Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio

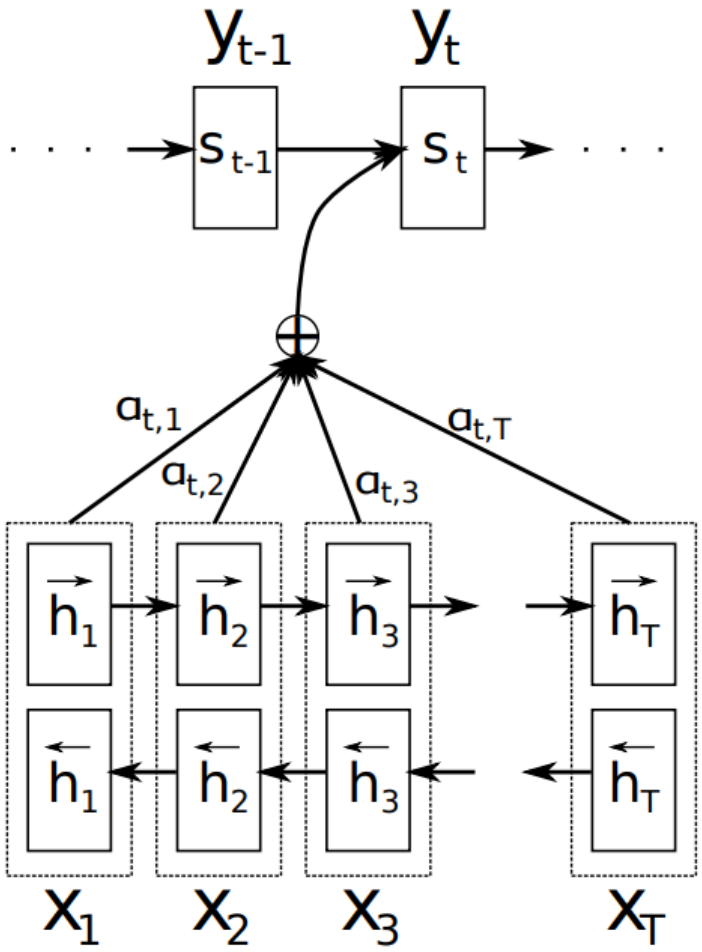
“Neural Machine Trranslation By Jointly Learning To Align Translate"

박 희 경

RNNsearch

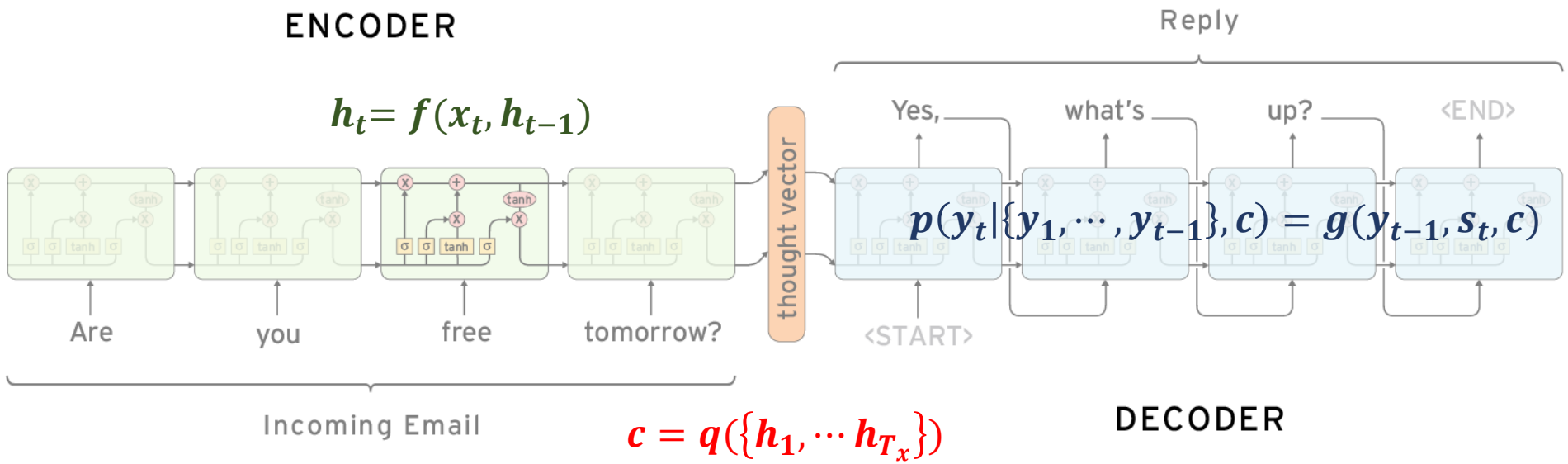
Introduction

- 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.



