Sequence to Sequence Models (Seq2Seq) Using RNNs and An Attention Mechanism

Alvaro Soto

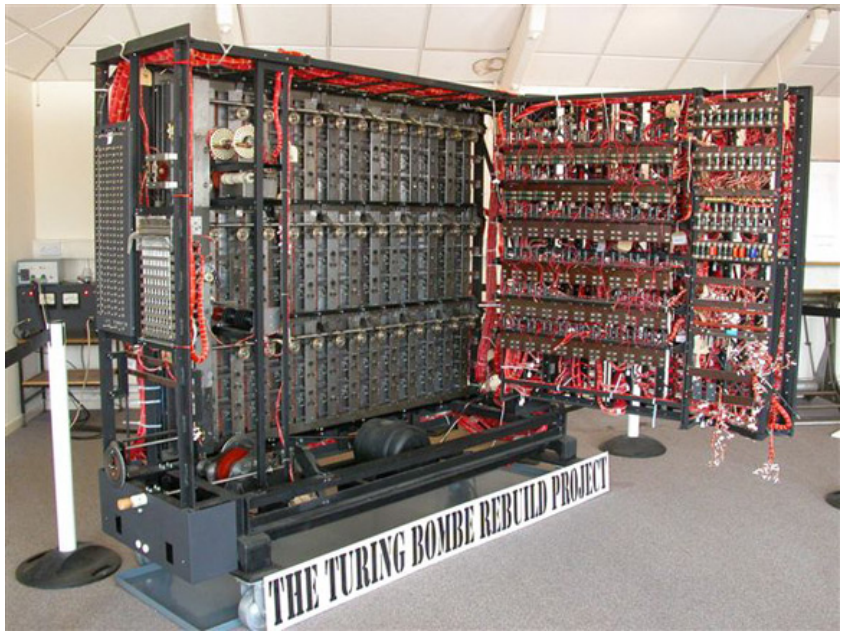
Computer Science Department, PUC

Language is a hallmark of human intelligence

Language is a medium to encode information

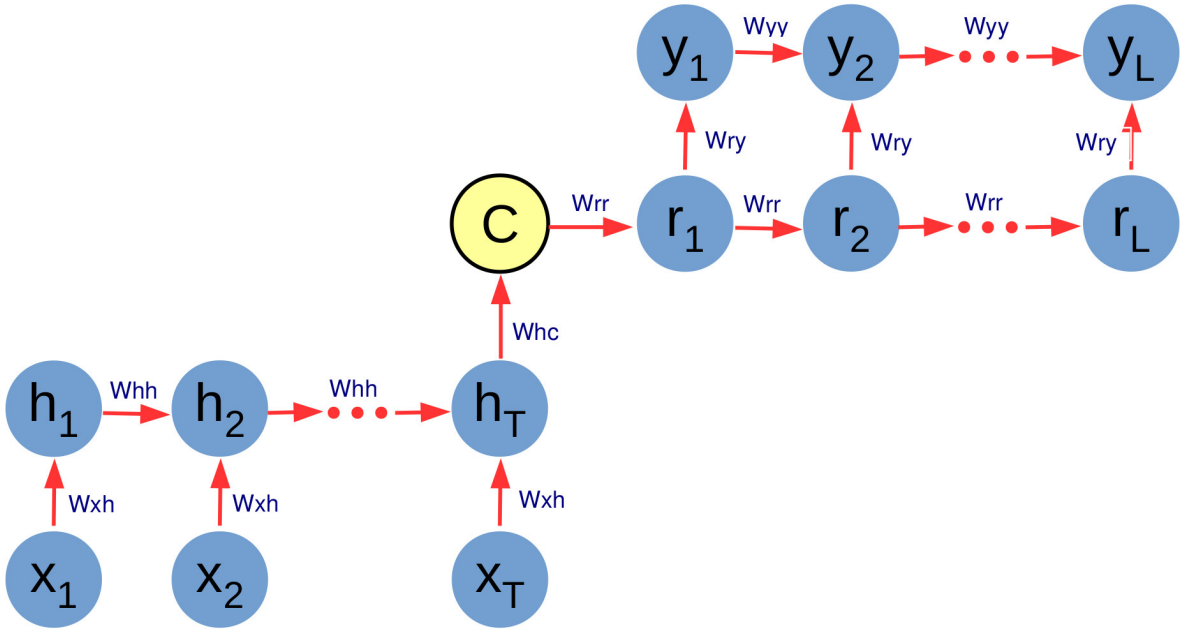
Language is a hallmark of human intelligence

Equally important is the ability to decode the information

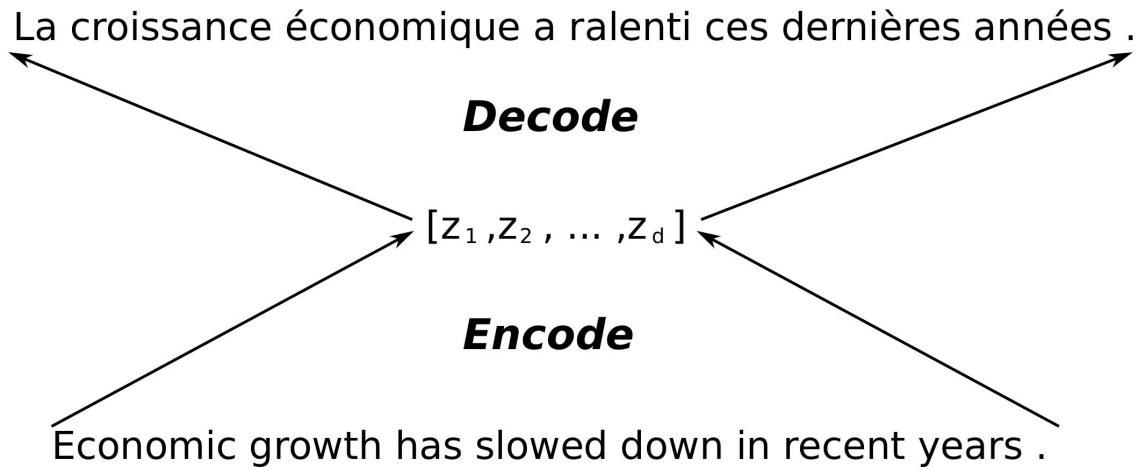


An encoding/decoding process can be modeled as a Seq2Seq convertion task

1. An encoder or input RNN processes the input sequence, and emits an intermediate fixed-length vector representation (ex. as a simple function of its final hidden state).
2. A decoder or output RNN is conditioned on the intermediate vector to generate a suitable output sequence.



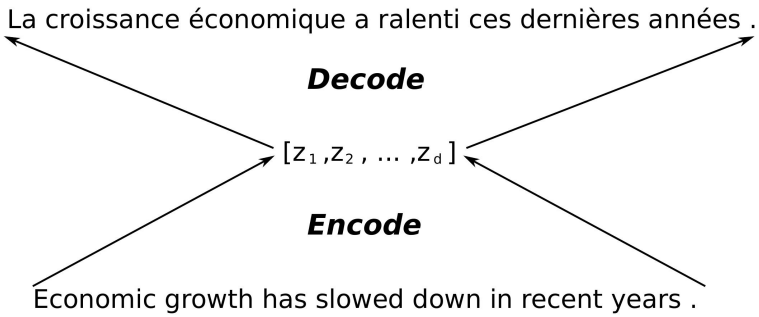
1. An encoder or input RNN processes the input sequence, and emits an intermediate fixed-length vector representations.
2. A decoder or output RNN is conditioned on the intermediate vector to generate a suitable output sequence.



As a main novelty, the intermediate representation $[z_1, \ldots, z_d]$ acts as a latent space, such that the input and output sequences can have different sizes.

We can think about the mid-level representation as a semantic embedding space that encodes relevant semantic of the input instances.

- Several ML applications can be cast as sequence-to-sequence transformation. Ex. speech recognition, machine translation, or question answering, among many others.
- The encoder-decoder approach is a fruitful model to handle these cases.



A woman with a little girl in a park, the woman is throwing a fresbee.

We can think about the mid-level representation as a semantic embedding space that encodes relevant semantic of the input instances?

