

NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE ‘ICLR 15’

Dzmitry Bahdanau
Jacobs University Bremen

KyungHyun Cho, **Yoshua Bengio***
Universite de Montreal

Presenter: Happy Buzaaba
University of Tsukuba
D2 KDE Member

Supervisor: **Prof Toshiyuki Amagasa**

Date: 2019-04-22

Outline

- Introduction
- Previous work (Basic RNN encoder-decoder)
- Learning to Align and Translate (Attention Mechanism)
- Experiments and results
- Discussion
- Conclusion

Overview Introduction: Machine Translation

- Model inputs source sentence as 1-of-k coded word vectors

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

and outputs a translated sentence of 1-of-k coded word vectors

$$y = (y_1, \dots, y_{T_y}), y_i \in R^{K_y},$$

T_x and T_y : length of source and target languages respectively.

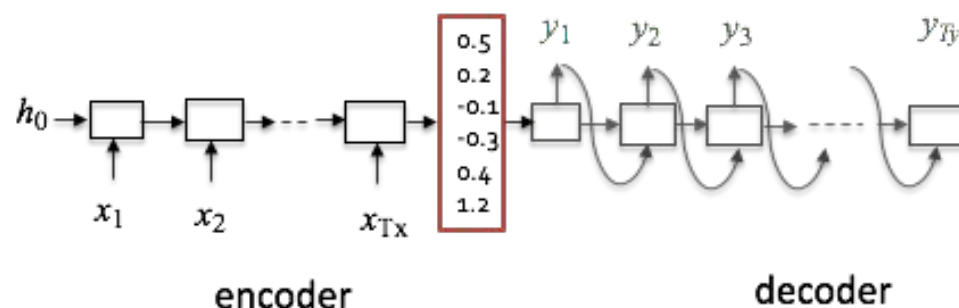
K_x and K_y : vocabulary sizes of source and target languages respectively,

Example

x_1 x_2 x_3 x_4 x_5
Jane visite l'Afrique en Septembre

Jane is visiting Africa in September.

y_1 y_2 y_3 y_4 y_5 y_6



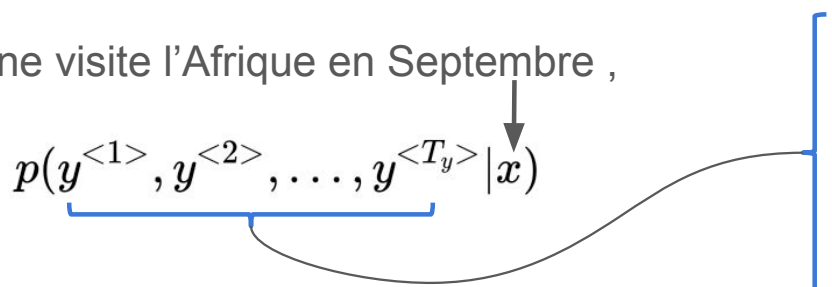
Encoder: Finds the encoding of the input/source sentence

Decoder: Uses the encoding to generate target sentence translation

Overview Introduction

Given:

Jane visite l'Afrique en Septembre ,

$$p(y^{<1>}, y^{<2>}, \dots, y^{<T_y>} | x)$$


- Jane is visiting Africa in September.
- Jane is going to be visiting Africa in September.
- ...
- In September, Jane will visit Africa.

Note: From the probabilistic perspective, **Translation** is equivalent to finding a **target sentence y** that **maximizes** the conditional probability of y given a **source sentence x**.

$$\operatorname{argmax}_{y^{<1>}, \dots, y^{<T_y>}} p(y^{<1>}, \dots, y^{<T_y>} | x)$$

- A common algorithm used in machine translation to find a value of y that maximizes conditional probability is **Beam search**.

Previous work: Basic RNN Encoder-decoder definition

- **RNN-encoder:** Map Input sentence x into a fixed-length context vector c .

$$x = (x_1, \dots, x_{T_x})$$

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

$h_t \in R^n$: encoder hidden state at time t

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

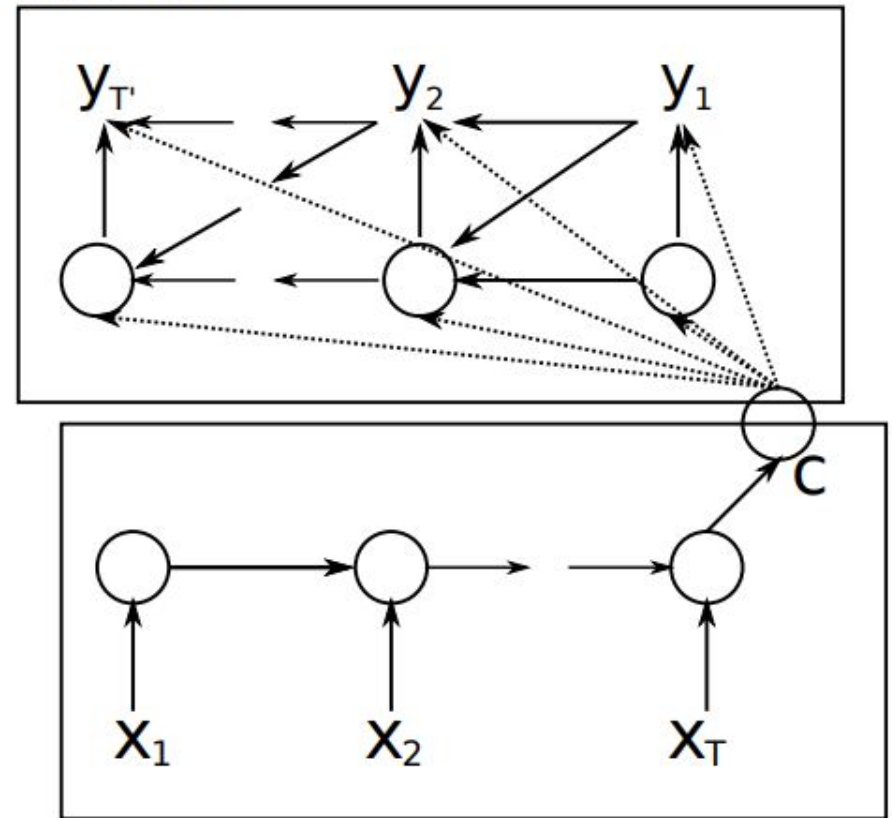
- **RNN-decoder:** Predict the output sentence y by maximizing the probability.