RNNsearch

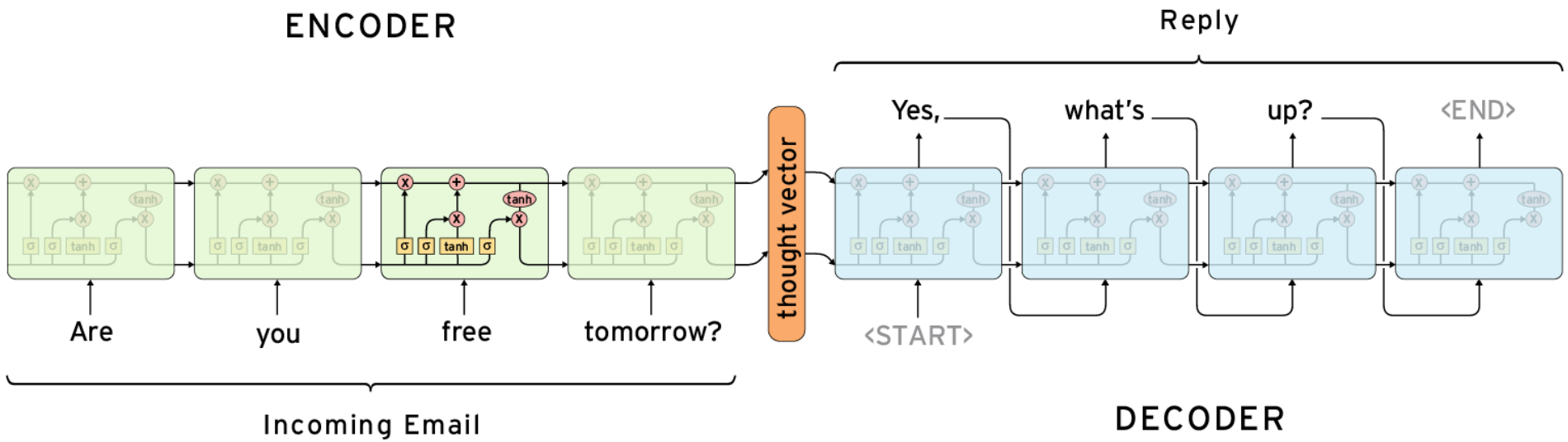
Motivation (RNN Encoder – Decoder)



- The encoder read the input sentence (a sequence of vectors) into a fixed vector c.
- c is a vector generated from a the sequence of the hidden states.
- The decoder is often trained to predict the next word $y_{t'}$, given the context vector c and all the previously predicted words $\{y_1, \cdots, y_{t'-1}\}$.

RNNsearch

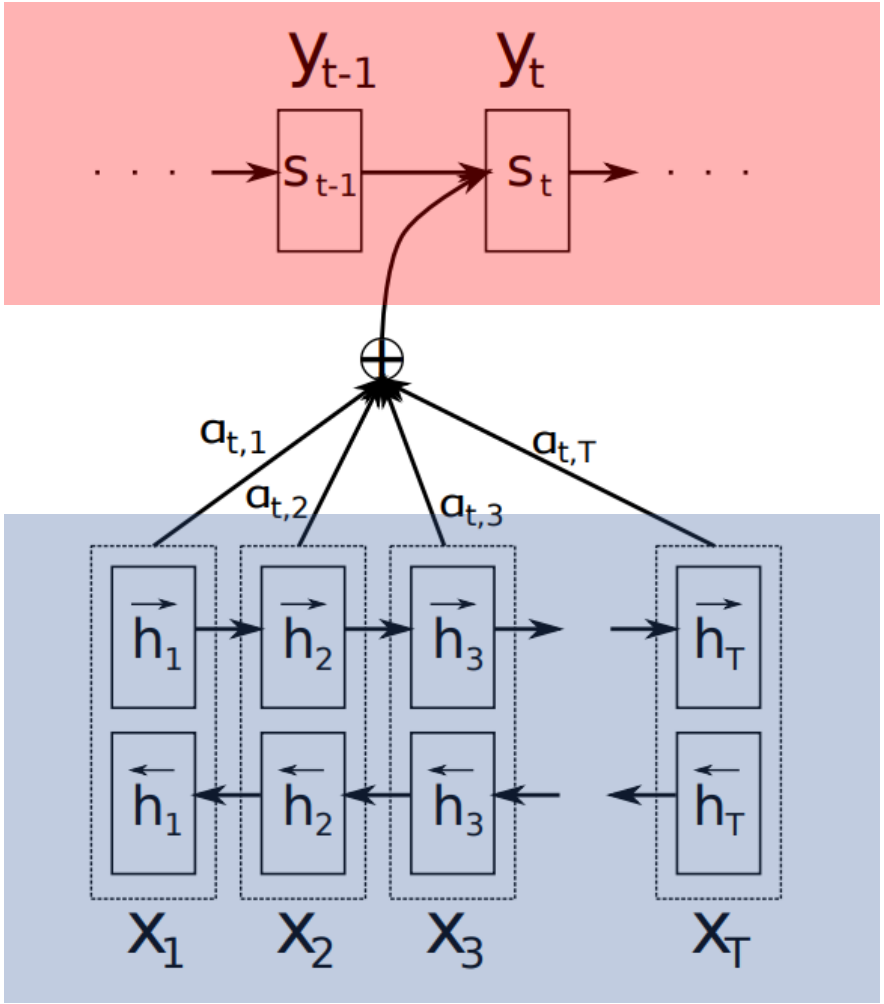
Motivation (RNN Encoder - Decoder)



- A potential issue with this encoder-decoder approach is that a neural network needs to be able to comporess all the necessary information of a source sentence into a fixed-length vector.
- This may make it difficult for the neural network to cope with long sentences, especially those that are longer than the sentences in the training corpus.