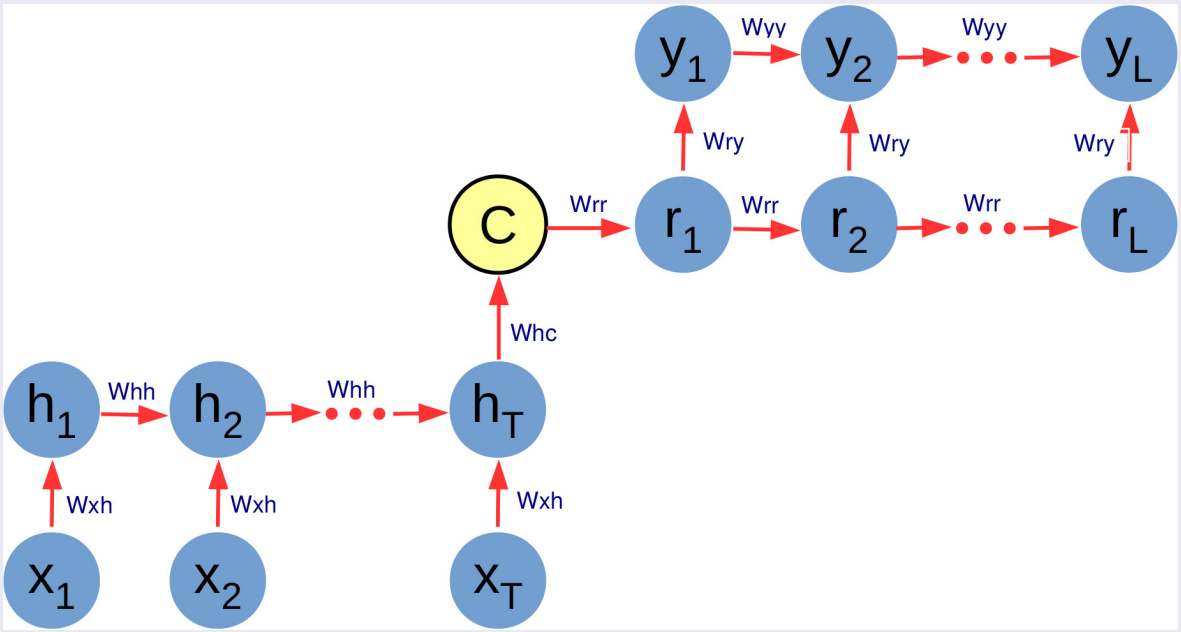
As an example, in the case of text inputs:

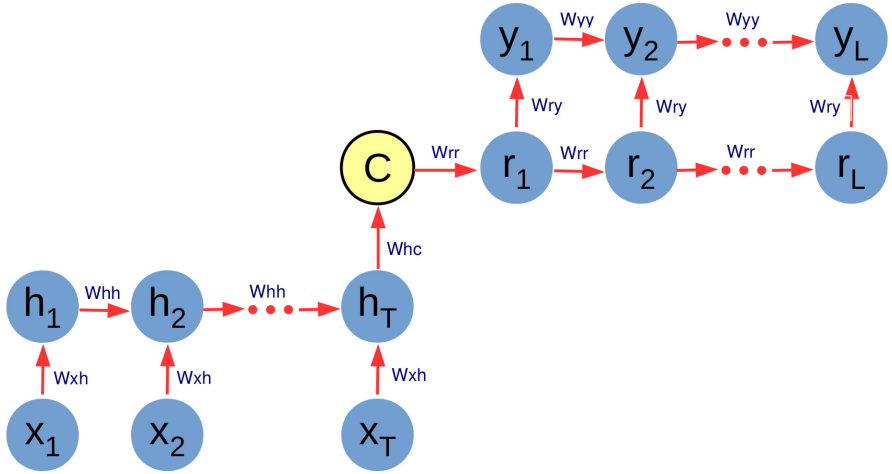
- The intermediate embedded space maps sentences with similar meaning close to each other, and far from unrelated sentences.
- Furthermore, due to the sequential construction, the encoding is sensitive to word order, a capability not possible in BoW models.

- Encoder and decoder RNNs are jointly trained to maximize the average conditional log probability over all pairs of training sequences $(x_i, y_i) \in TS$, i.e., we maximize:

$$\frac{1}{|TS|} \sum_{(x_i, y_i) \in TS} \log P(y_i | x_i)$$

- One of the simplest configuration considers an encoder that outputs a context vector C that is then transformed to a sequence by a decoder network:





In this case, the model is given by:

1. Encoder RNN

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1})$$

2. Decoder RNN

$$\text{if}(t = 1)$$

$$y_1 = \sigma(W_{ry}r_1)$$

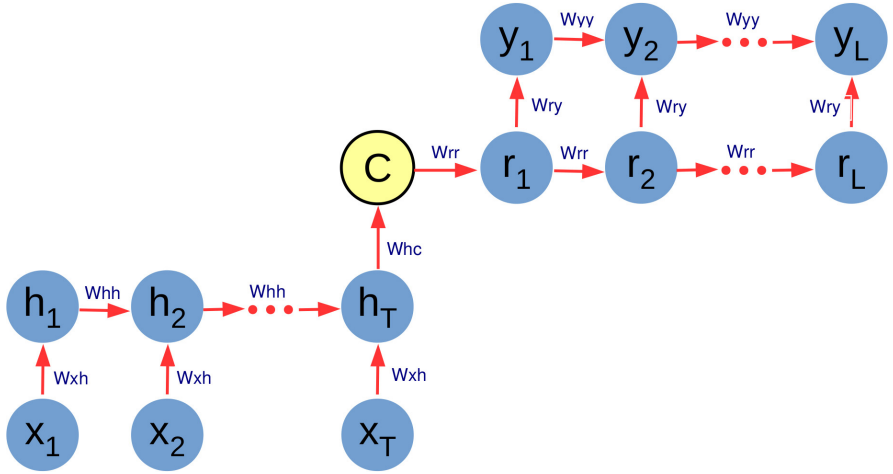
$$r_1 = \sigma(W_{rr}C)$$

$$C = \sigma(W_{hC}h_T)$$

$$\text{if}(t > 1)$$

$$y_t = \sigma(W_{ry}r_t + W_{yy}y_{t-1})$$

$$r_t = \sigma(W_{rr}r_{t-1})$$



In this case, the model is given by:

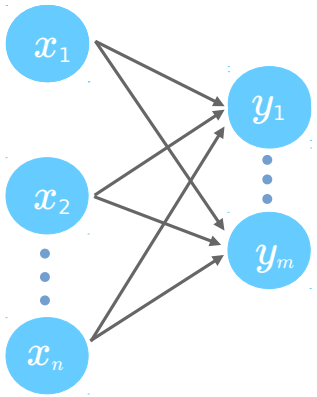
1. Encoder RNN

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1})$$

2. Decoder RNN

$$\text{if}(t = 1)$$
$$y_1 = \sigma(W_{ry}r_1)$$
$$r_1 = \sigma(W_{rr}C)$$
$$C = \sigma(W_{hC}h_T)$$
$$\text{if}(t > 1)$$
$$y_t = \sigma(W_{ry}r_t + W_{yy}y_{t-1})$$
$$r_t = \sigma(W_{rr}r_{t-1})$$

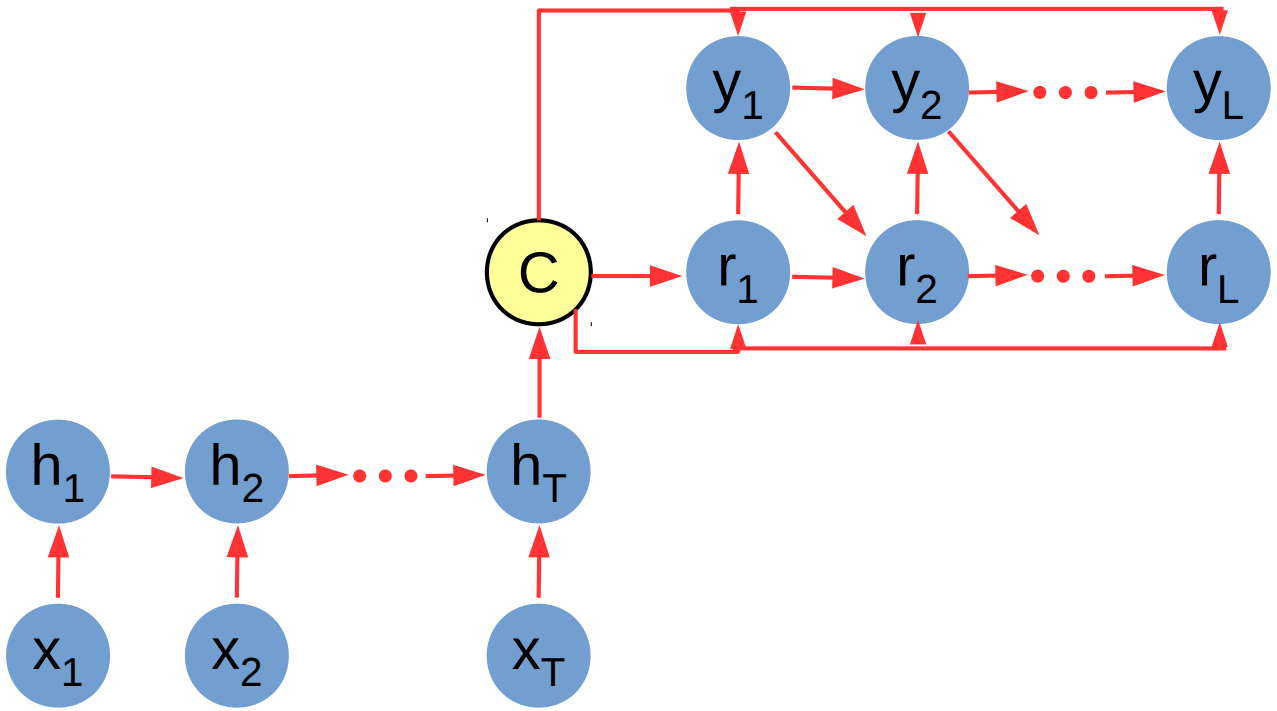
In general, each model W_{ij} is given by a MLP, i.e., a fully connected NN:



