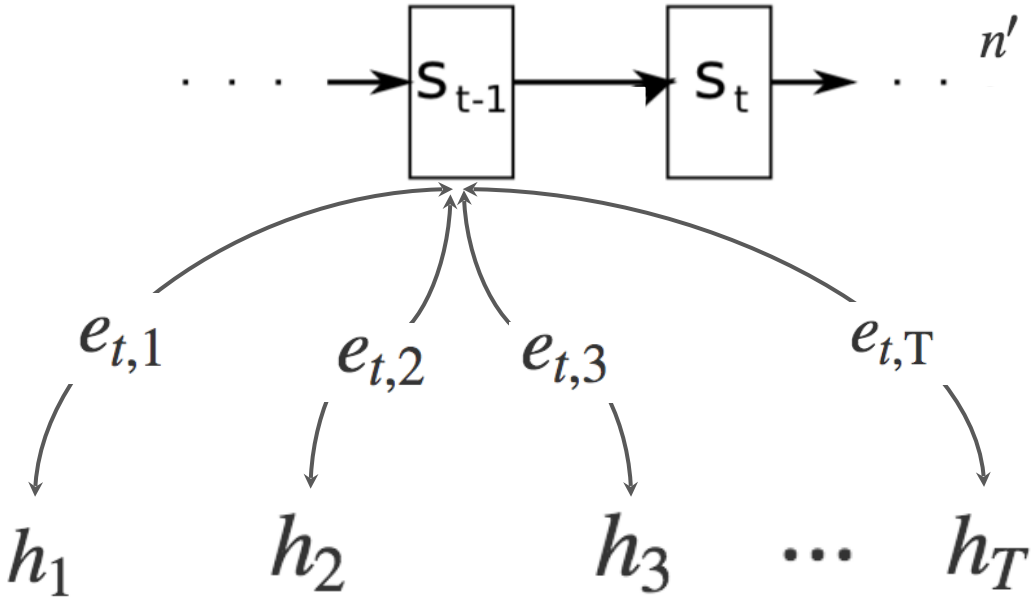
$W_a \in \mathbb{R}^{n' \times n}$

$U_a \in \mathbb{R}^{n' \times 2n}$

$v_a \in \mathbb{R}^{n'}$

n : Number of hidden units

n' : Number of Alignment hidden size

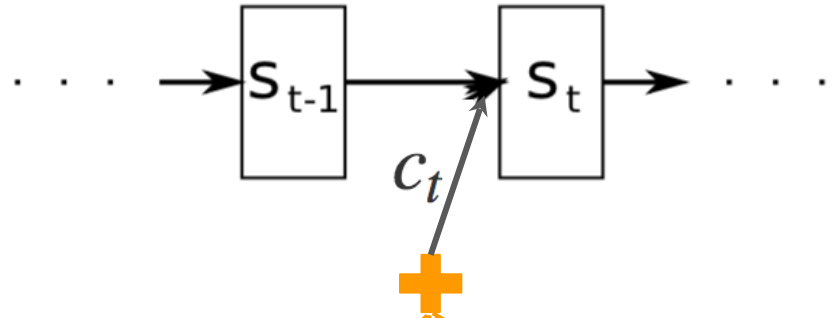


$e_{t,1}$

How Match between 1-th source word and t-th target word

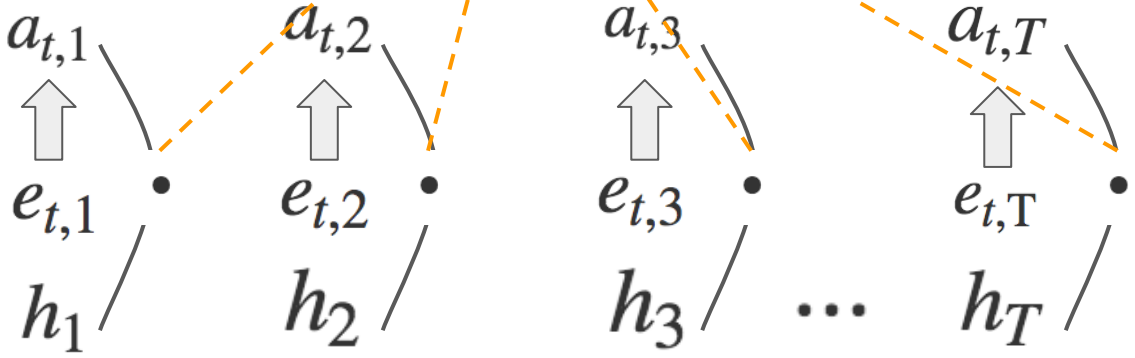
Attention Mechanism

3. $c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$



Regard α_{ij} as weight for each h

2. $\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$

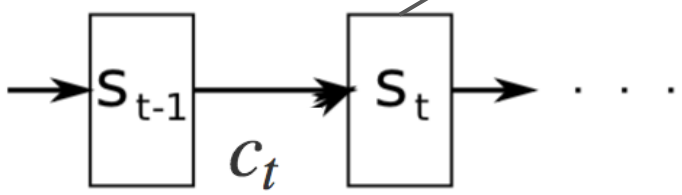


• : Dot product

Attention Mechanism