RNNsearch

Outline

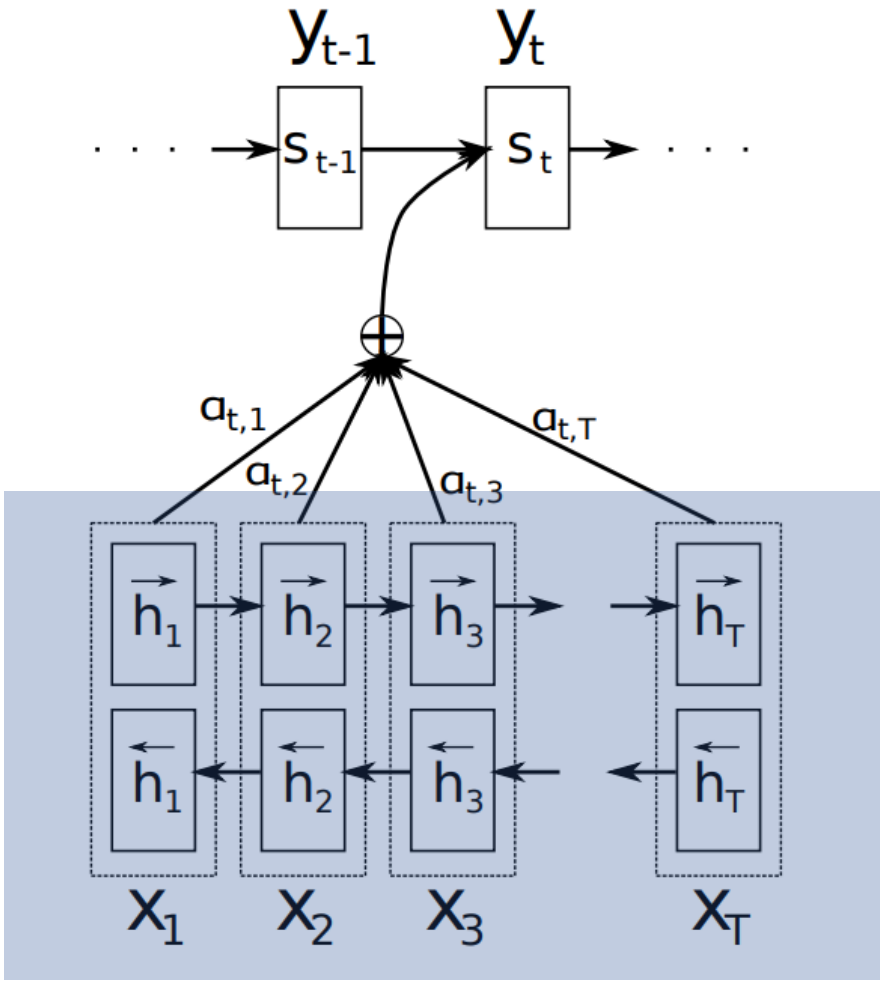


Decoder
: General Description
- Decoder decides parts fo the source sentence to pay attention to

Encoder
: Bidirectional RNN for annotating Sequences
- The annotation h_j contains the summaries of both the preceding words and the following words.

RNNsearch

Learning to Align and Translate



Encoder (BiRNN)
: Bidirectional RNN for Annotating Sequences.

- A BiRNN consists of forward $\overrightarrow{f}$ and backward $\overleftarrow{f}$ RNN.