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

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

s_t : decoder hidden state at time t

Decoder



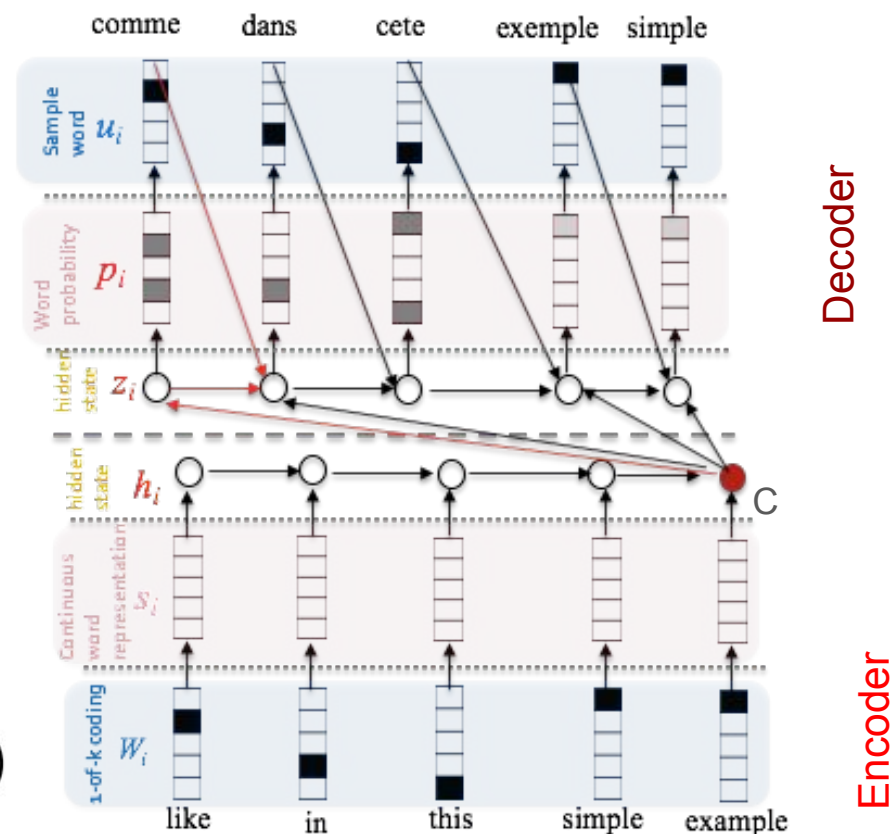
Encoder

Basic RNN Encoder-decoder

Compute every hidden state in decoder from

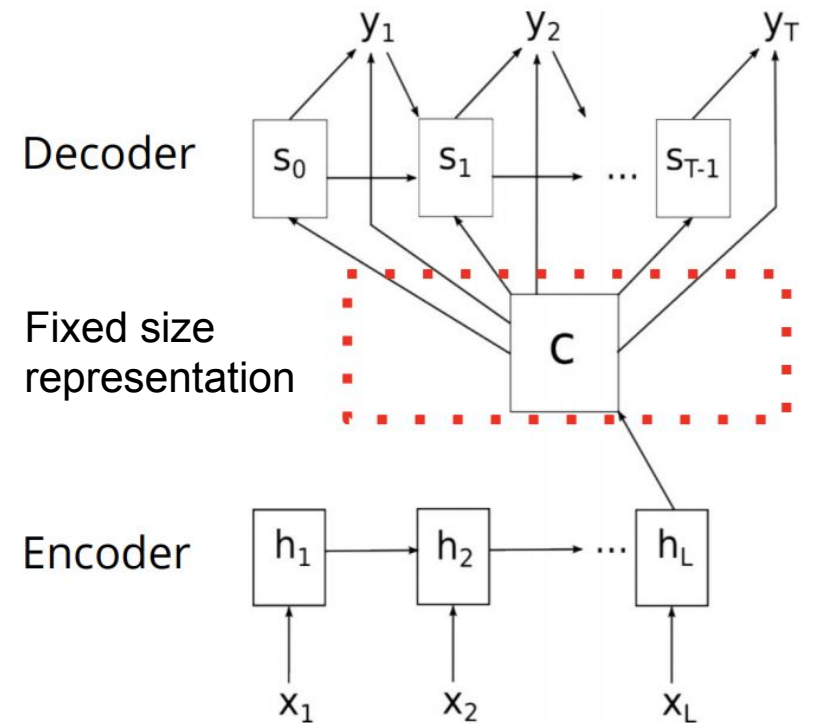
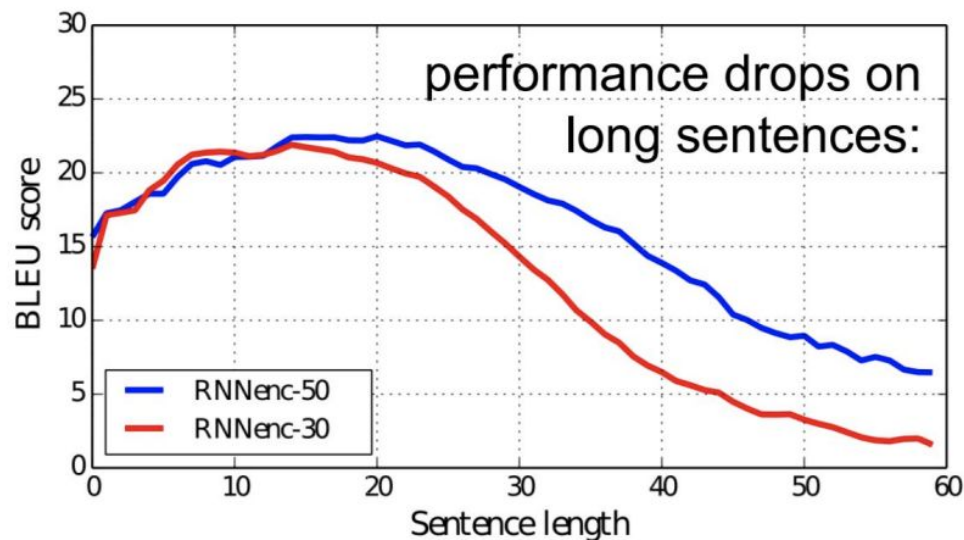
- Previous hidden state (standard RNN)
- Last hidden vector of the encoder C
- Previous predicted output word

Decoder hidden state at time t $s_t = (y_{t-1}, s_{t-1}, c)$



RNN Encoder-decoder: **Short falls**

- Has to remember the whole sentence Even with LSTM/GRU
- Fixed size representation can be a bottleneck
- Humans do it differently Need to attend to each individual word



Basic RNN Encoder-decoder: **shortfall** example

- Deviation in the end of long sentences

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

“based on his state of health???”