Output layer can be freely define

$$s_0 = \tanh(W_s \overleftarrow{h}_1)$$
$$W_s \in \mathbb{R}^{n \times n}$$



$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$
$$= (1 - z_i) \circ s_{i-1} + z_i \circ \tilde{s}_i$$

$$\tilde{s}_i = \tanh\left(WEy_{i-1} + U\left[r_i \circ s_{i-1}\right] + Cc_i\right)$$
$$z_i = \sigma\left(W_z Ey_{i-1} + U_z s_{i-1} + C_z c_i\right)$$
$$r_i = \sigma\left(W_r Ey_{i-1} + U_r s_{i-1} + C_r c_i\right)$$

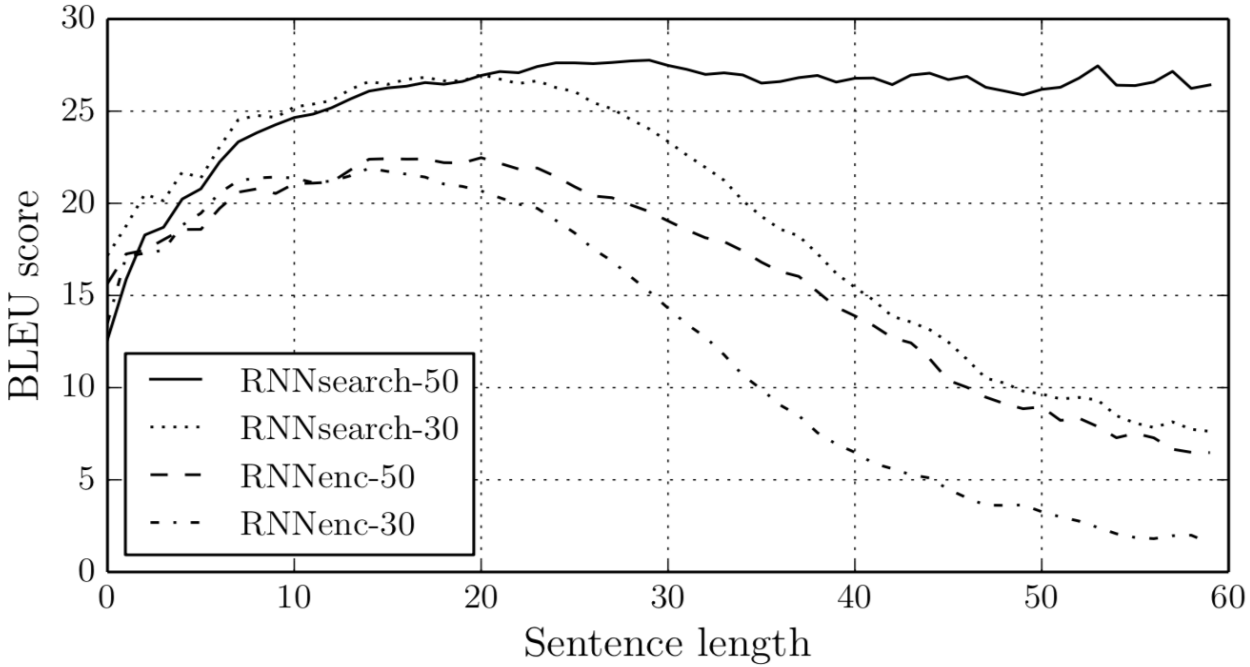
$$U, U_z, U_r \in \mathbb{R}^{n \times n}$$
$$C, C_z, C_r \in \mathbb{R}^{n \times 2n}$$
$$W, W_z, W_r \in \mathbb{R}^{n \times m}$$