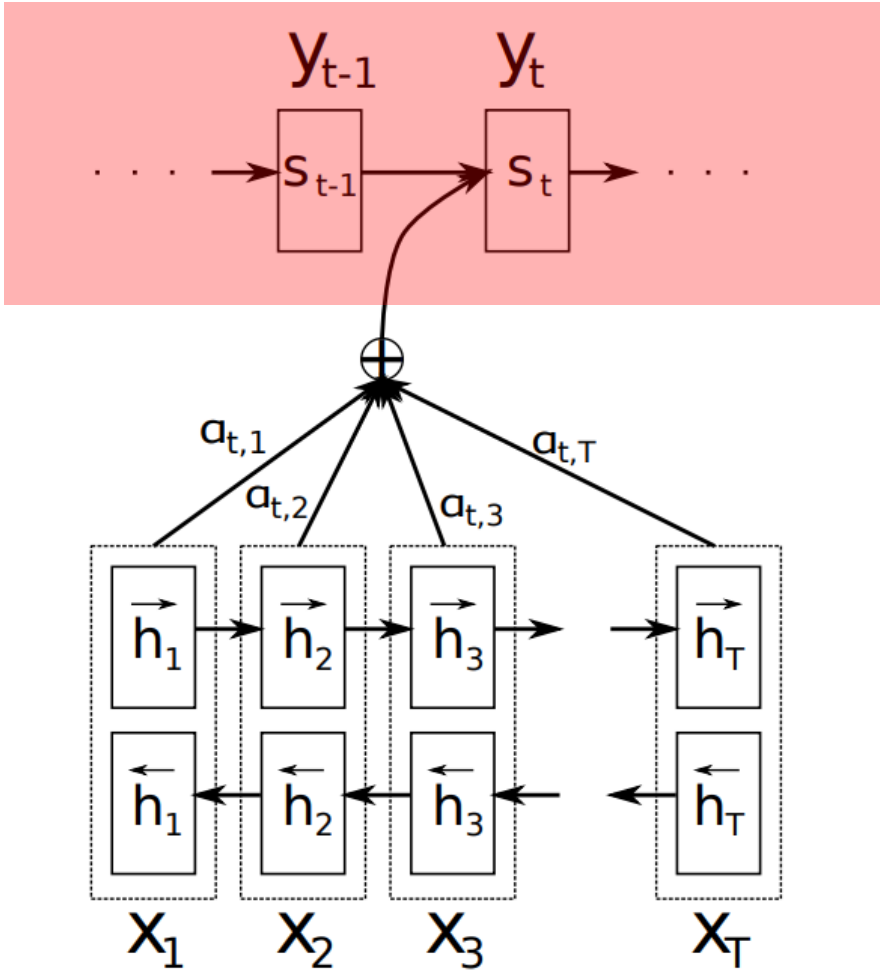
- We obtain an annotation for each word x_j by concatenating the forward hidden state $\overrightarrow{h_j}$ and the backward one $\overleftarrow{h_j}$, i.e., $h_j = [\overrightarrow{h_j^T} ; \overleftarrow{h_j^T}]^T$.

- In this way, the annotation h_j contains the summaries of both the preceding words and following words.

- Due to the endency of RNNs to better represent recent inputs, the annotation h_j will be focused on the words w_j aroud.

RNNsearch

Learning to Align and Translate



Decoder
: General Description

- Each conditional probability :

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

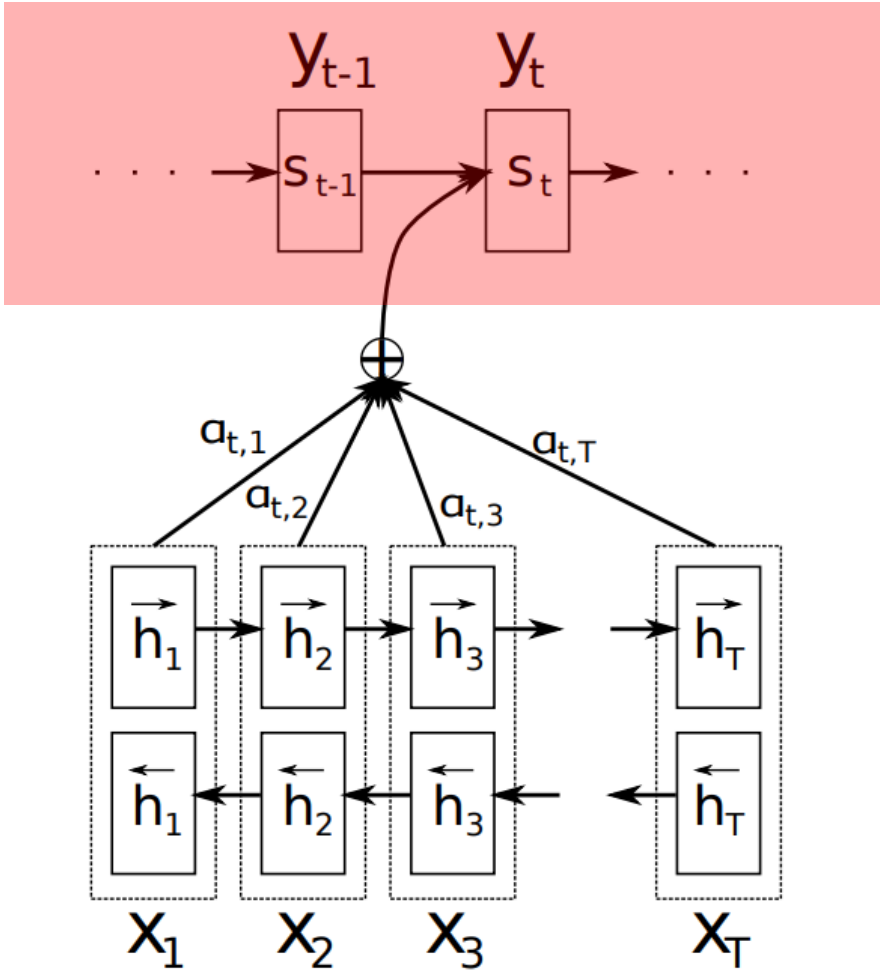
- RNN hidden state for time i

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

- The probability is conditioned on a distinct context vector c_i for each target word y_i