$$\begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix} = \sigma \begin{pmatrix} w_{11} & \cdots & w_{1n} \\ \vdots & \vdots & \vdots \\ w_{m1} & \cdots & w_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$$

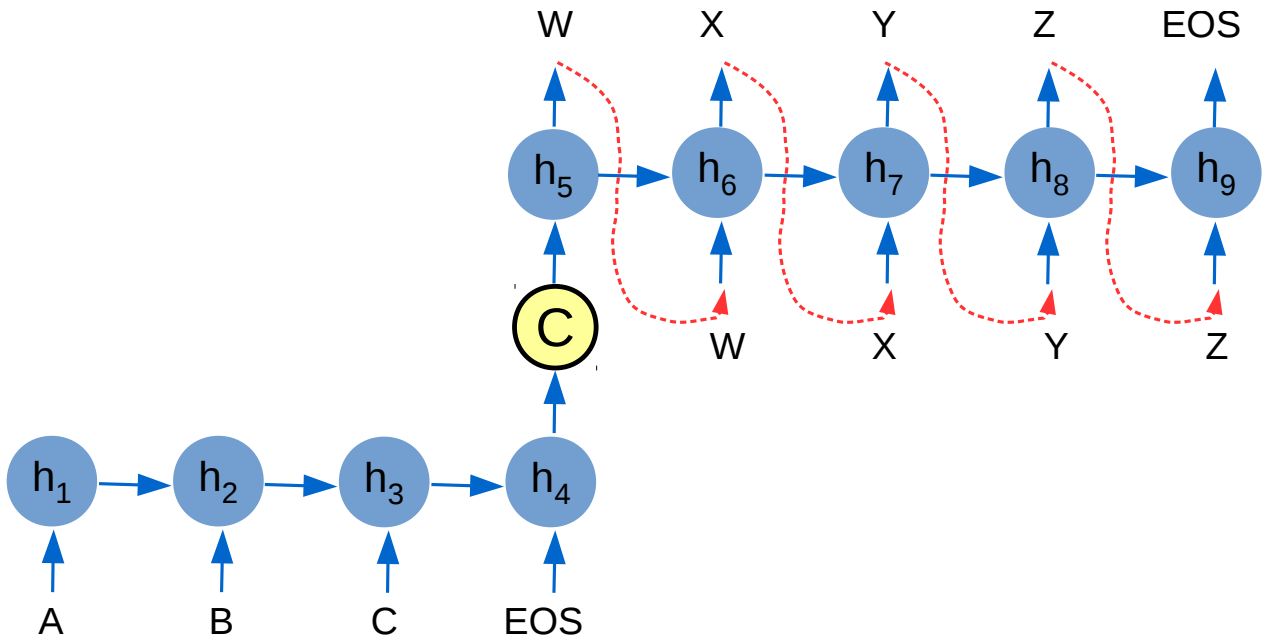
$$y = Wx$$

More complex configurations can also be used.



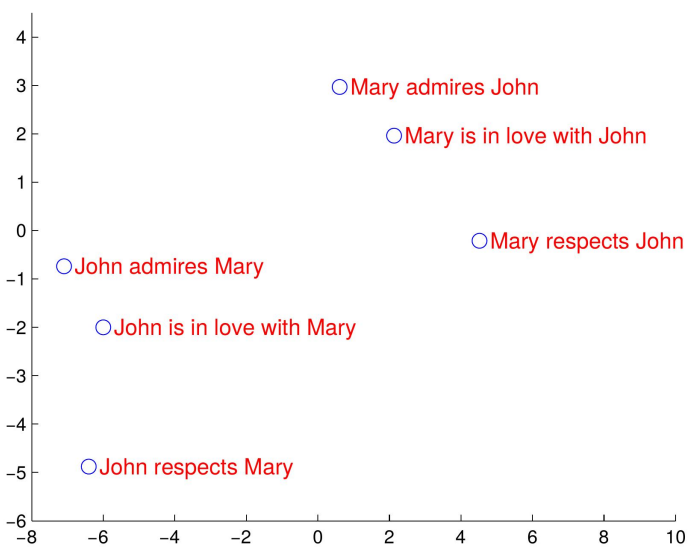
Ex. Language Translation (Sutskever et al.,2014)

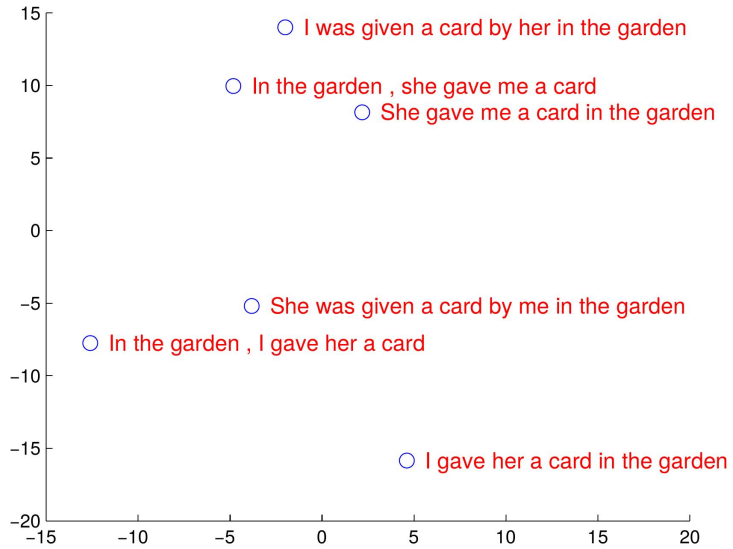
- Textual language translation application.
- Encoder reads an input sentence (ex. "ABC") and decoder produces an output sentence (ex."WXYZ").
- Each RNN consists of a deep LSTM model. Specifically, 4-layers deep.
- Contextual vector C corresponds to last hidden state of encoder LSTM.
- A special end-of-sentence symbol "< EOS >" is used to stop the decoding.
- During training, ground truth outputs are used as inputs to the decoder.



Ex. Language Translation (Sutskever et al.,2014)

- They show that the intermediate context vector C encodes each input sentence using a semantic space. Specifically, a space where sentences with similar meaning are close to each other.
- Figure shows a 2D PCA projection with examples of the context vector for different input sentences.





- Note that phrases are clustered by meaning, considering word order. This type of representation would be difficult to capture with a BoW model.