Be a little different from GRU

Comparison

Training : Sentence of length up to 30 words (RNNencdec-30, RNNsearch-30)
Sentence of length up to 50 words (RNNencdec-50, RNNsearch-50)

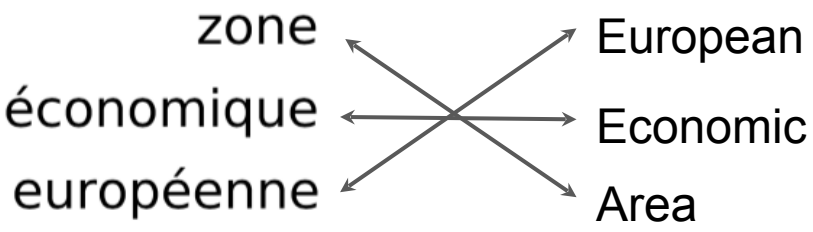
Testing : On the full test set which includes sentences having unknow words

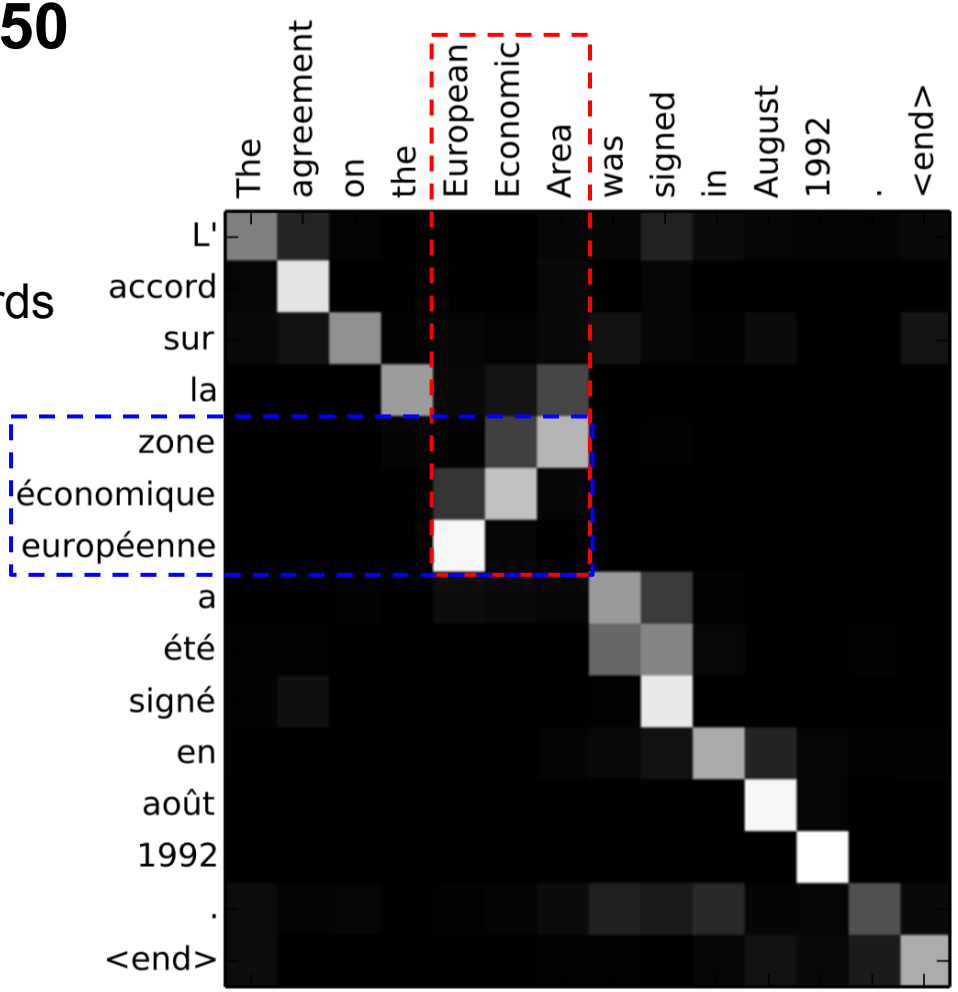


Performance on RNNsearch-50

Cond：Arbitary sentence on test set

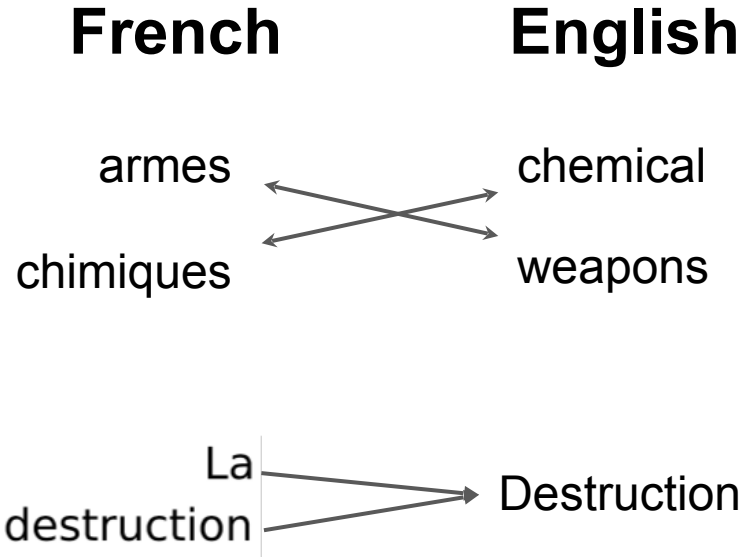
Includes sentences having unknown words

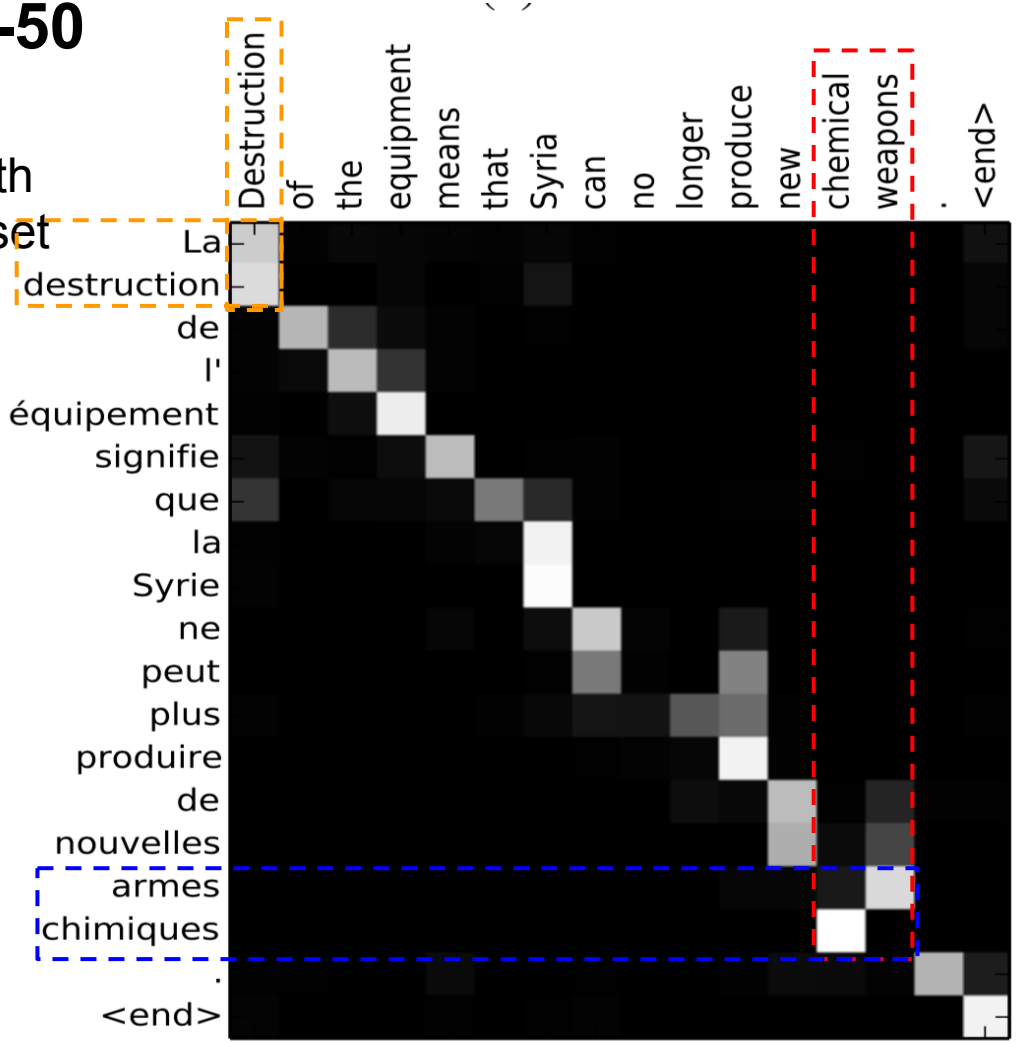




Performance on RNNsearch-50