RNNsearch

Learning to Align and Translate



Decoder
: General Description

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

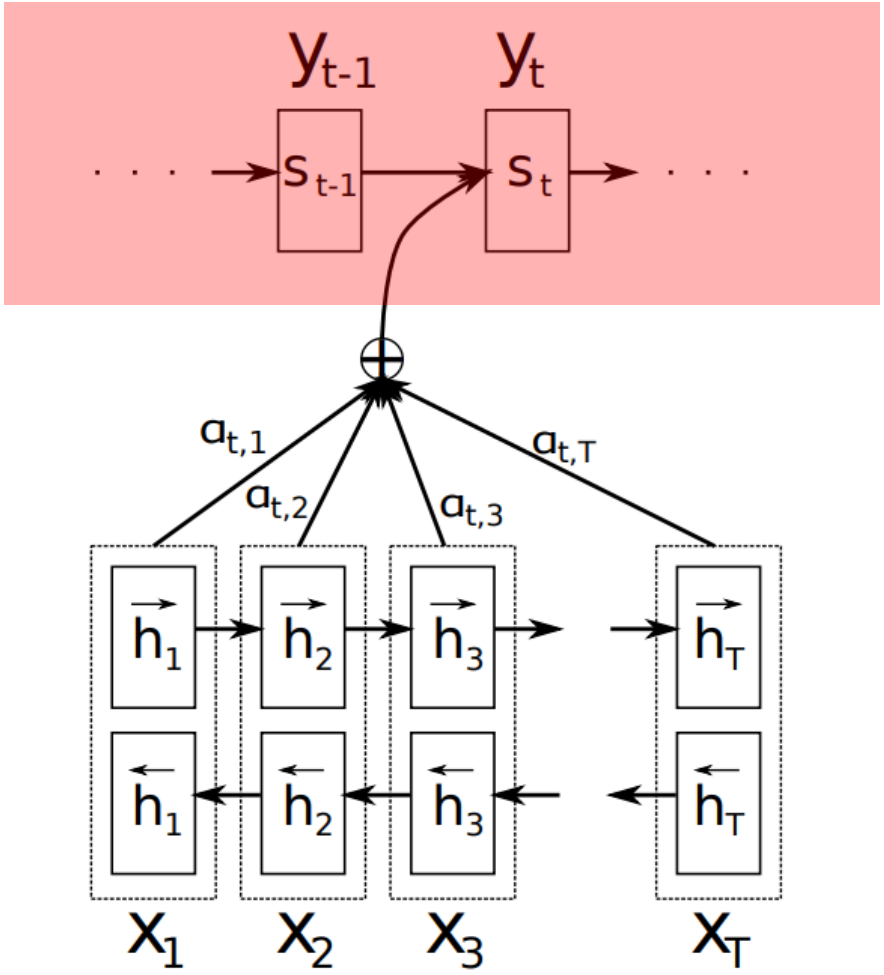
- The context vector c_i computed as a weighted sum of these annotations h_i.

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

- We can understand the approach of taking a weighted sum of all the annotations as computing an expected annotation, where the expectation is over possible alignments.

RNNsearch

Learning to Align and Translate



Decoder
: General Description

- The weight α_{ij} of each annotation h_j is computed by

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

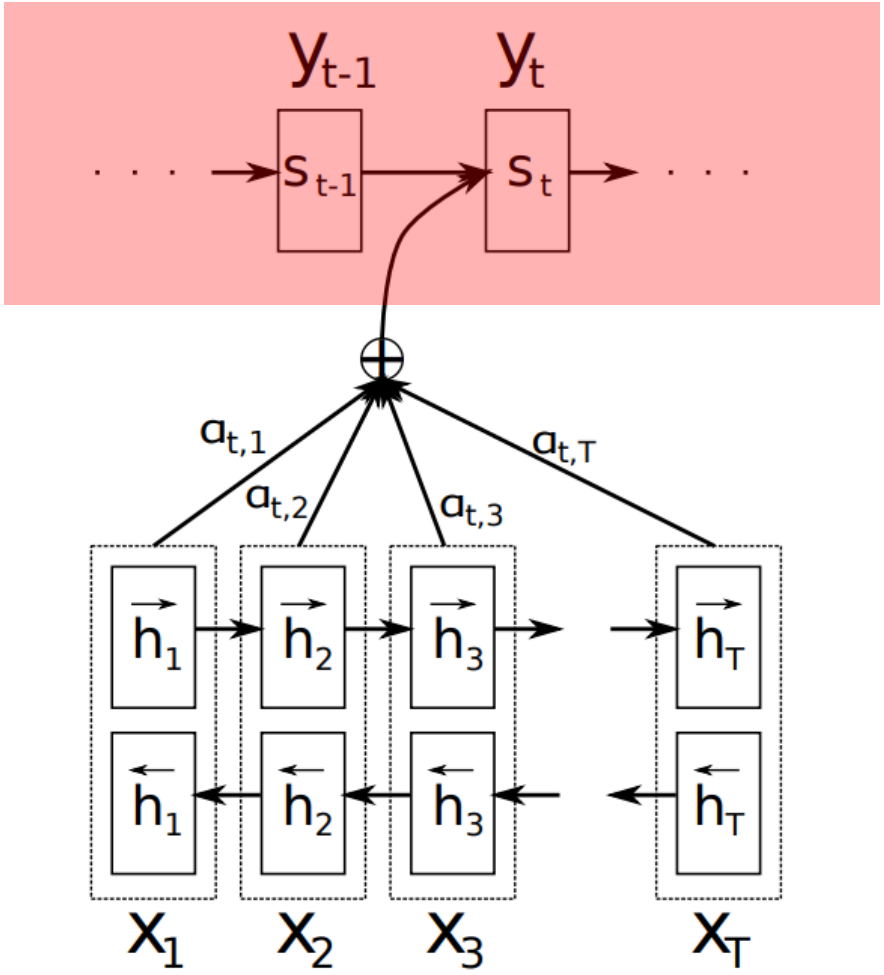
- Alignment model which scores how well the inputs around position j and the output at posion i match.

