

COLX-531: Neural Machine Translation

Muhammad Abdul-Mageed

muhmmad.mageed@ubc.ca

Deep Learning & NLP Lab

The University of British Columbia

Table of Contents

1 Seq2Seq Models in NMT

2 Attention

Machine Translation

- **Phrase-Based Machine Translation (PBMT)**: Many small sub-components that are tuned separately (e.g., Koehn et al. [2003])

Machine Translation

- **Neural Machine Translation (NMT)**: A large neural network that reads a sentence and outputs a translation
- **Encoder-decoder approach** (e.g. Sutskever et al. (2014)):
 - **Encoder** encodes a source sentence into a *fixed-length* vector
 - **Decoder** outputs a translation from the encoded vector
- **Encoder–decoder system** **jointly trained** to maximize the probability of a correct translation given a source sentence

MT as Seq2Seq Learning: English-French



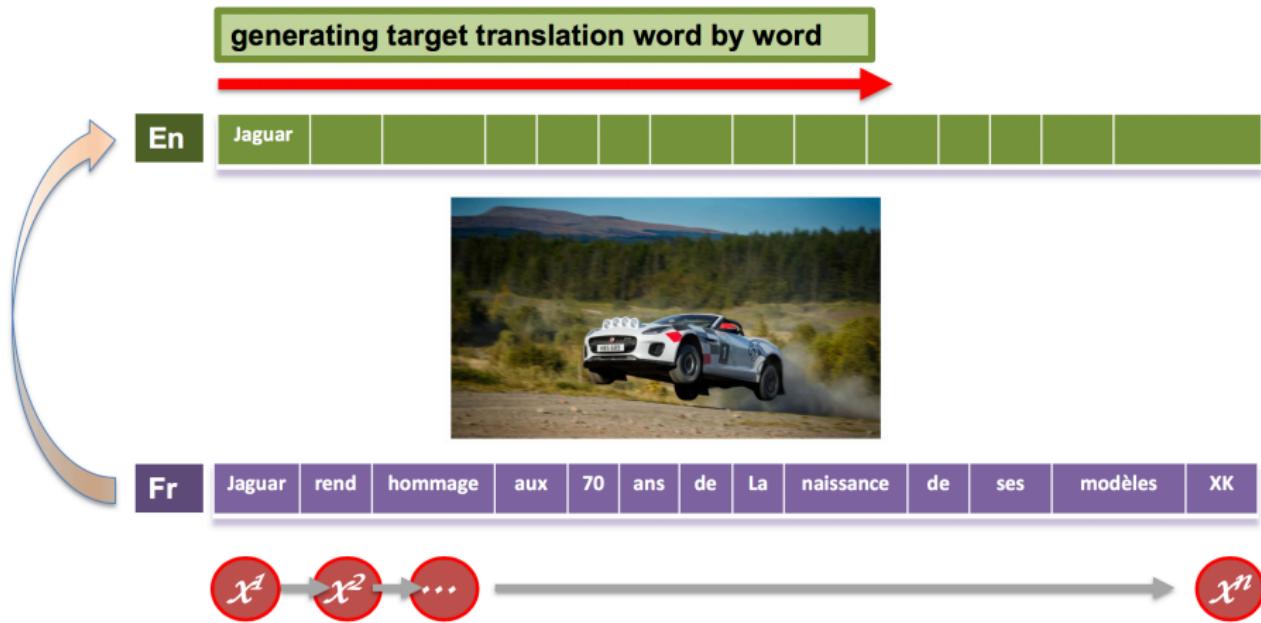
En Jaguar | pays | homage | to | the | 70 | years | of | the | birth | of | its | XK | models

Fr Jaguar | rend | hommage | aux | 70 | ans | de | La | naissance | de | ses | modèles | XK

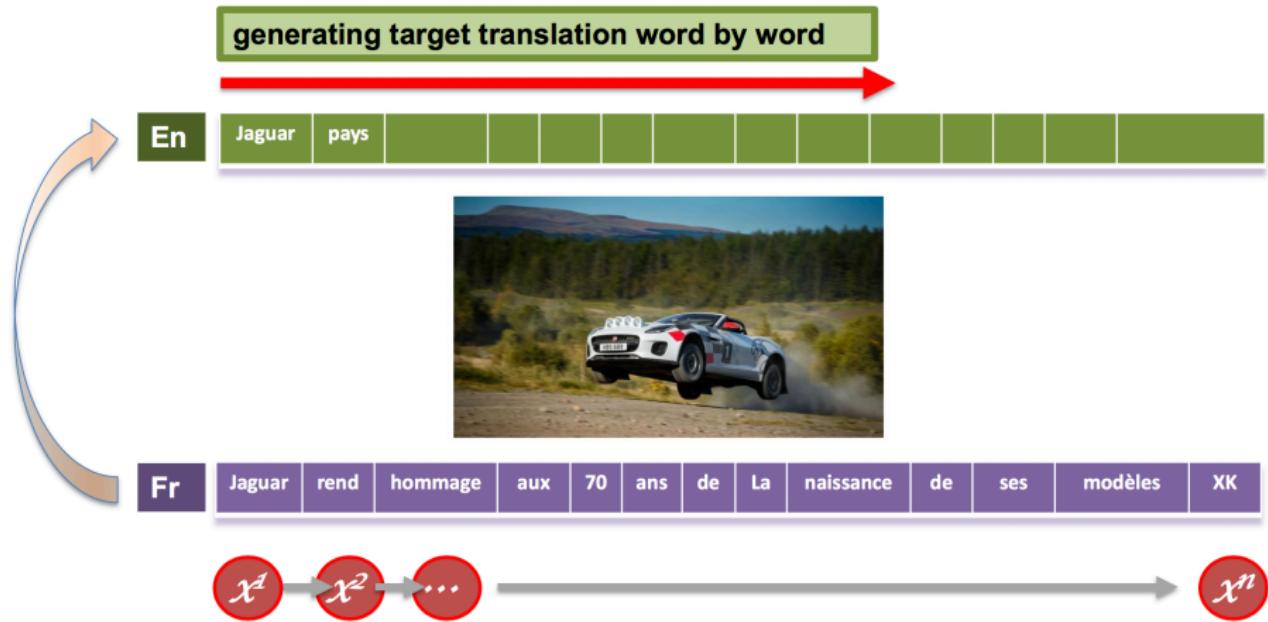
A blue curved arrow points from the English sentence back to the French sentence.