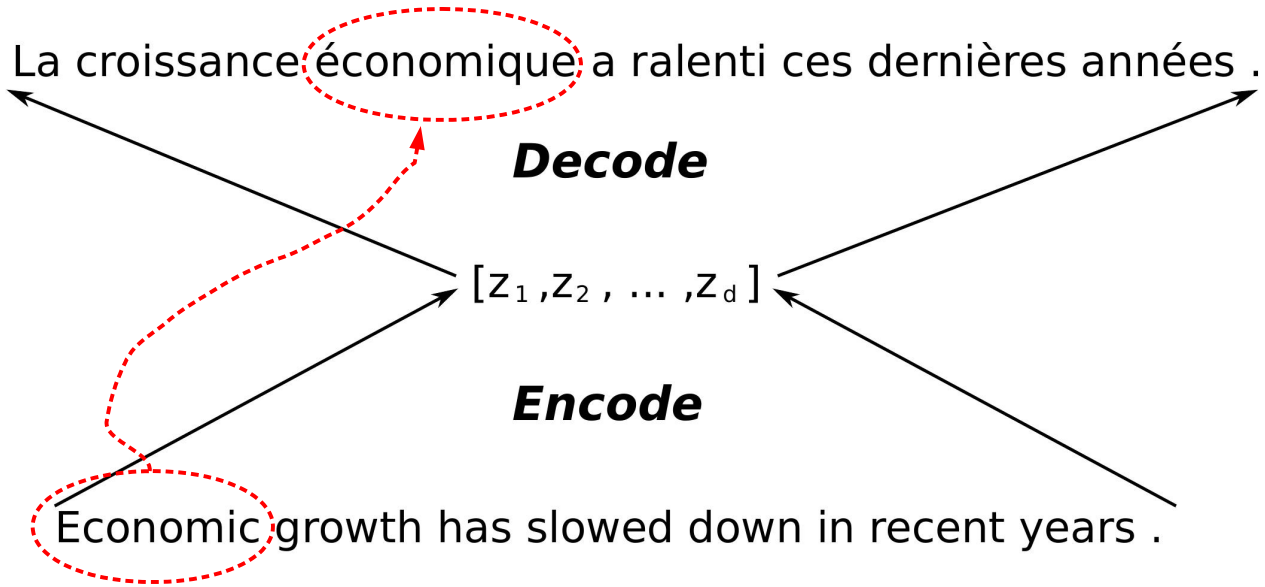
Seq2Seq Models With Attention

- In previous Seq2Seq model, encoder encodes input sequence into a fixed embedding or context vector C.
- This context vector encodes all relevant information from the input. I.e., it is the only connection from the encoder to the decoder.

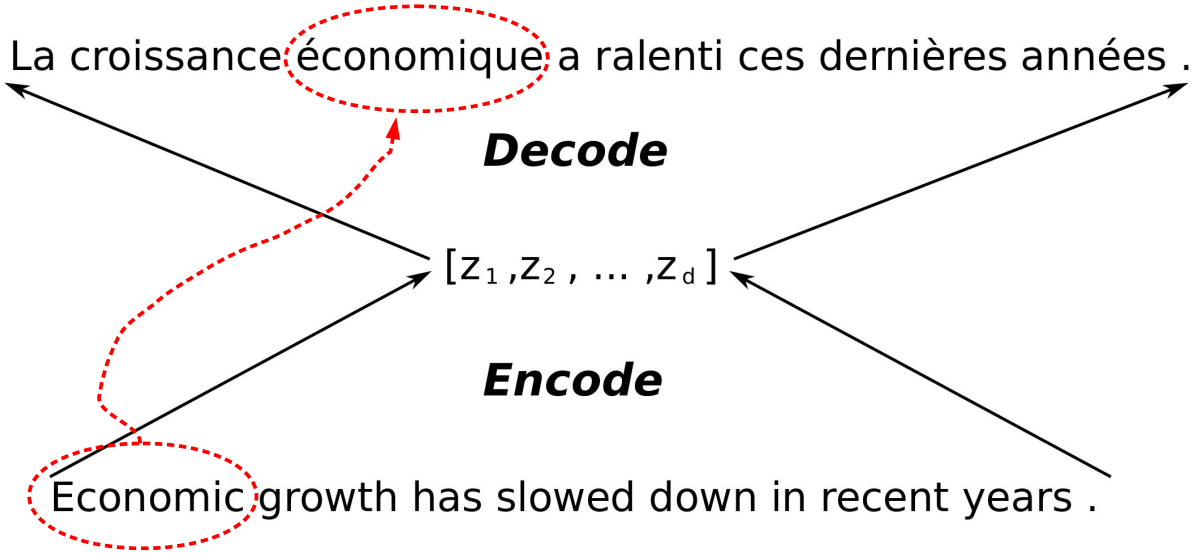
Seq2Seq Models With Attention

- In previous Seq2Seq model, encoder encodes input sequence into a fixed embedding or context vector C.
- This context vector encodes all relevant information from the input. I.e., it is the only connection from the encoder to the decoder.

Problem: As the decoding process progresses, it needs to consider different parts from the input.