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

- We parametrize the alignment model a as feedforward neural network which is jointly trained with all the other componetns of the proposed system.

RNNsearch

Learning to Align and Translate



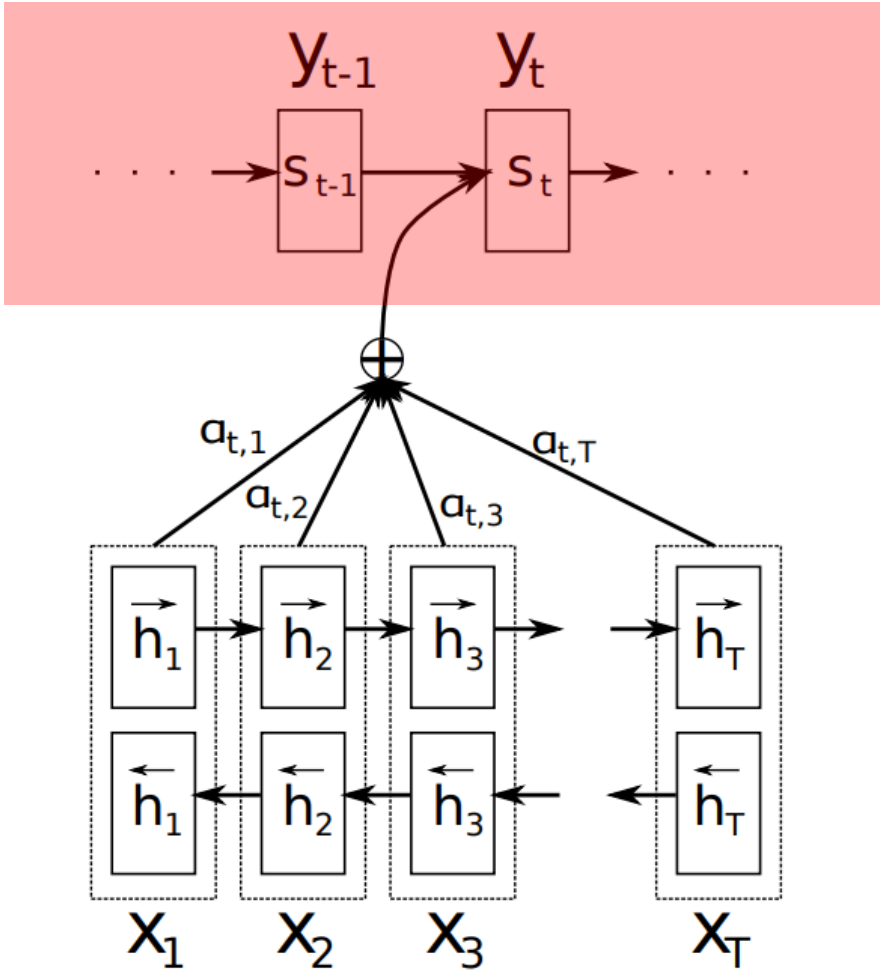
Decoder
: General Description

- We parametrize the alignment model a as feedforward neural network which is jointly trained with all the other componetns of the proposed system.

- This gradient can be used to train the alignment model as well as the whole translation model jointly.