Learning to Align and Translate: Attention Mechanism

- Neural machine translation models often encode source sentence into a fixed-length vector from which a decoder generates translation. However, they do suffer on long sentences.
- Rather than building a single context vector out of the encoder's last hidden state, the secret sauce invented by attention is to create shortcuts between the context vector and the entire source input. The weights of these shortcut connections are customizable for each output element.

Learning to Align and Translate

Key Idea

Translate sentence part by part.

The agreement on European Economic Area was signed in 1992

L'accord sur

L'accord sur l'Espace économique européen a été signé en ???

Have such hints computed by the network itself!

Main Contribution

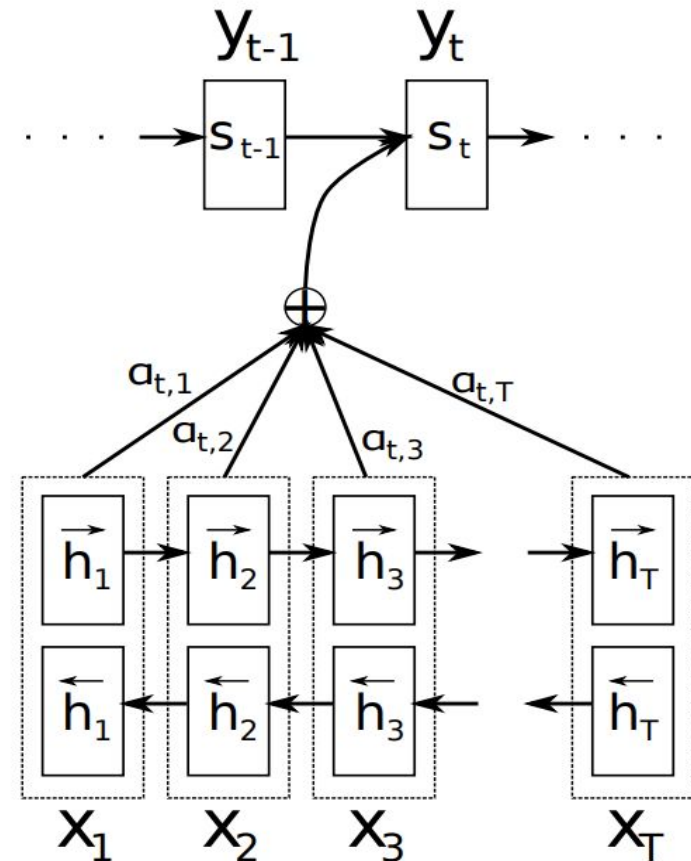
Propose a novel architecture for NMT

The **encoder**: a **bi-directional RNN**

- The hidden state should encode information from both previous and following words.

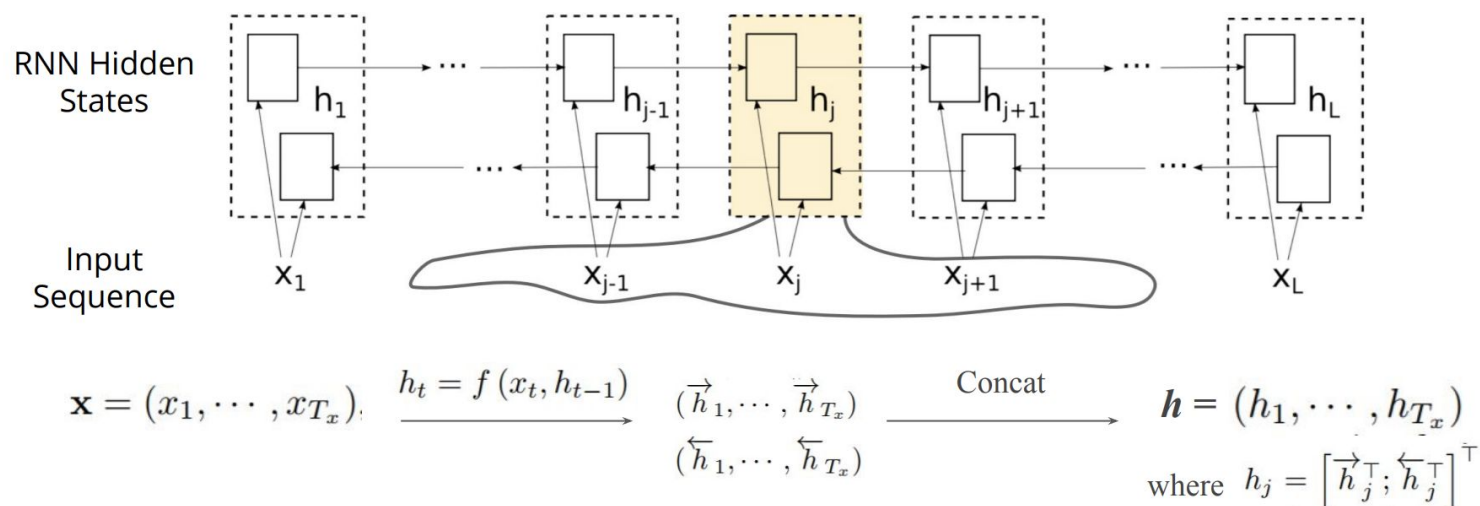
The **decoder**: **proposed an extension (attention) model**