Figure: Translation by Google, November, 20, 2018.

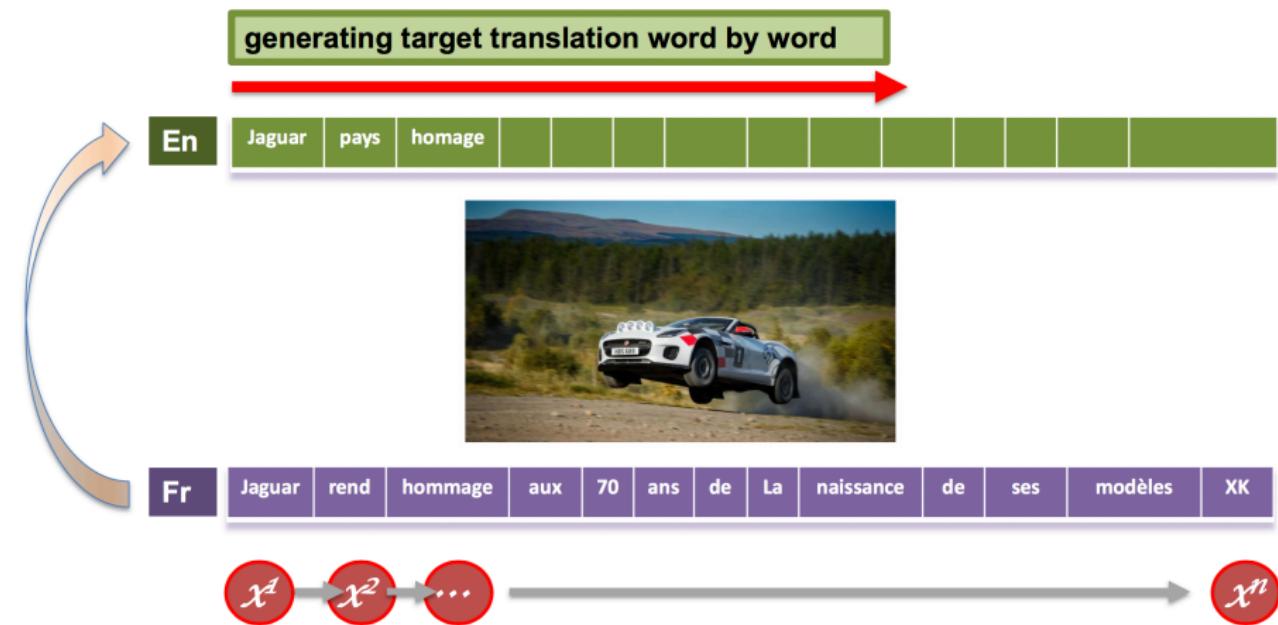
Generating Target (En) One Word at a Time: w_1



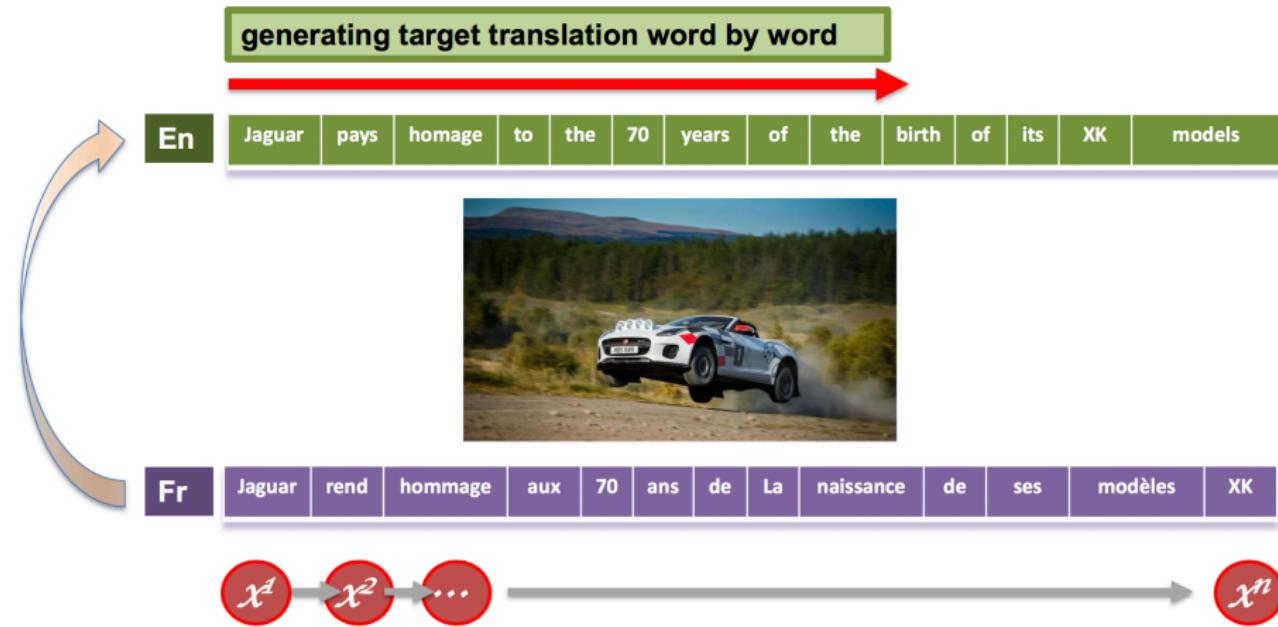
Generating Target (En) One Word at a Time: $w_1 w_2$