RNNsearch

Learning to Align and Translate



Decoder
: General Description

- RNN hidden state for time i

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

- The context vector c_i

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

- The weight α_{ij} of each annotation h_j

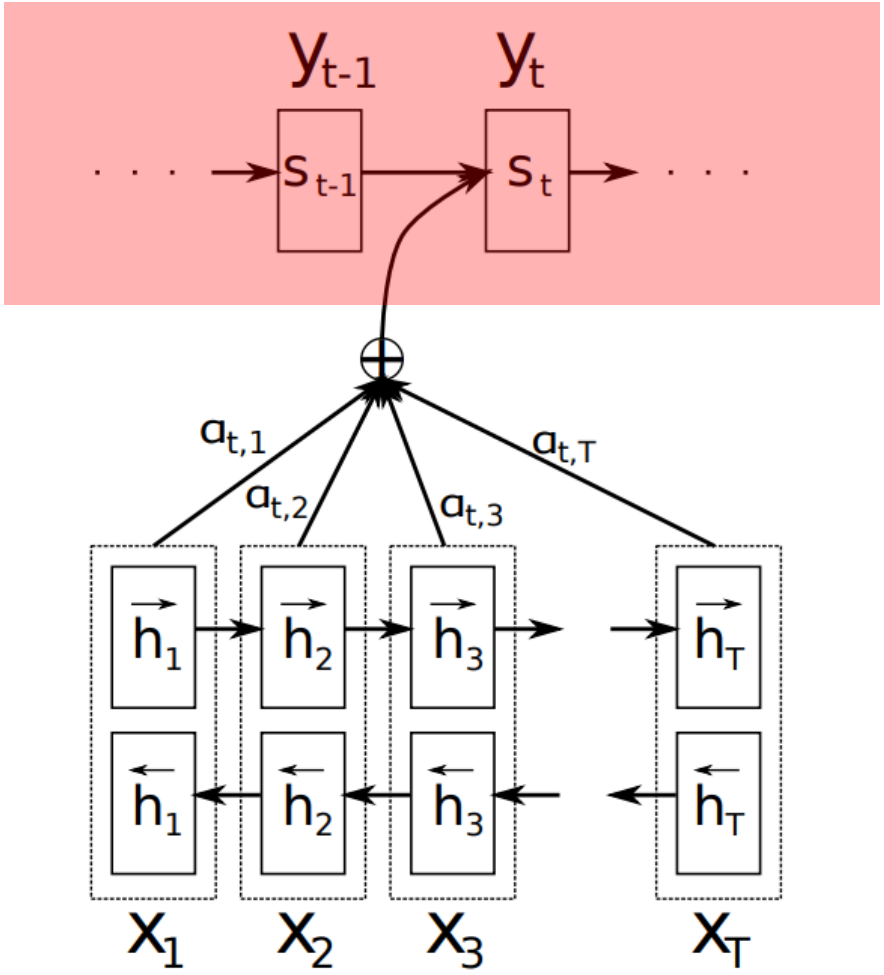
$$\alpha_{ij} = \frac{\exp(e_{ik})}{\sum_{k=1}^{T_x} \exp(e_{ik})}$$

- Alignment model scores

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

RNNsearch

Learning to Align and Translate



Decoder
: General Description

- The probability α_{ij}, or its associated energy e_{ij}, reflects the importance of the annotation h_i with respect to the previous hidden state s_{i-1} in deciding the next state s_i and generating y_i.

- Intuitively, this implements a mechanism of attention.

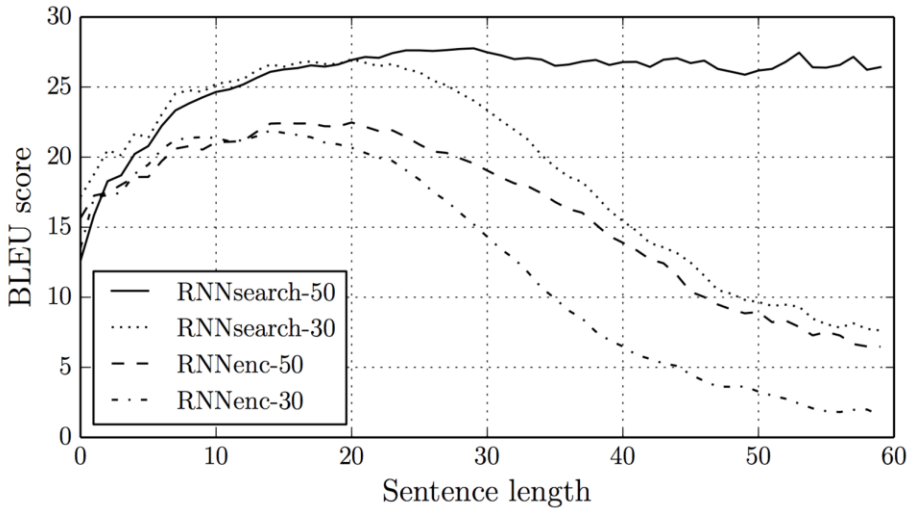
- The decoder decides parts of the source sentence to pay attention to.

- We relieve the encoder from the burden of having to encode all information in the source sentence into a fixed length vector.