- attention mechanism: **weighted sum of the input hidden states**



New encoder

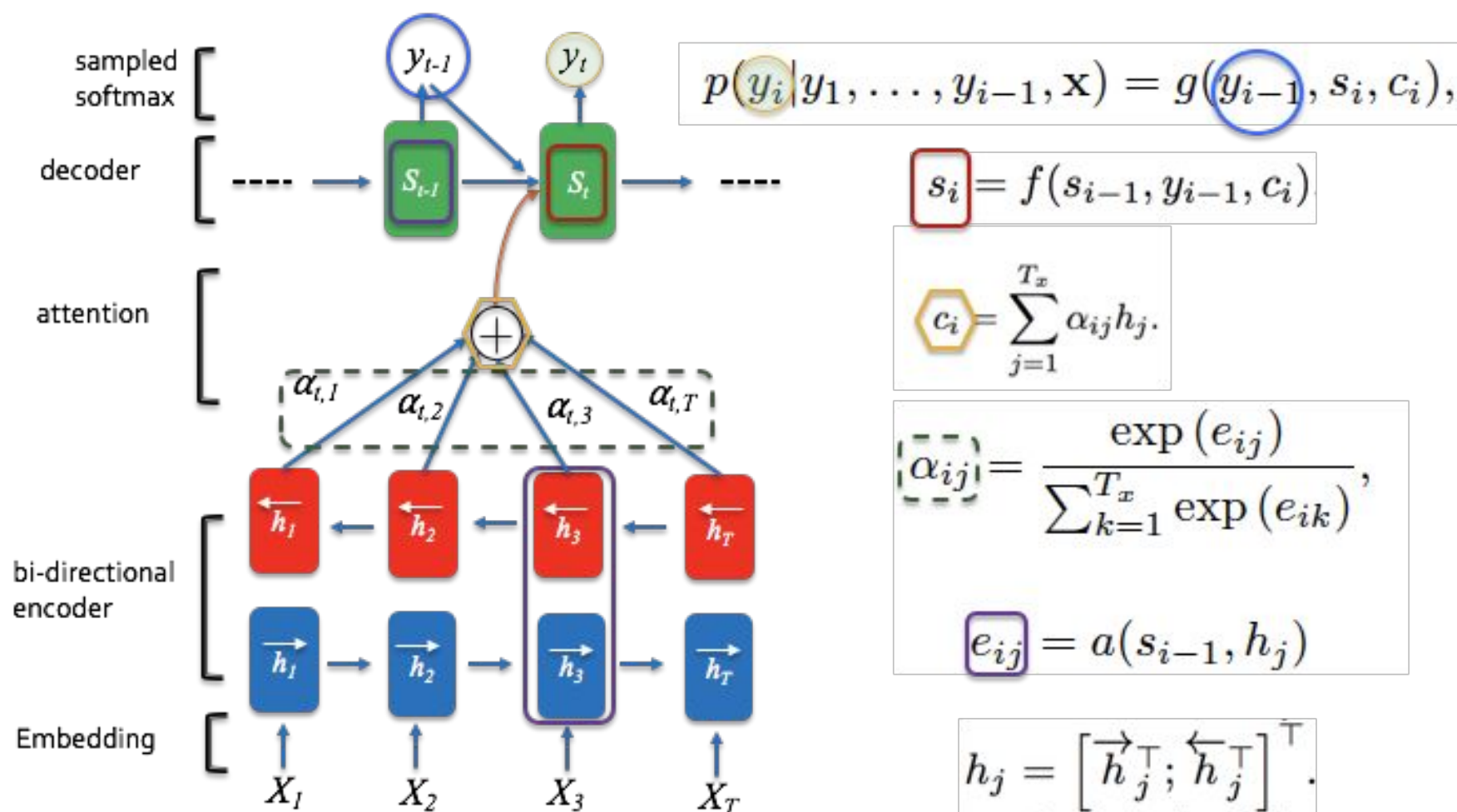
- **Bidirectional RNN**



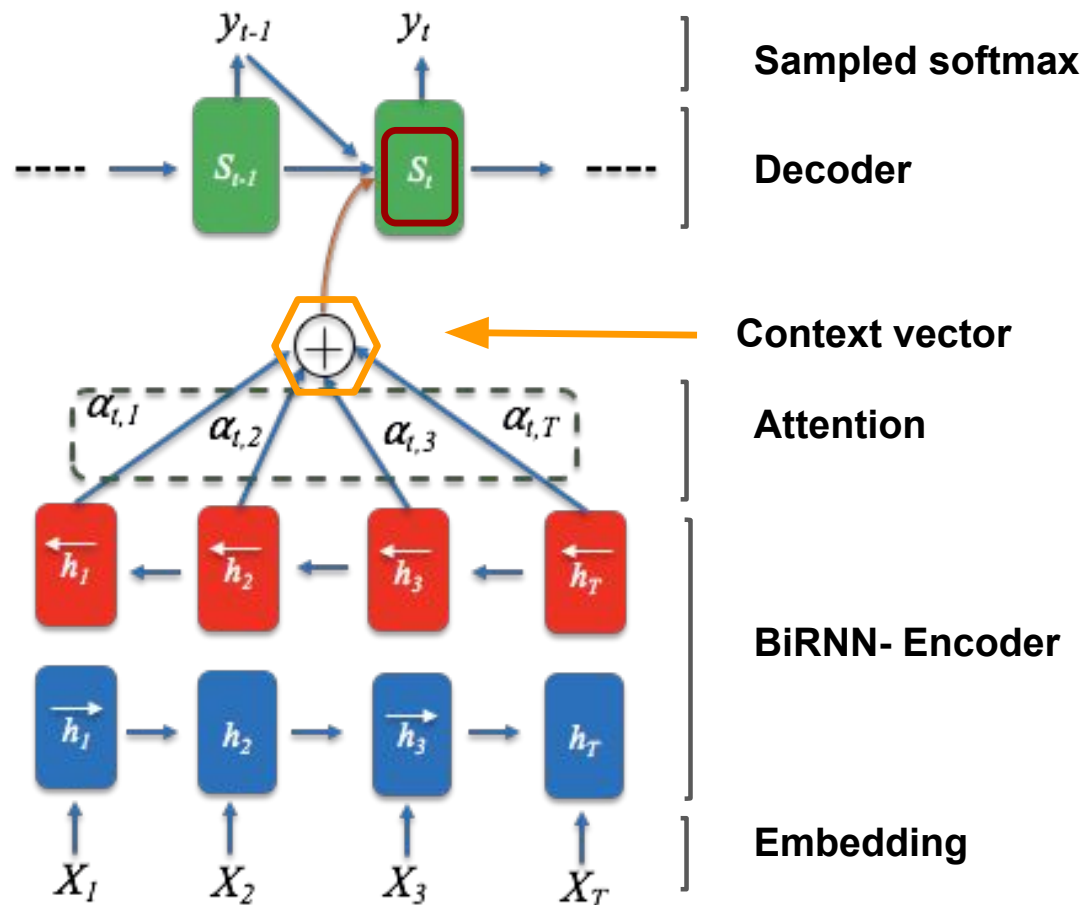
- The motivation is to include both the preceding and following words in the annotation of one word