Generating Target (En) One Word at a Time: $w_1 w_2 w_3$



Generating Target (En) One Word at a Time: $w_1 w_2 \dots w_n$



1: Encoder

Encoder reads the input sentence \mathbf{x} in to a vector \mathbf{c}

$$\mathbf{x} = (x_1, \dots, x_T)$$

Common to use an **RNN**:

$$h_t = f(x_t, h_{t-1})$$

and

$$\mathbf{c} = q(h_1, \dots, h_{T_x})$$

- where h_t is a hidden state at time step t and \mathbf{c} vector of hidden states, and f and q are nonlinear functions (e.g., LSTM)

Decoder

- Decoder trained to predict next word y_t , given (1) the context vector \mathbf{c} and (2) all the previously predicted words $\{y_1, \dots, y_{t-1}\}$
- i.e., The decoder defines a probability over the translation \mathbf{y} by decomposing the joint probability into the ordered conditionals:

2: Decoder

$$p(\mathbf{y}) = \prod_{t=1}^T p(y_t | \{y_1, \dots, y_{t-1}\}, \mathbf{c})$$

where (\mathbf{y} is gold word):

$$\mathbf{y} = (y_1, \dots, y_{T_y})$$

3: Decoder *Contd.*

With an RNN, each conditional probability is modelled as:

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

where ***g*** is a nonlinear function that outputs the probability of y_t and s_t is the hidden state of the RNN.

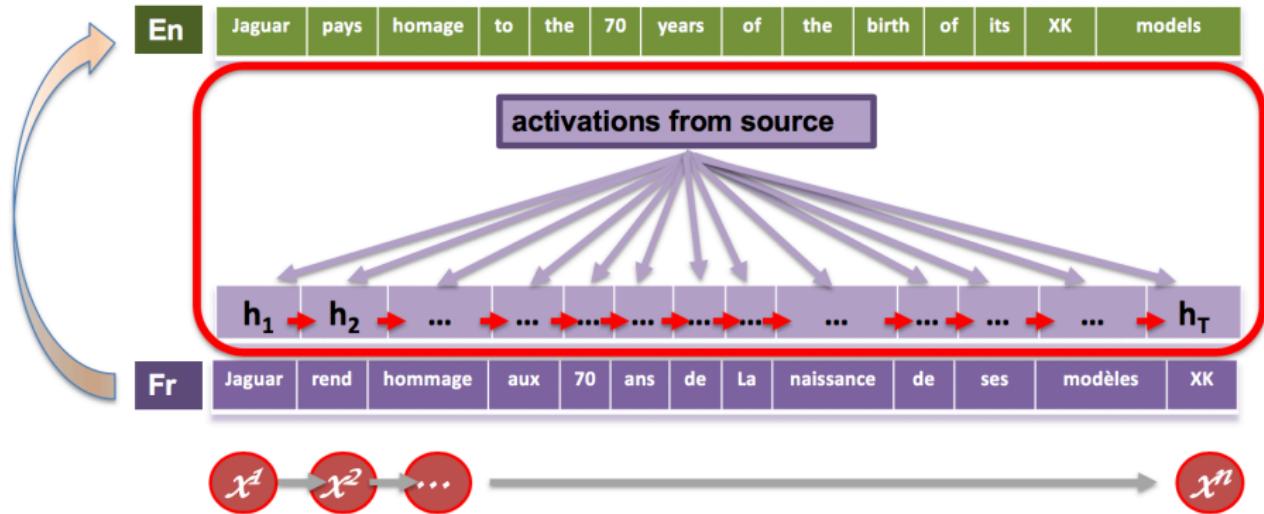
Compressing Long Sentences

- **Issue:** Difficult to compress all the necessary information of a source sentence into a **fixed-length vector**, especially for **long sentences** (Cho et al., 2014)