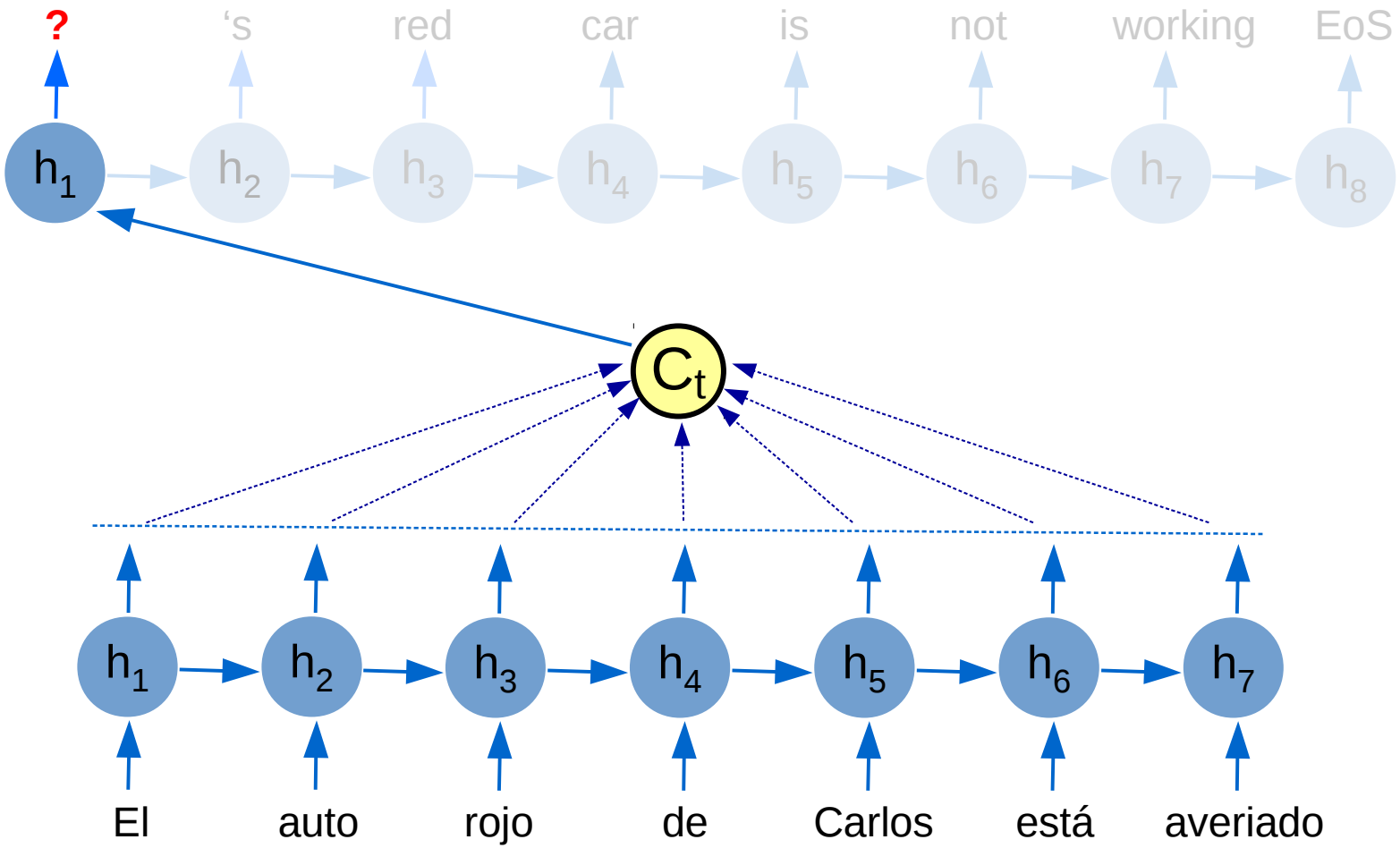
Idea: Add an Attention Mechanism

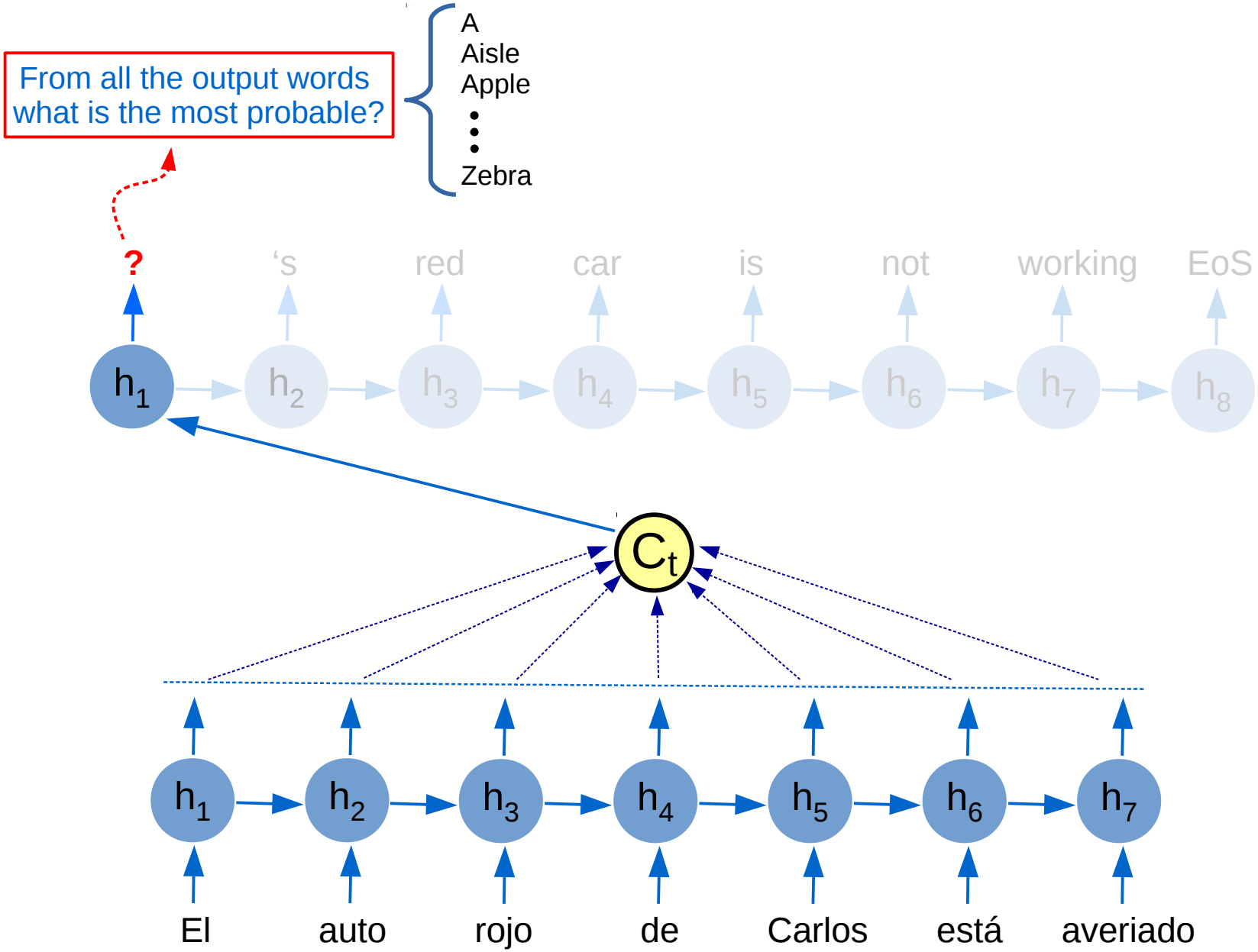


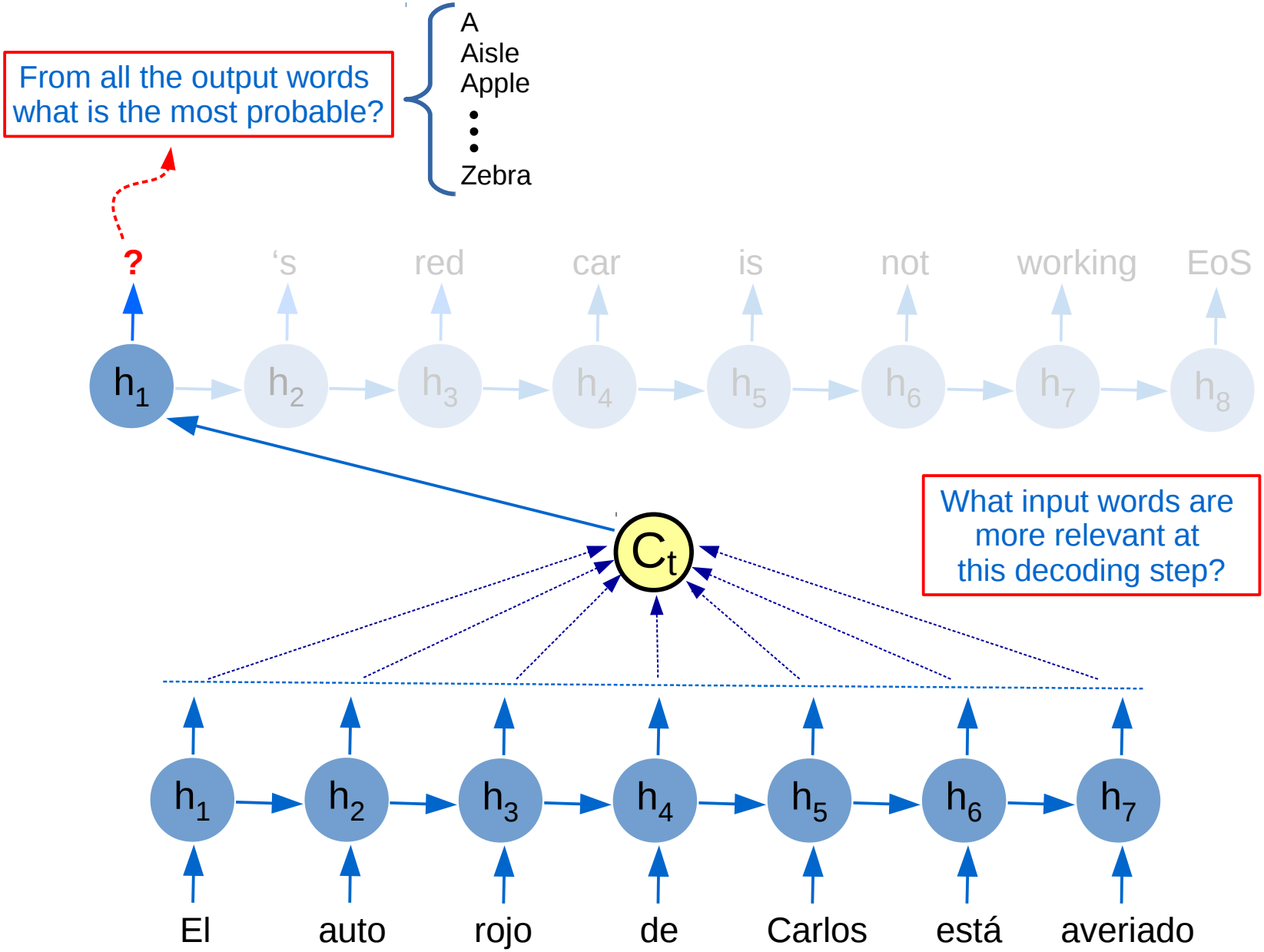
- Add an attention mechanism to adaptively identify the part of the input that the decoder needs to consider to generate the next output.
- In other words, the attention mechanism should align the output with the corresponding (relevant) data from the input sequence.
- At each decoding step, the decoder has access to selective pieces of information from the input sequence.