Image source: <https://iclr.cc/archive/www/lib/exe/fetch.php%3Fmedia=iclr2015:bahdanau-iclr2015.pdf>

Learning to Align and Translate



New architecture



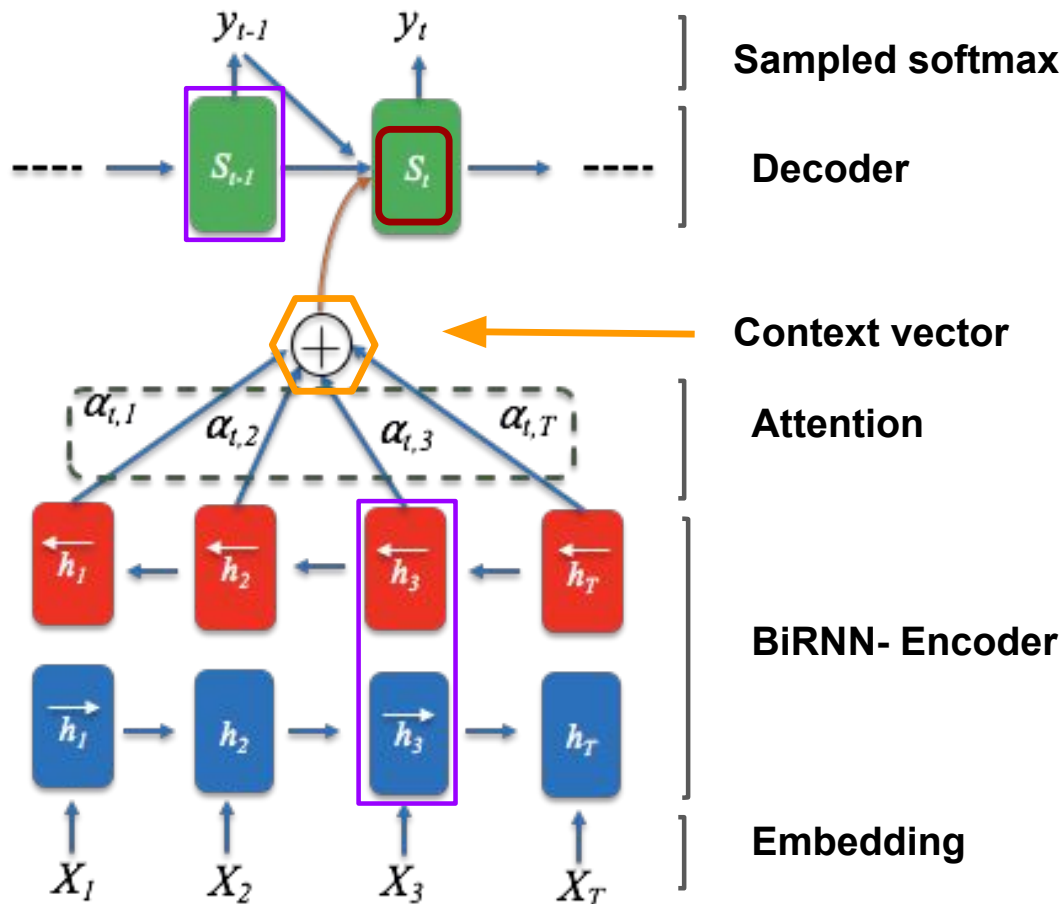
1. The decoder hidden state for the output word at time step t is

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

2. The context vector is the sum of the hidden states of the input sequence, weighted by alignment scores

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

New architecture



3. The weight for each annotation/hidden state is computed as a softmax of some predefined alignment score

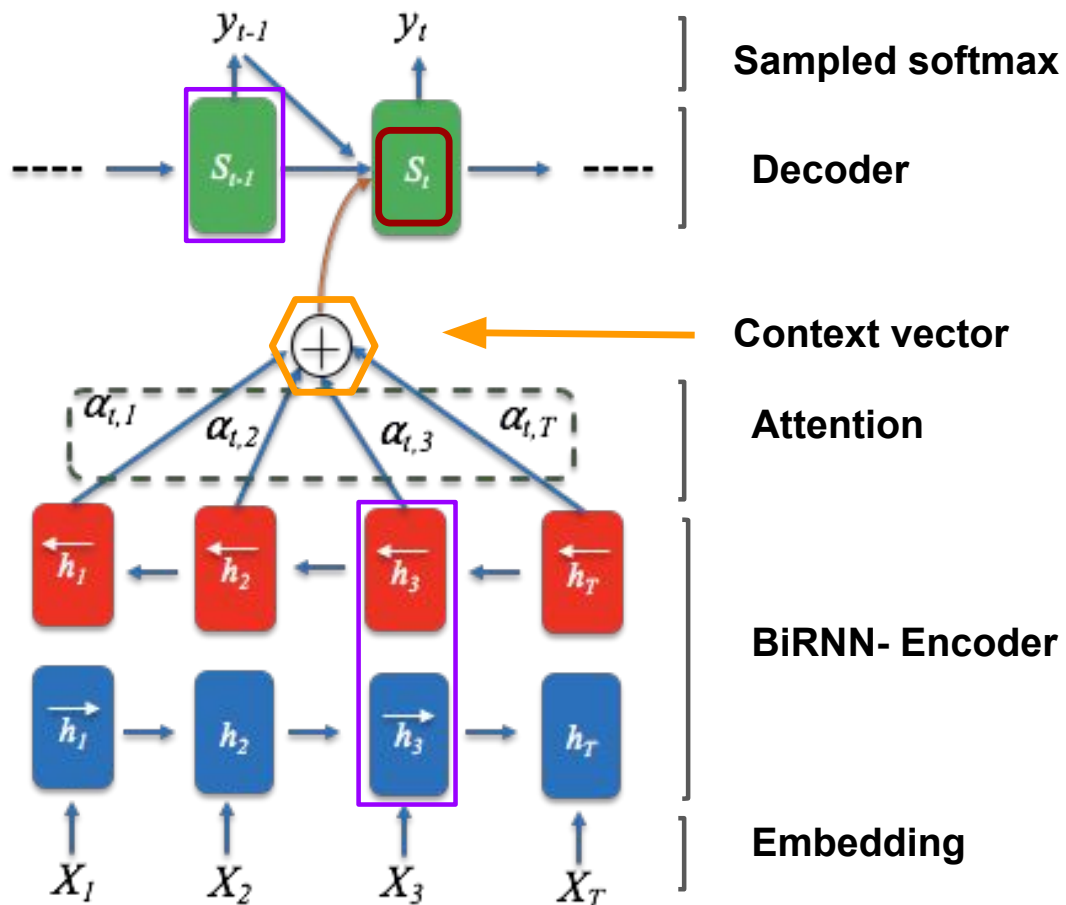
$$[\alpha_{ij}] = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