RNNsearch

Results : Quantitative Results

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

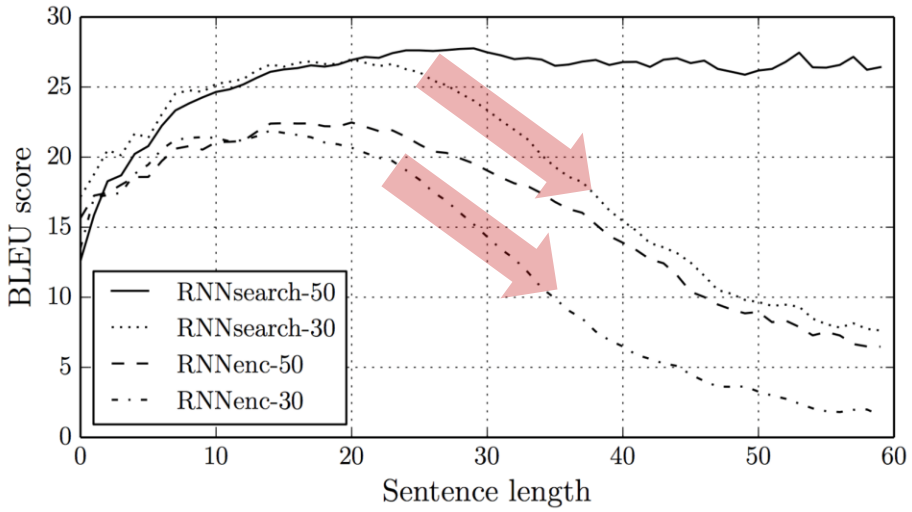


- The proposed RNNsearch outperforms the convetional RNNencdec.
- The performance of the RNNsearch is as high as that of the conventional phrase-based translation system (Moses), when only the sentences consisting of known words are considered.
- We conjectured that this limitation may make the basic encoder-decoder approach to underperform with long sentences.

RNNsearch

Results : Quantitative Results

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

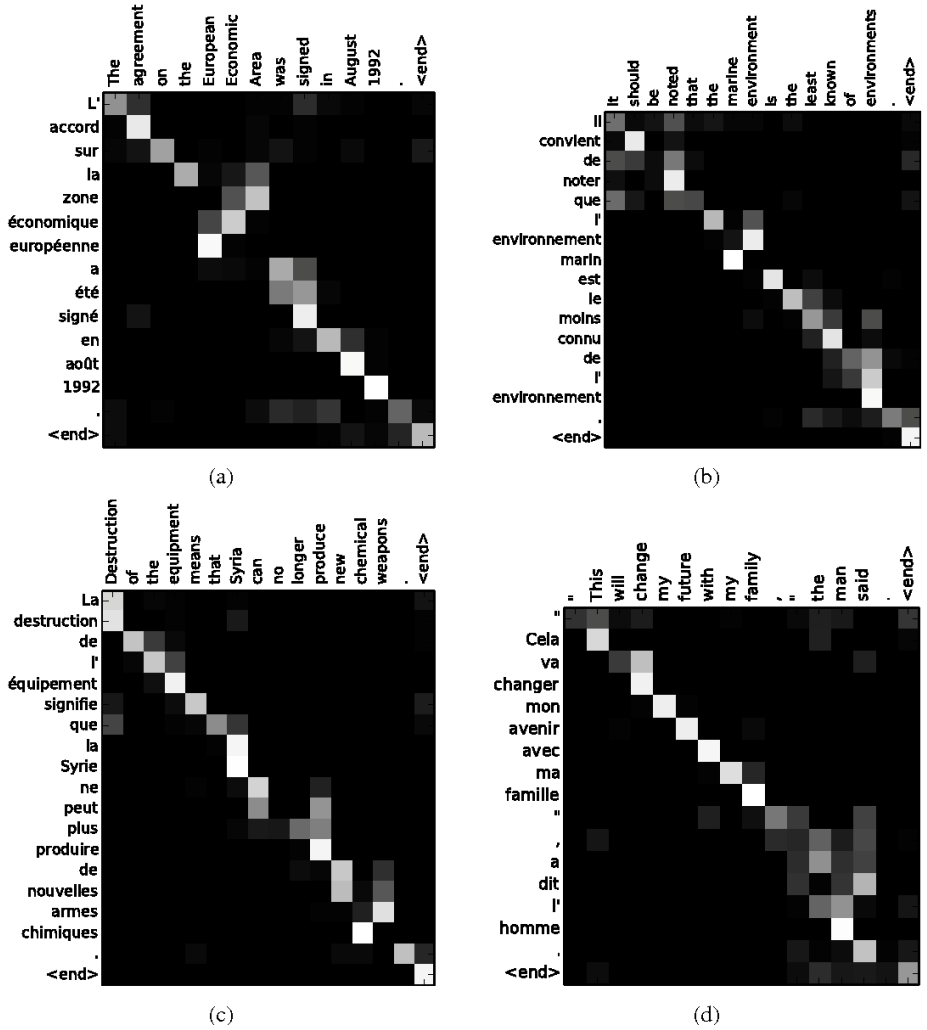


- We see that the performance of RNNencdec dramatically drops as the length of the sentences increases.

- On the other hand, RNNsearch are more robust to the length of the sentences.

RNNsearch

Results : Qualitative Analysis



- The proposed approache probides an intuitive way to inspect the (soft-)alignment between the words in a generated translation and those in a source sentence by visualizing the annotation weights α_{ij}.

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

RNNsearch

Conclusion