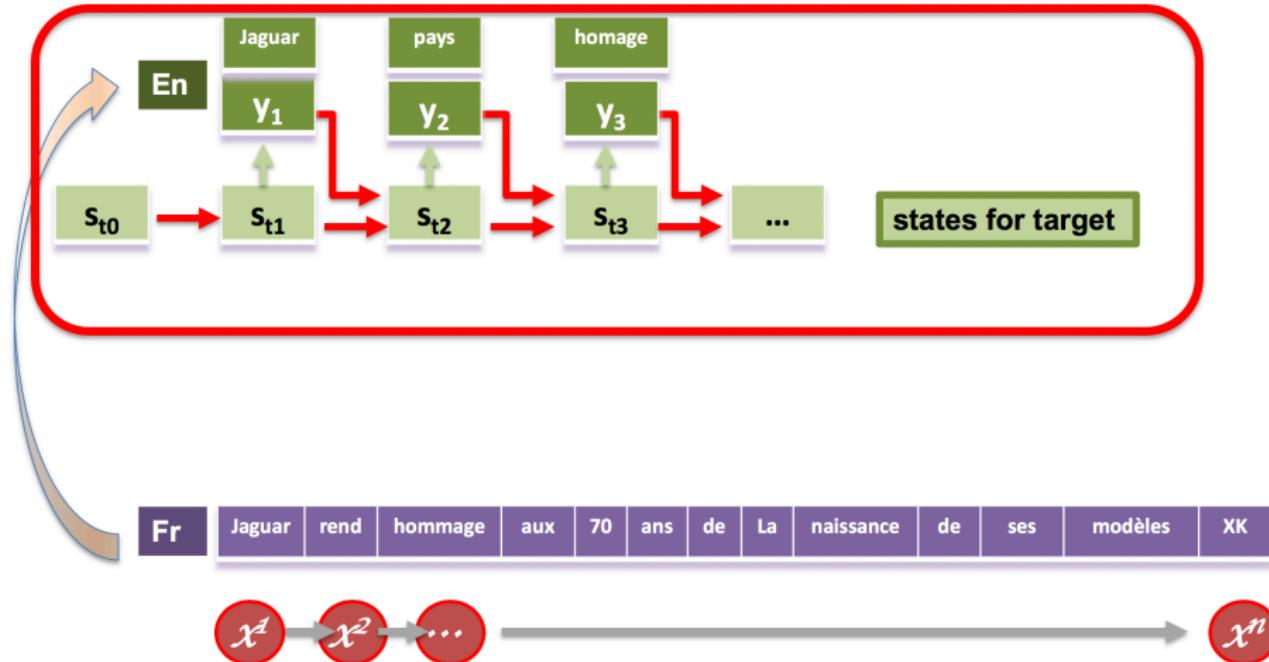
Bahdanau et al. (2015): Learn to ‘align’ and translate jointly

- Generate a word in a translation,
- (**Soft-**)search for a set of positions in a source sentence where the most relevant information is concentrated
- Predict a target word based on the context vectors associated with (1) these source positions and (2) all the previous generated target words
- Copes better with long sentences