4. Where the alignment assigns a score to the input at position j and output at position i based on how well they match

$$e_{i,j} = a(s_{i-1}, h_j) = \text{score}(s_{i-1}, h_j)$$

NB: The weights defines how much of each source hidden state should be considered for each output

New architecture



4. Where the alignment assigns a score to the input at position j and output at position i based on how well they match

$$e_{i,j} = a(s_{i-1}, h_j) = \text{score}(s_{i-1}, h_j)$$

Assuming tanh is used as non-linear activation function, the score function is in the following form:

$$\text{score}(s_{i-1}, h_j) = v_a^T \tanh(W_a s_{i-1} + U_a h_j)$$

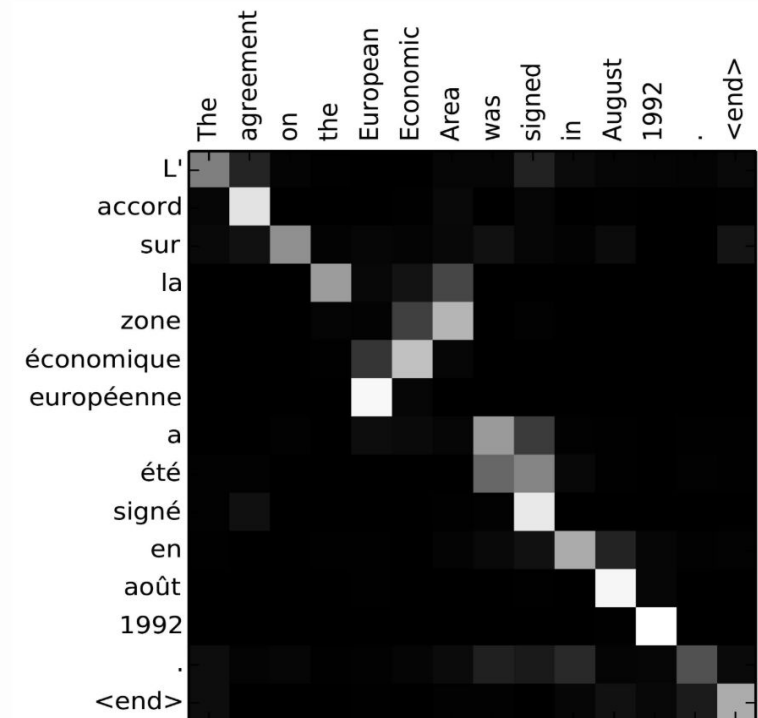
v_a, W_a, U_a are weight matrices to be learned in the alignment model

Alignment matrix: Explicitly shows the correlation between source and target words

Alignment matrix of “L’accord sur l’Espace économique européen a été signé en août 1992” (French) and its English translation “The agreement on the European Economic Area was signed in August 1992”

Each row of a matrix indicates the weights associated with the annotations.

From this we see which positions in the source sentence were considered more important when generating the target word.



Intuitions

- [Our model] does not attempt to encode the whole input sentence into a fixed-length vector. Instead, it encodes the input sentence into a sequence of vectors and chooses a subset of these vectors adaptively while decoding the translation.
- Each time the proposed model generates a word in a translation, it (soft-) searches for a set of positions in a source sentence where the most relevant information is concentrated.
- With this new approach the information can be spread throughout the sequence of annotations, which can be selectively retrieved by the decoder accordingly.

Experiment: English to French

Model

- RNN search, 1000 units

Baseline:

- RNN encoder-decoder, 1000 units
- Moses, a SMT system (Koehn et al. 2007)
- 1000 units (size of hidden layer)
- Word Embedding dimensionality 650
- Beam width 12

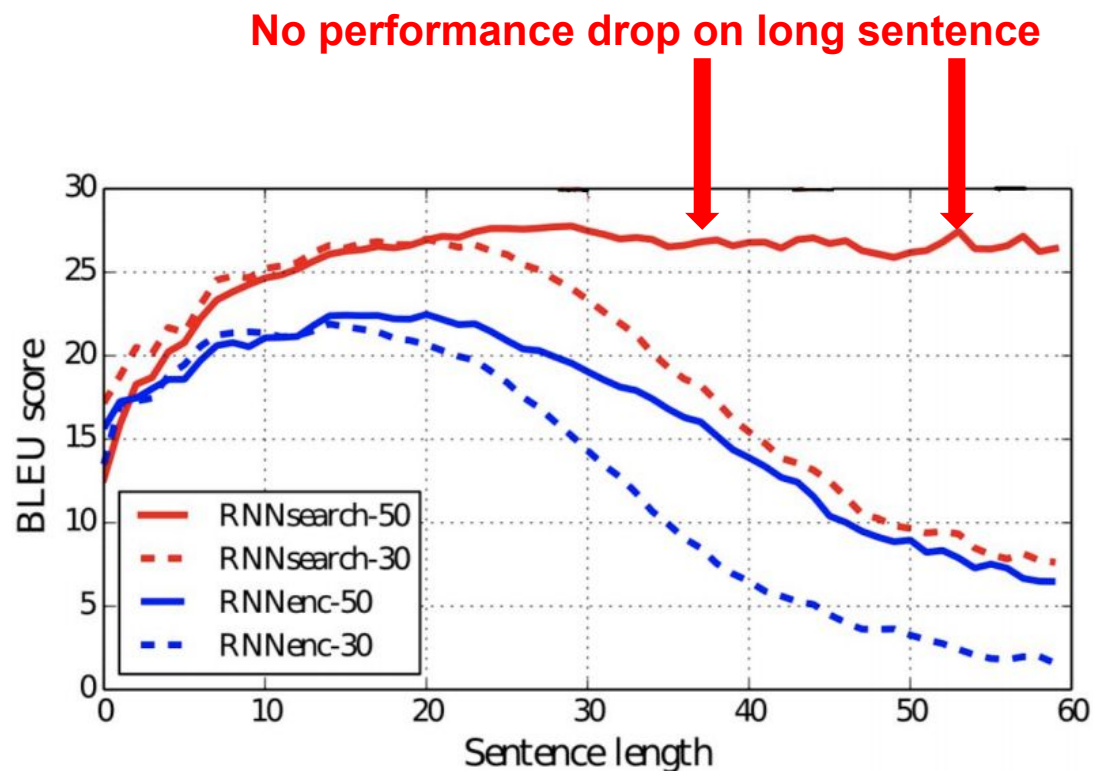
Data: <http://www.statmt.org/wmt14/translation-task.html> ACL WMT'14