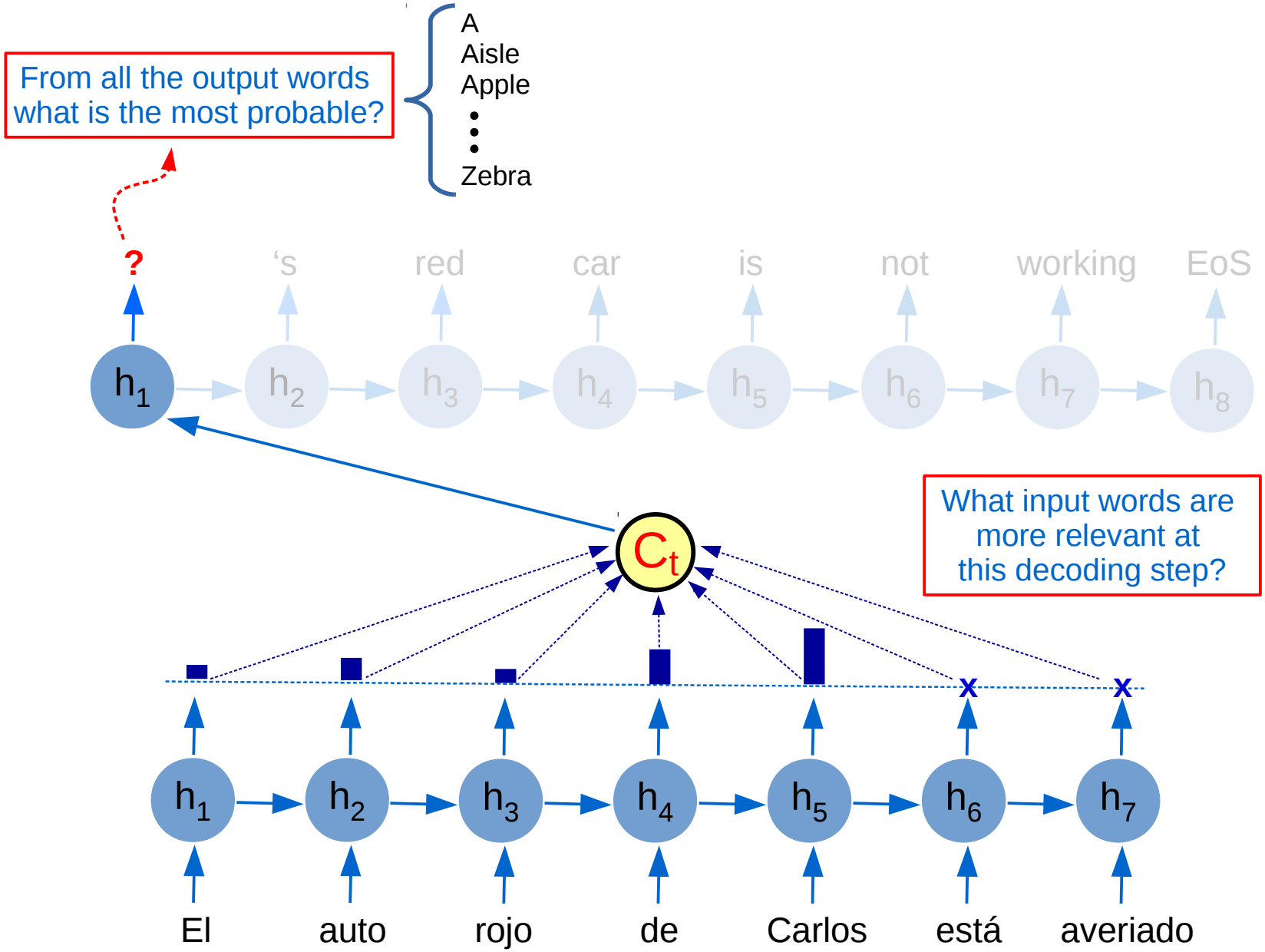
Example

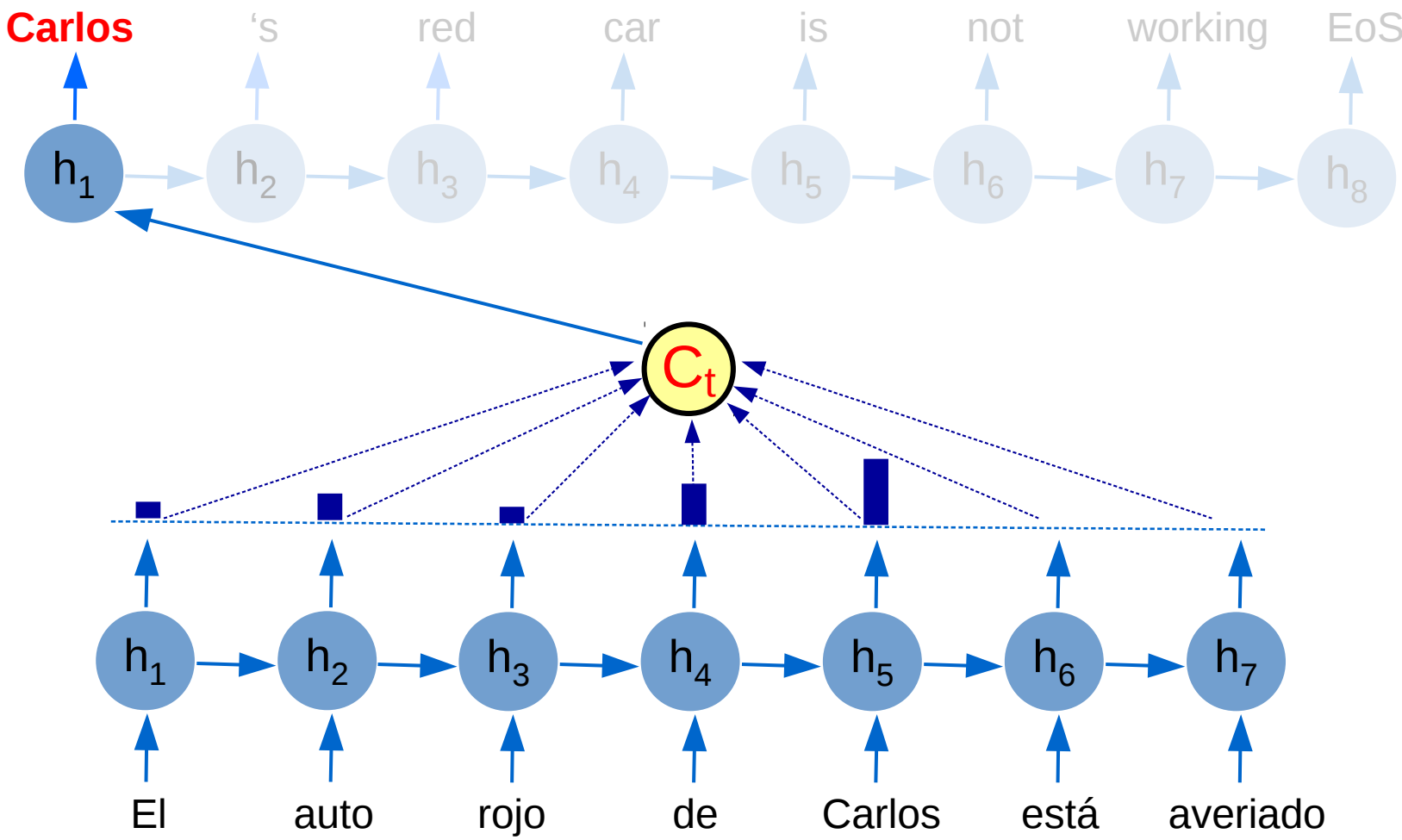
- Input: El auto rojo de Carlos está averiado.
- Output: Carlos 's red car is not working.