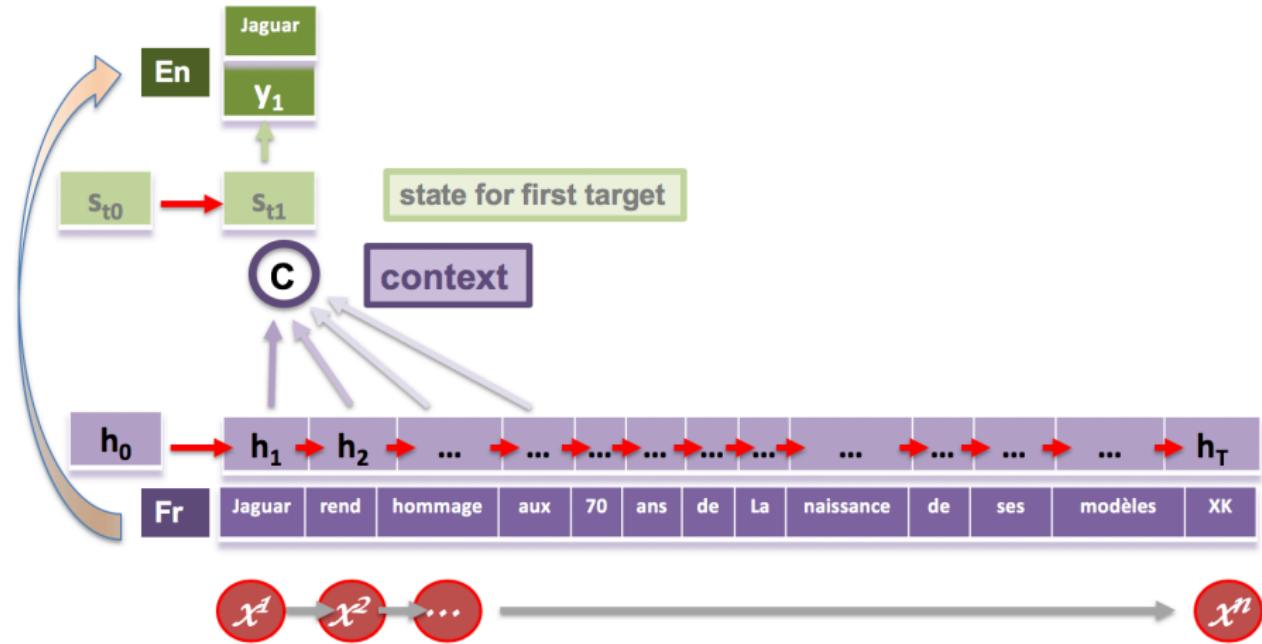
Source Language Activations



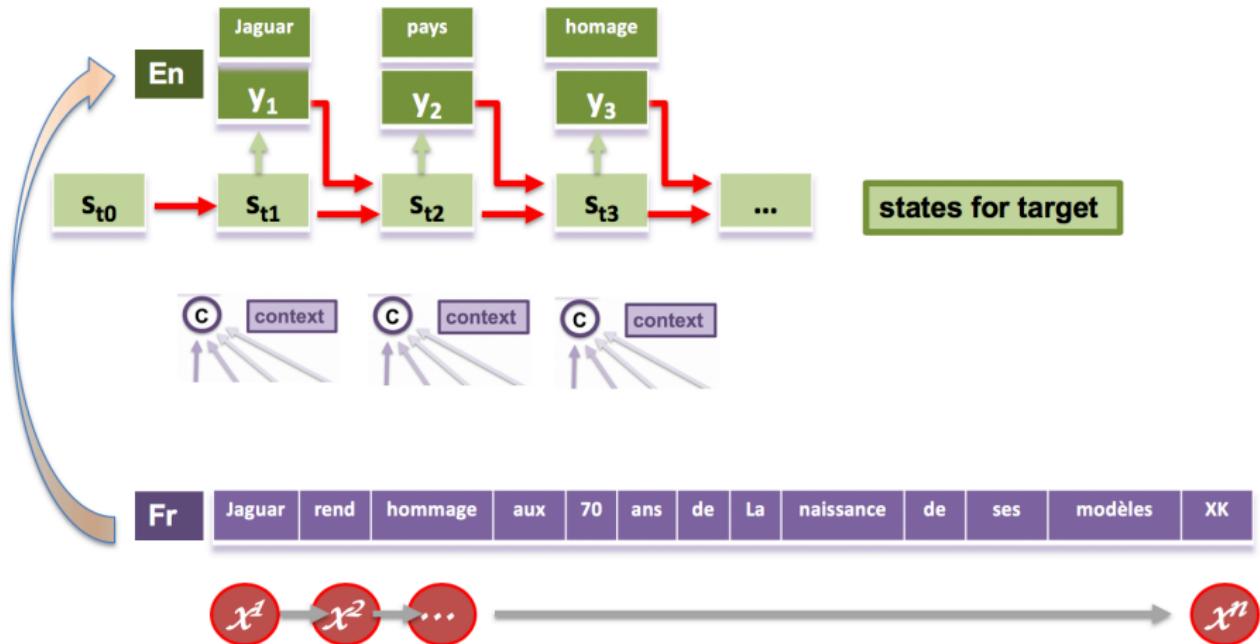
Target (Language) States



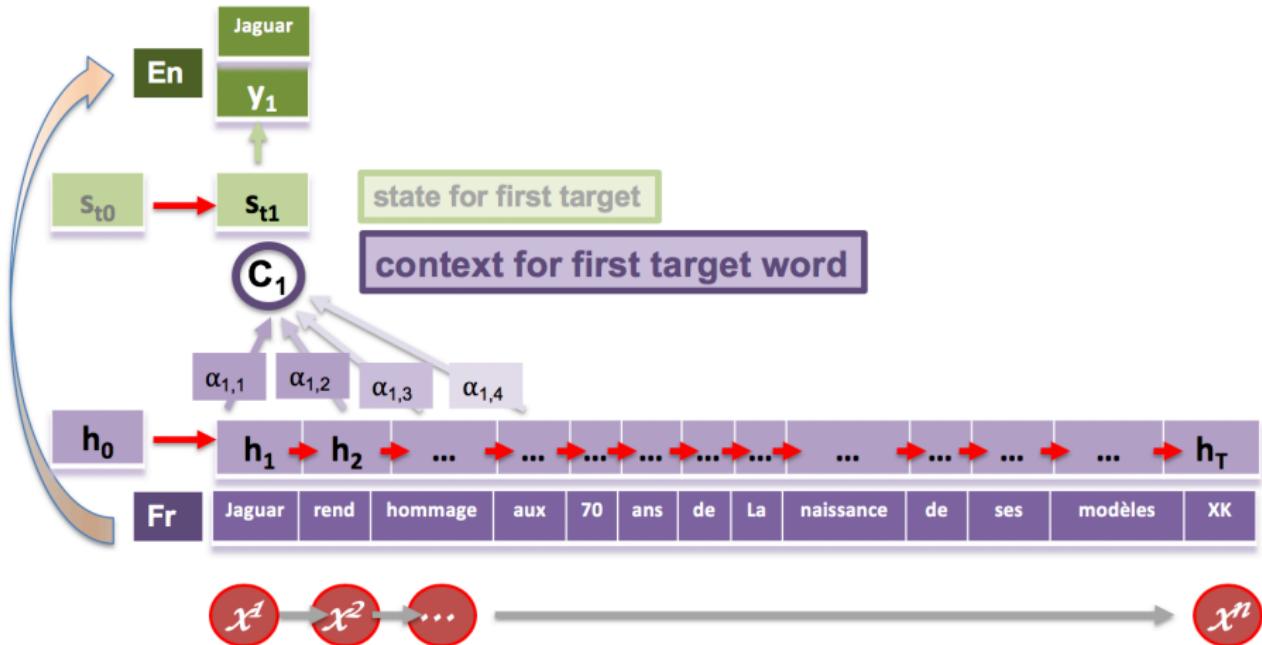
Context



Context Per Every Target State



Context For State One



What is in Context C1?