- English to French translation, 348 million words
- 30,000 most frequent words + UNK token to train the model

Training:

- SGDw/ Adadelter
- Mini-batch: 80 sentences
- Train the model for approximately 5 days (Beam search+Blue)

Quantitative results



BLEU scores of the generated translations on the test set with respect to the lengths of the sentences. The results are on the full test set which includes sentences having unknown words to the models

Quantitative results

RNNsearch much better than RNN encoder-decoder

Model	All	No UNK ^o
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 SMT	33.30	35.63

Without unknown words, comparable with SMT

BLEU scores of the trained models computed on the test set. The second and third columns show respectively the scores on all the sentences and, on the sentences without any unknown word in themselves and in the reference translations.

RNNsearch-50*, was trained much longer until the performance on the development set stopped improving.

(^o) We disallowed the models to generate [UNK] tokens when only the sentences having no unknown words were evaluated (last column)

Qualitative Results: Translation

An admitting privilege is the right of a doctor to admit a patient to a hospital or a medical centre to carry out a diagnosis or a procedure, based on his status as a health care worker at a hospital.

New model
RNNsearch-50

Un privilège d'admission est le droit d'un médecin d'admettre un patient à un hôpital ou un centre médical pour effectuer un diagnostic ou une procédure, selon son statut de travailleur des soins de santé à l'hôpital.

Correct!

Encoder-decoder
RNNencdec-50

Un privilège d'admission est le droit d'un médecin de reconnaître un patient à l'hôpital ou un centre médical d'un diagnostic ou de prendre un diagnostic en fonction de son état de santé.

“based on his state of health???”

Discussion: Qualitative Analysis Alignment

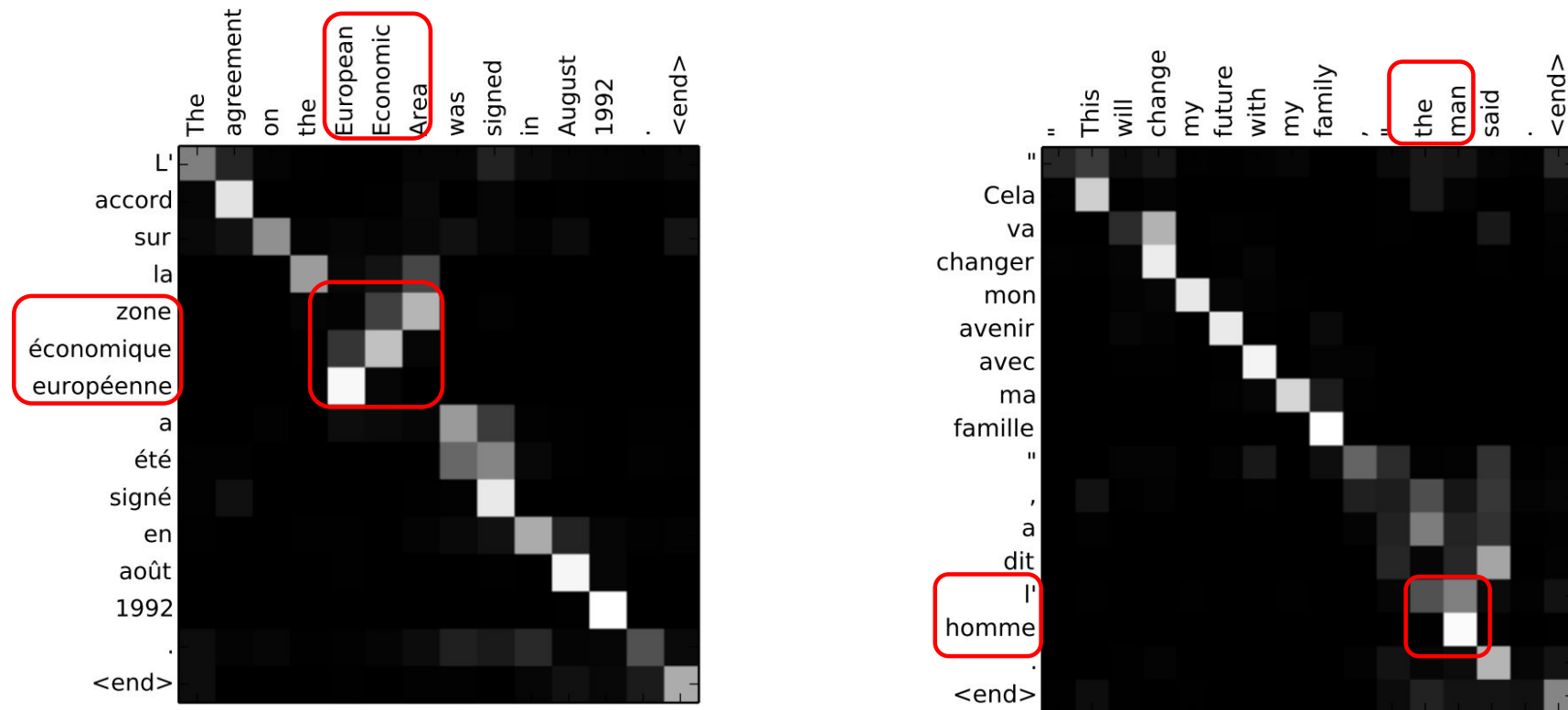


Figure shows sample alignments found by RNNsearch-50. The x-axis and y-axis of each plot correspond to the words in the source sentence (English) and the generated translation (French), respectively. Each pixel shows the weight α_{ij} of the annotation of the j -th source word for the i -th target word in grayscale (0: black, 1: white).