Cond :
Without any unknown words and of length between 10 and 20 words from the test set





Performance

BLEU scores of the models computed on the test set

All：All the test set

No UNK：Without any unknown words on test set

RNNsearch-50*：Train longer until the performance on the valid stopped improving

Model	All	No UNK°
RNNencdec-30	13.93	24.19
RNNsearch-30	21.50	31.44
RNNencdec-50	17.82	26.71
RNNsearch-50	26.75	34.16
RNNsearch-50⋆	28.45	36.15
Moses	33.30	35.63

Moses：http://www.statmt.org/moses/?n=Moses.Overview

Next...

Effective Approaches to Attention-based Neural Machine Translation

Minh-Thang Luong **Hieu Pham** **Christopher D. Manning**
Computer Science Department Stanford University, Stanford, CA, 94305
{lmthang,hyhieu,manning}@stanford.edu

New alignment model

$$\text{score}(\boldsymbol{h}_t, \bar{\boldsymbol{h}}_s) = \begin{cases} \boldsymbol{h}_t^\top \bar{\boldsymbol{h}}_s & \textit{dot} \\ \boldsymbol{h}_t^\top \boldsymbol{W}_{\boldsymbol{a}} \bar{\boldsymbol{h}}_s & \textit{general} \\ \boldsymbol{W}_{\boldsymbol{a}} [\boldsymbol{h}_t; \bar{\boldsymbol{h}}_s] & \textit{concat} \end{cases}$$