4: Context C1

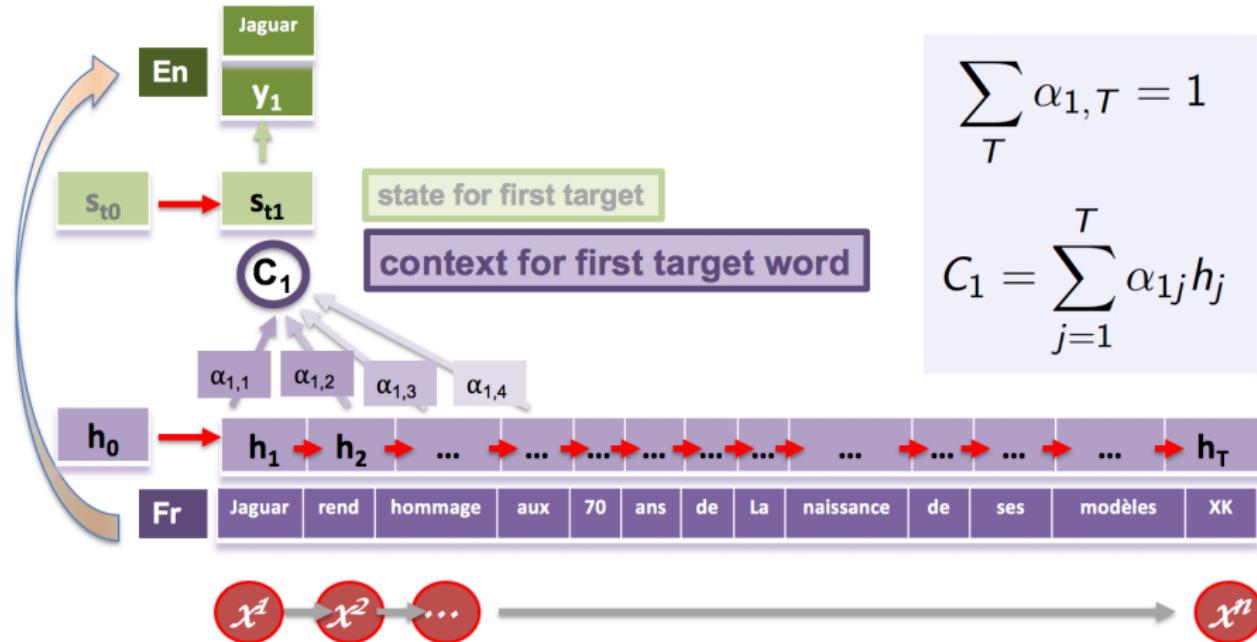
$$\sum_{\tau} \alpha_{1,\tau} = 1$$

$$C_1 = \sum_{j=1}^T \alpha_{1j} h_j$$

Context Ingredients

- α : attention weights
- h : activation from source

Context For State One Enhanced



What is in Context C2?