Discussion: Qualitative Analysis Alignment

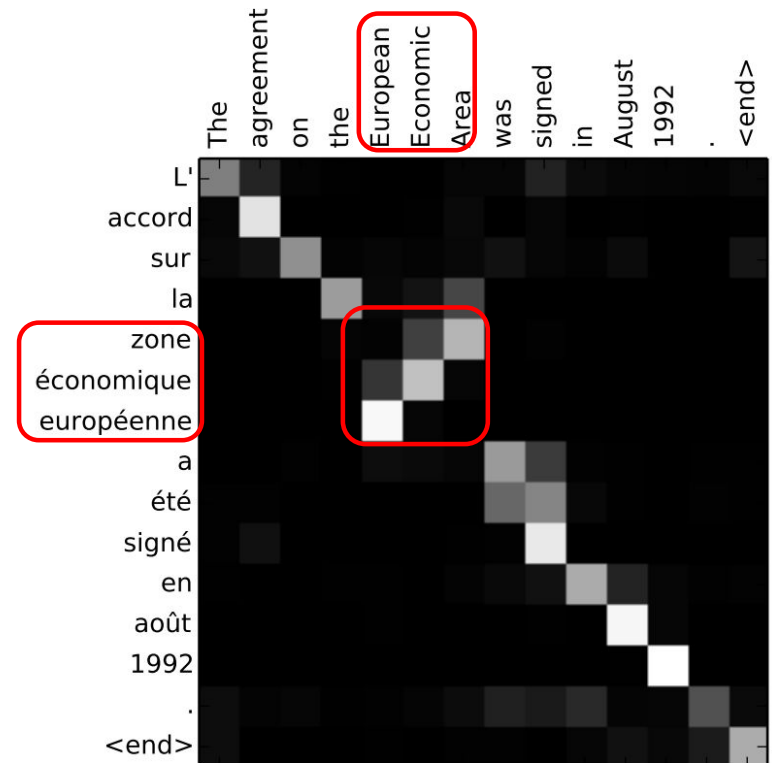
Alignment of words between English and French is largely monotonic. We see strong weights along the diagonal of matrix.

However, we also observe a number of non-monotonic alignments. Adjectives and nouns are typically ordered differently between French and English.

Example:

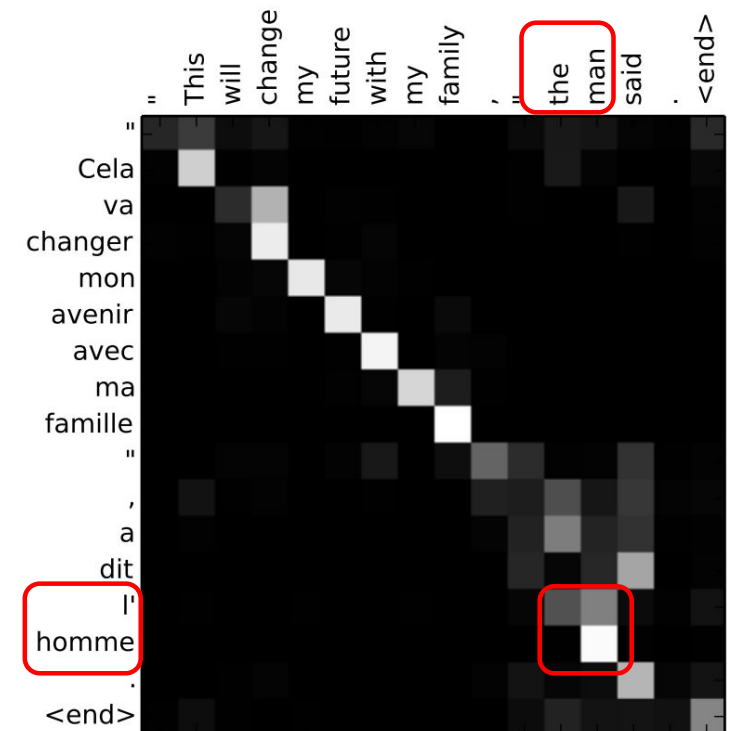
The model correctly translates a phrase [European Economic Area] into [zone économique européen].

The model was also able to correctly align [zone] with [Area], jumping over the two words ([European] and [Economic]), and then looked one word back at a time to complete the whole phrase [zone économique européen].



Discussion: Qualitative Analysis Alignment

- The strength of the soft-alignment, opposed to a hard-alignment, is evident. For instance, Consider the source phrase [the man] which was translated into [l'homme].
- Any hard alignment will map [the] to [l'] and [man] to [homme]. This is not helpful for translation, as one must consider the word following [the] to determine whether it should be translated into [le], [la], [les] or [l'].
- Our soft-alignment solves this issue naturally by letting the model look at both [the] and [man], and in this example, we see that the model was able to correctly translate [the] into [l'].



Conclusion

- We extended the basic encoder–decoder by letting a model (soft-)search for a set of input words, or their annotations computed by an encoder, when generating each target word. This frees the model from having to encode a whole source sentence into a fixed-length vector, and also lets the model focus only on information relevant to the generation of the next target word.
- The experiment revealed that the proposed RNNsearch outperforms the conventional encoder–decoder model (RNNencdec) significantly, regardless of the sentence length. From the qualitative analysis we were able to conclude that the model can correctly align each target word with the relevant words, or their annotations, in the source sentence as it generated a correct translation.
- The proposed approach achieved a translation performance comparable to the existing phrase-based statistical machine translation. It is a striking result, considering that the proposed architecture, has only been proposed recently. We believe the architecture proposed here is a promising step toward better machine translation and a better understanding of natural languages in general.