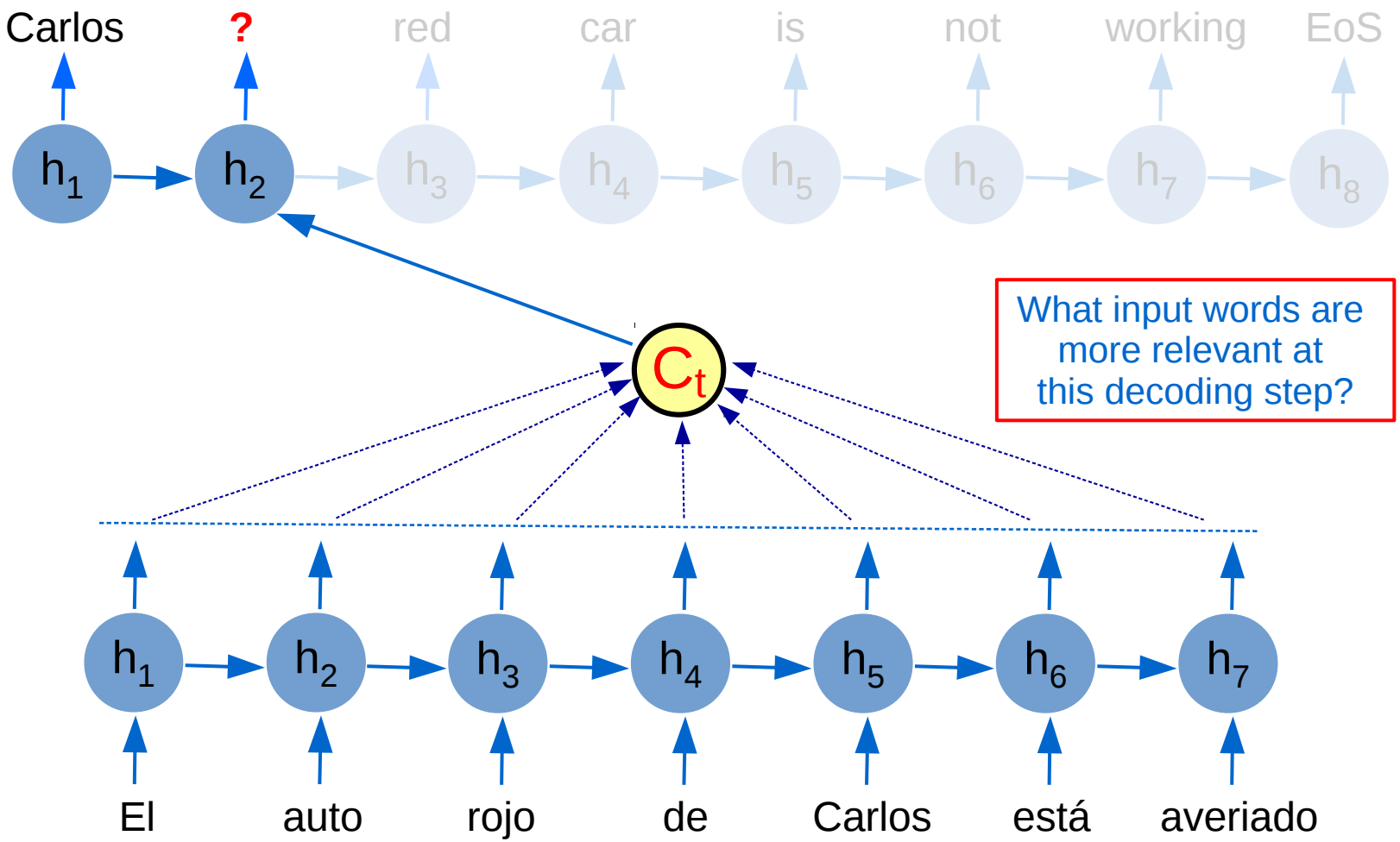


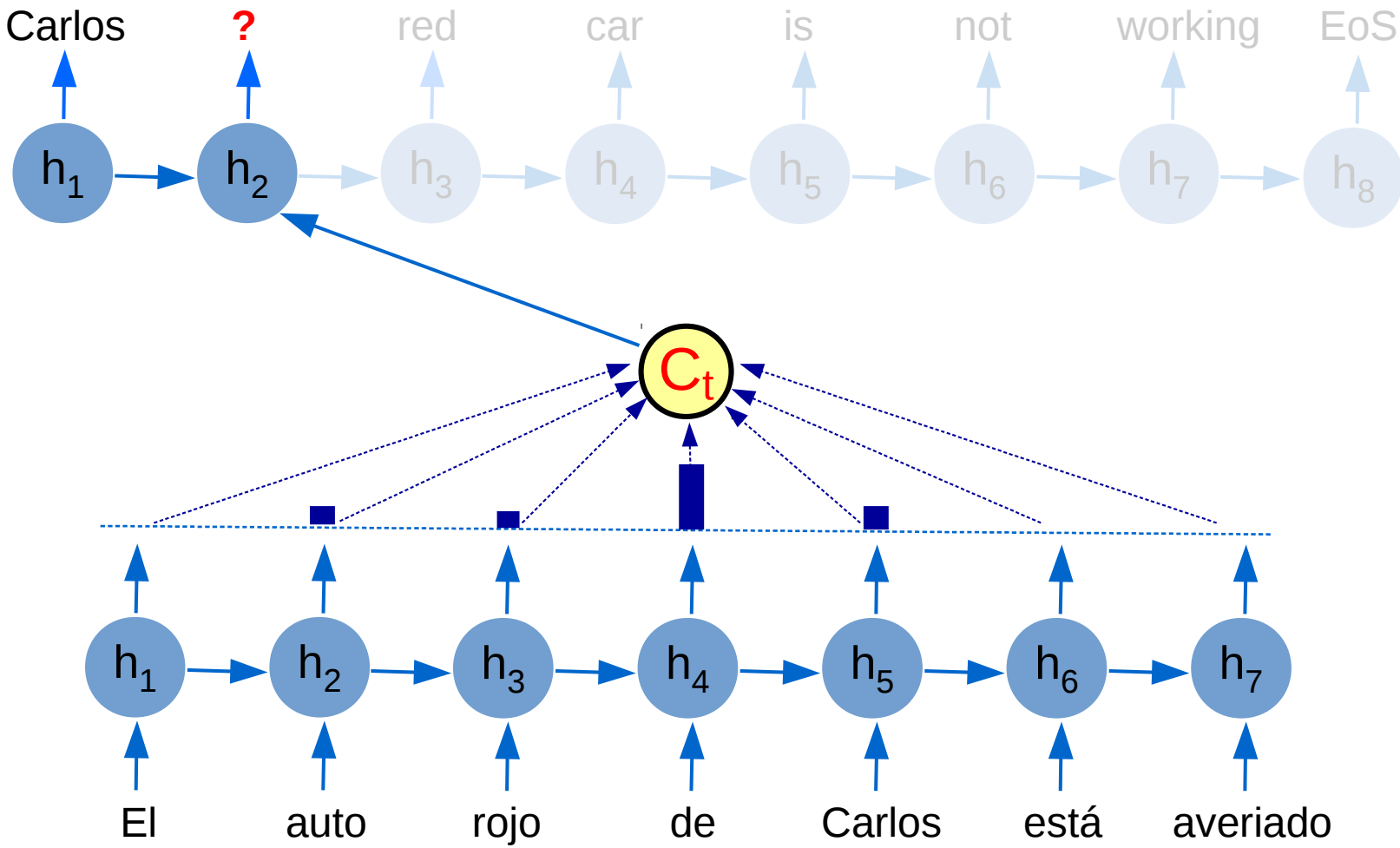

From all the output words what is the most probable?
A
Aisle
Apple
Zebra
?
's
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
El
auto
rojo
de
Carlos
está
averiado

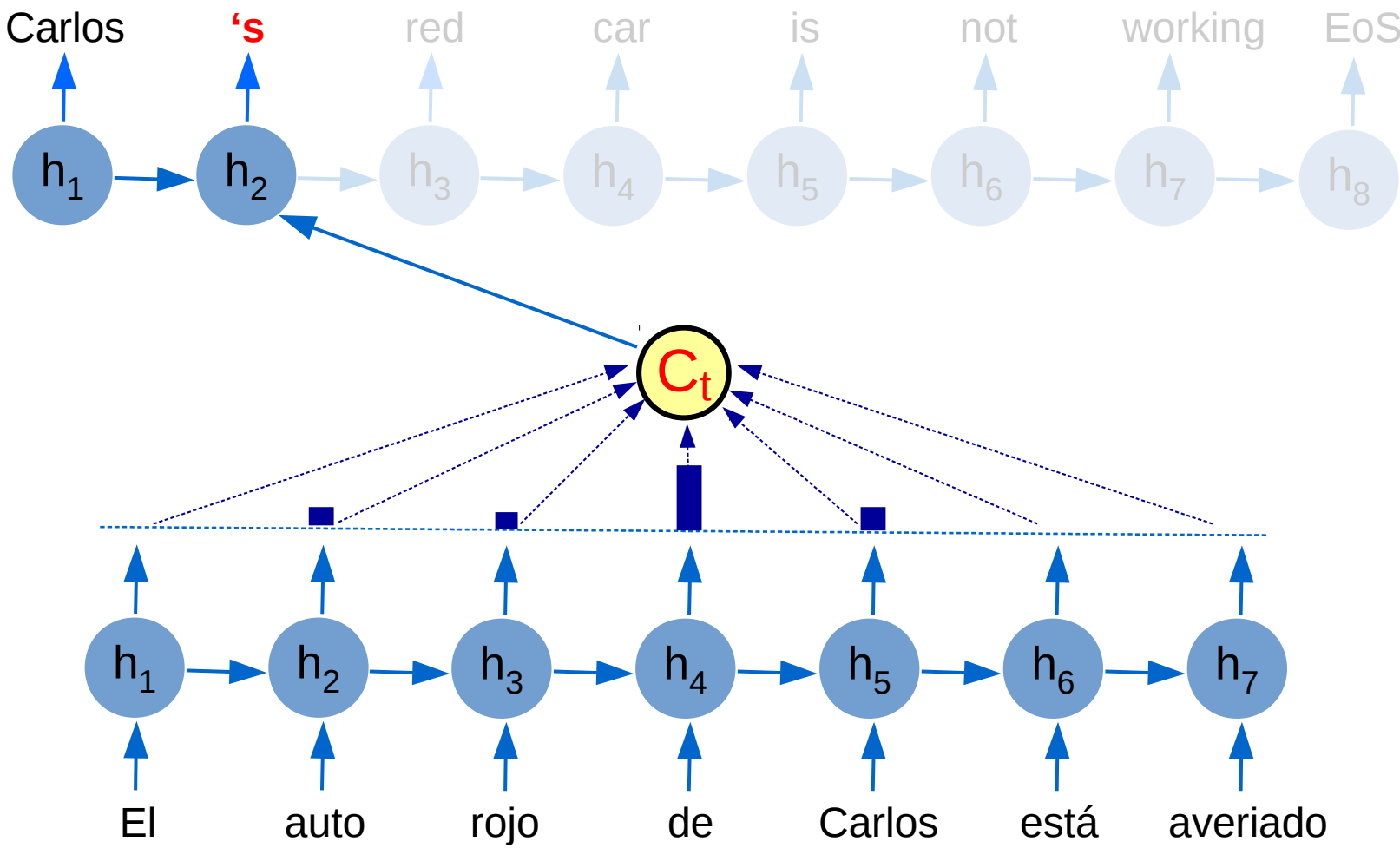

From all the output words what is the most probable?
A
Aisle
Apple
Zebra
?
's
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
What input words are more relevant at this decoding step?
El
auto
rojo
de
Carlos
está
averiado