5: Context C2

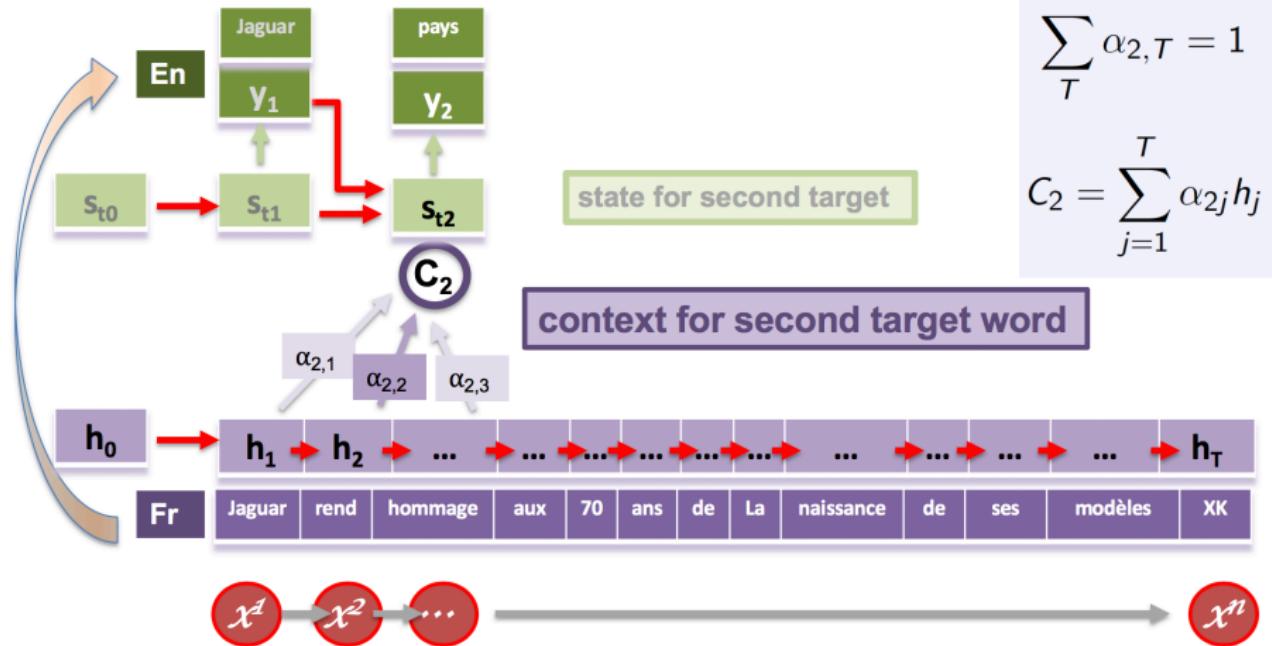
$$\sum_T \alpha_{2,T} = 1$$

$$C_2 = \sum_{j=1}^T \alpha_{2j} h_j$$

Context Ingredients

- α : attention weights
- h : activation from source

Context For State Two



What is in Context C3?

6: Context C3

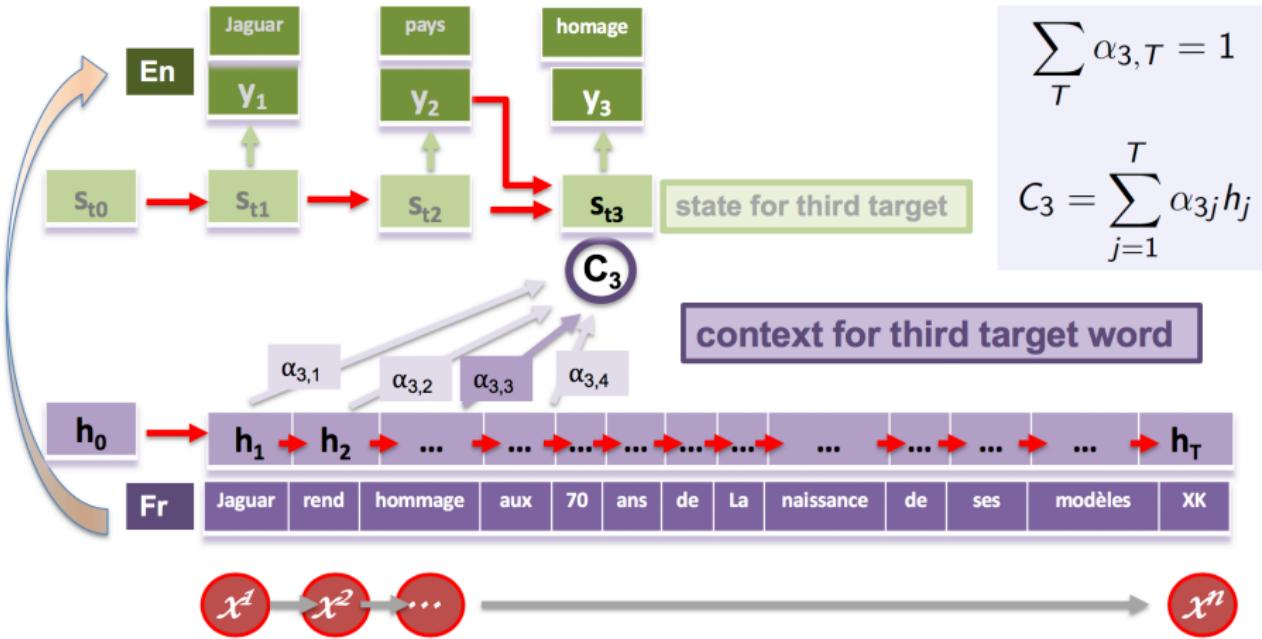
$$\sum_{\tau} \alpha_{3,\tau} = 1$$

$$C_3 = \sum_{j=1}^T \alpha_{3j} h_j$$

Context Ingredients

- α : attention weights
- h : activation from source

Context For State Three



Decoder: Conditioning on Distinct C_i For Each y_i

7: New Decoder

Recall, original decoder:

$$p(\mathbf{y}) = \prod_{t=1}^T p(y_t | \{y_1, \dots, y_{t-1}\}, \mathbf{c})$$

For new decoder, each conditional probability defined as:

$$p(y_i | y_1, \dots, y_{i-1}, \mathbf{x}) = g(y_{i-1}, \mathbf{s}_i, \mathbf{c}_i)$$

where \mathbf{s}_i is an RNN hidden state for time i , computed by:

$$\mathbf{s}_i = f(\mathbf{s}_{i-1}, y_{i-1}, \mathbf{c}_i)$$

- Note: Unlike previous encoder-decoder approach, here the probability is conditioned on a distinct context vector c_i for each target word y_i

What is in Context C?

Context Ingredients

- The context vector c_i depends on a sequence of annotations (h_1, \dots, h_{T_x}) to which the encoder maps the input sequence.

8: Context C

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

Context Ingredients

- α : attention weights
- h : activation from source

Weight α_{ij}

9: Weight α_{ij} for each annotation h_j

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

where

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

is an **alignment model** based on the RNN hidden state s_{i-1} (just before emitting y_i) and the j -th annotation h_j of the input sentence

Alignment Model Job

- **alignment model a** scores how well the **inputs around position j** and the **output at position i** match.

Computing Soft Alignment for Model α

Alignment Model

- The **alignment model** directly computes a soft alignment, which allows the gradient of the cost function to be backpropagated through
- This gradient can be used to train the alignment model as well as the whole translation model jointly

Role of Attention

- "By letting the decoder have an attention mechanism, we relieve the encoder from the burden of having to encode all information in the source sentence into a fixed-length vector. With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly"
(Bahdanau et al., 2015)

Bahdanau et al. (2015) Attention Model Visualization

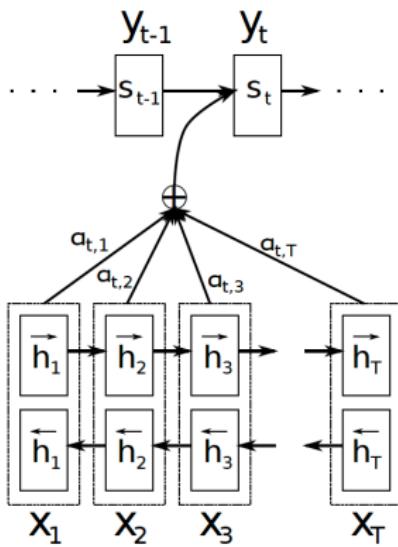


Figure: Bahdanau et al. (2015) attention model trying to generate the t -th target word y_t given a source sentence (x_1, x_2, \dots, x_T)

Bahdanau et al. (2015) Weight Visualization

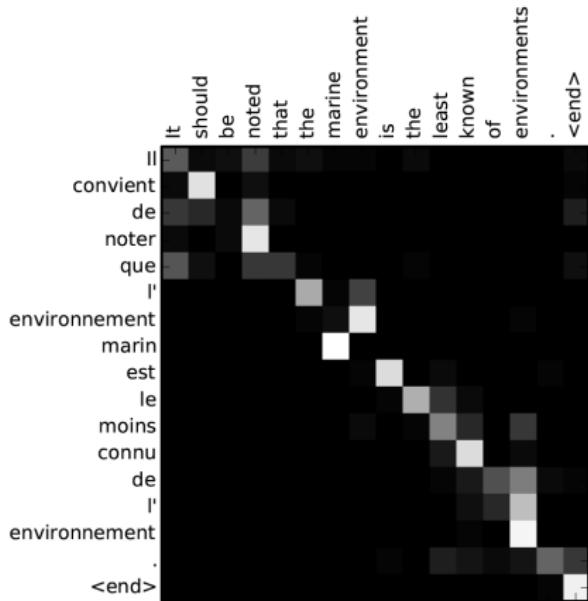
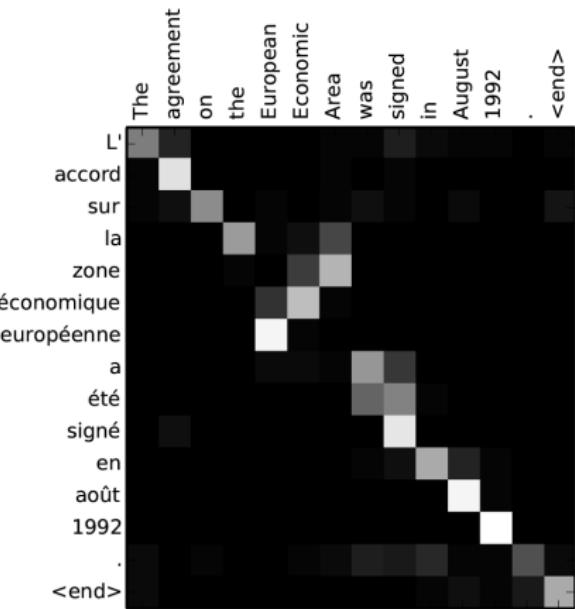


Figure: Each pixel shows the weight ij of the annotation of the j -th source word for the i -th target word (Bahdanau et al., 2015).

Coming Up: Transformer

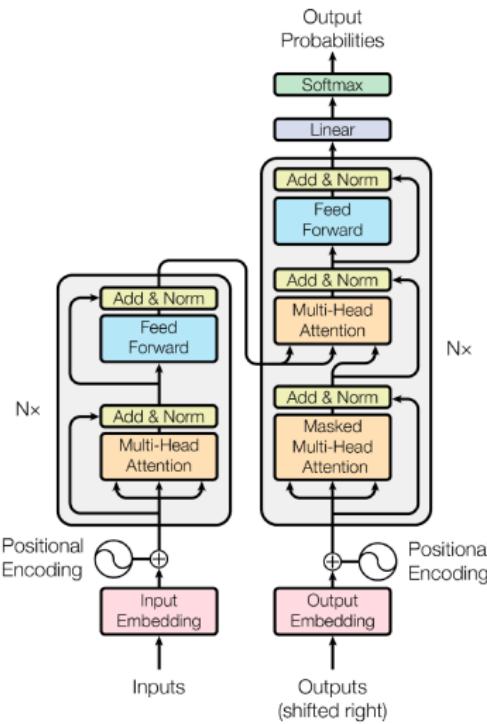


Figure: Transformer. Self-attention is key (Vaswani et al., 2017)