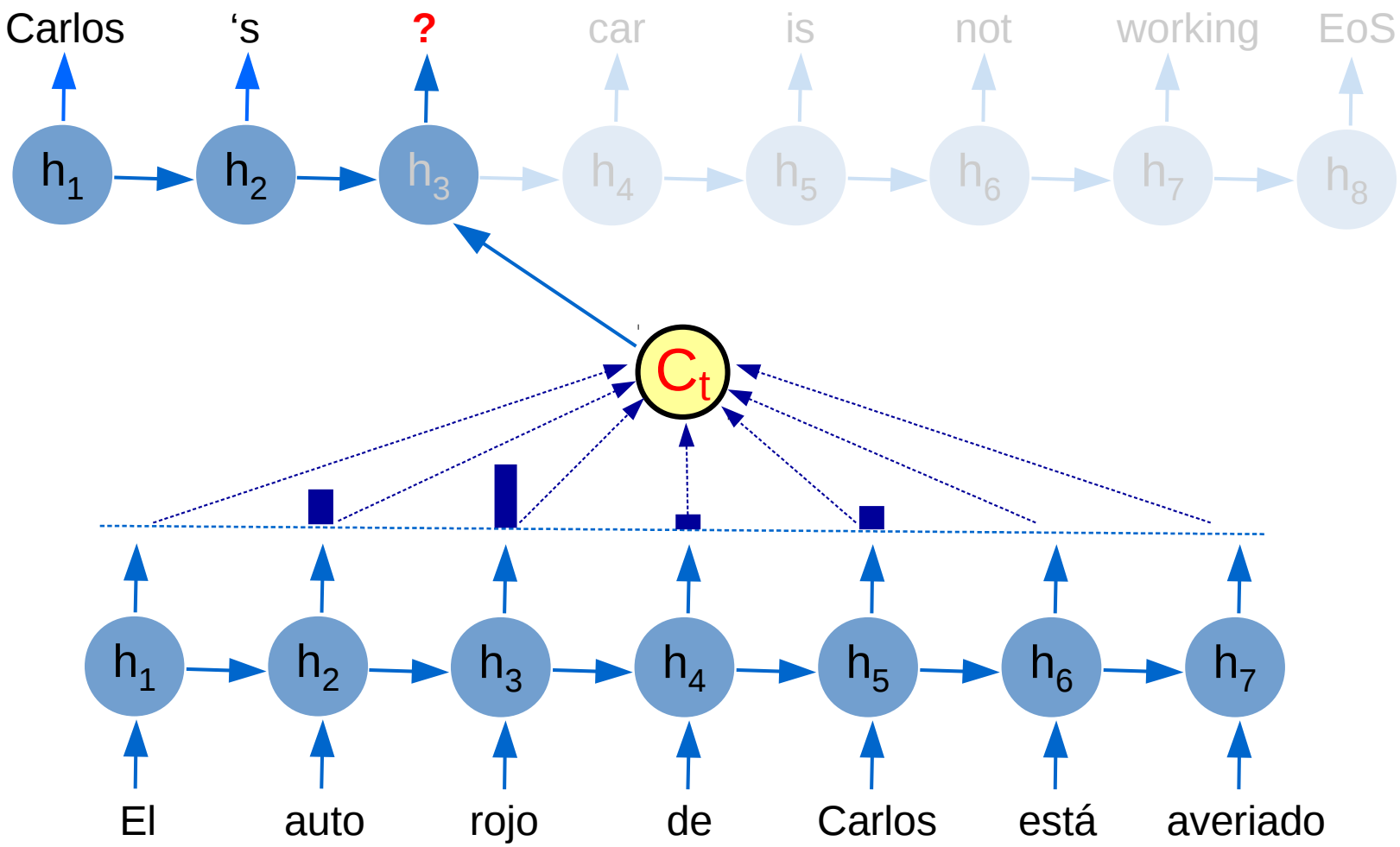

From all the output words what is the most probable?
A
Aisle
Apple
Zebra
?
's
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
What input words are more relevant at this decoding step?
Ct
El
auto
rojo
de
Carlos
está
averiado

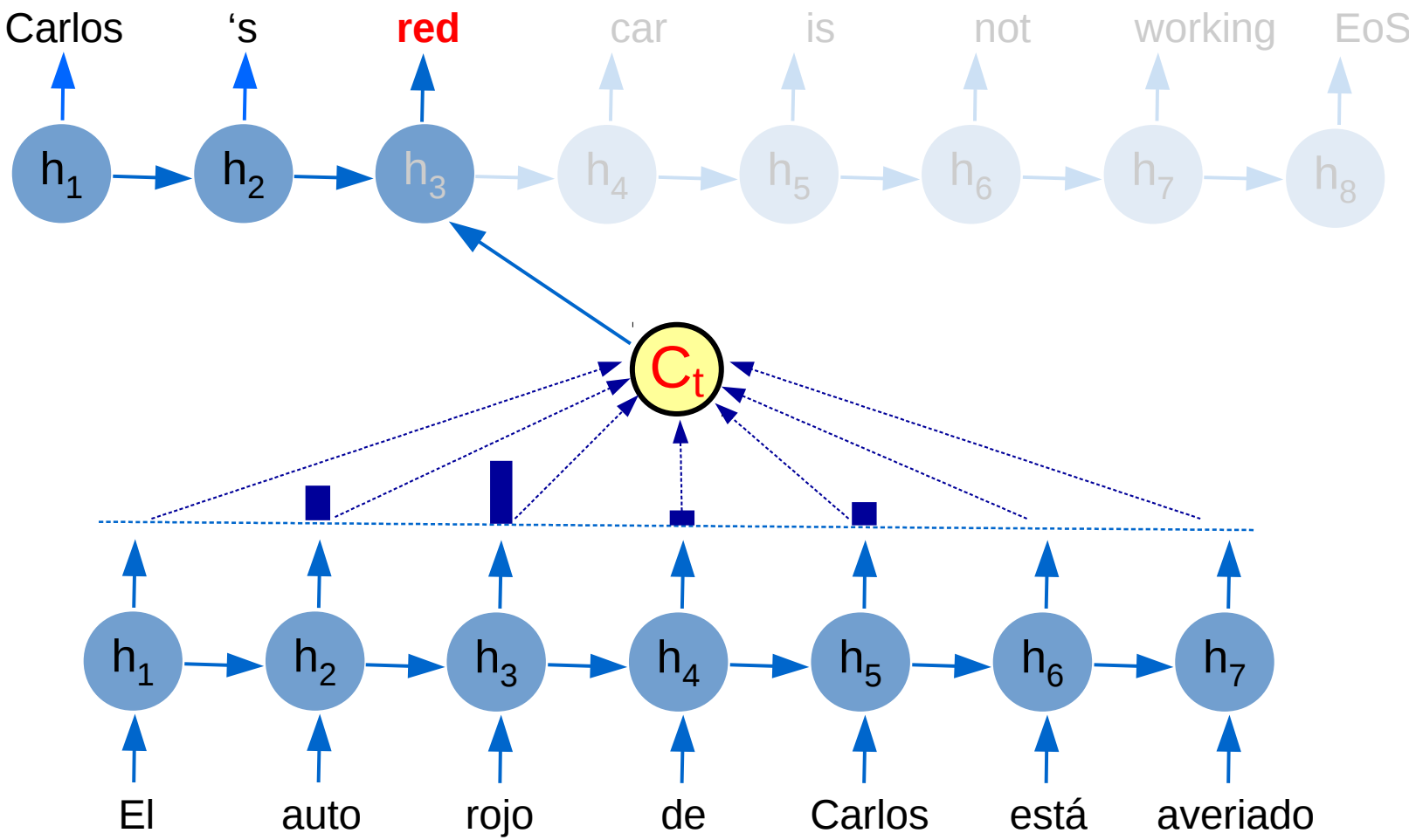

Carlos
's
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
h1
h2
h3
h4
h5
h6
h7
El
auto
rojo
de
Carlos
está
averiado

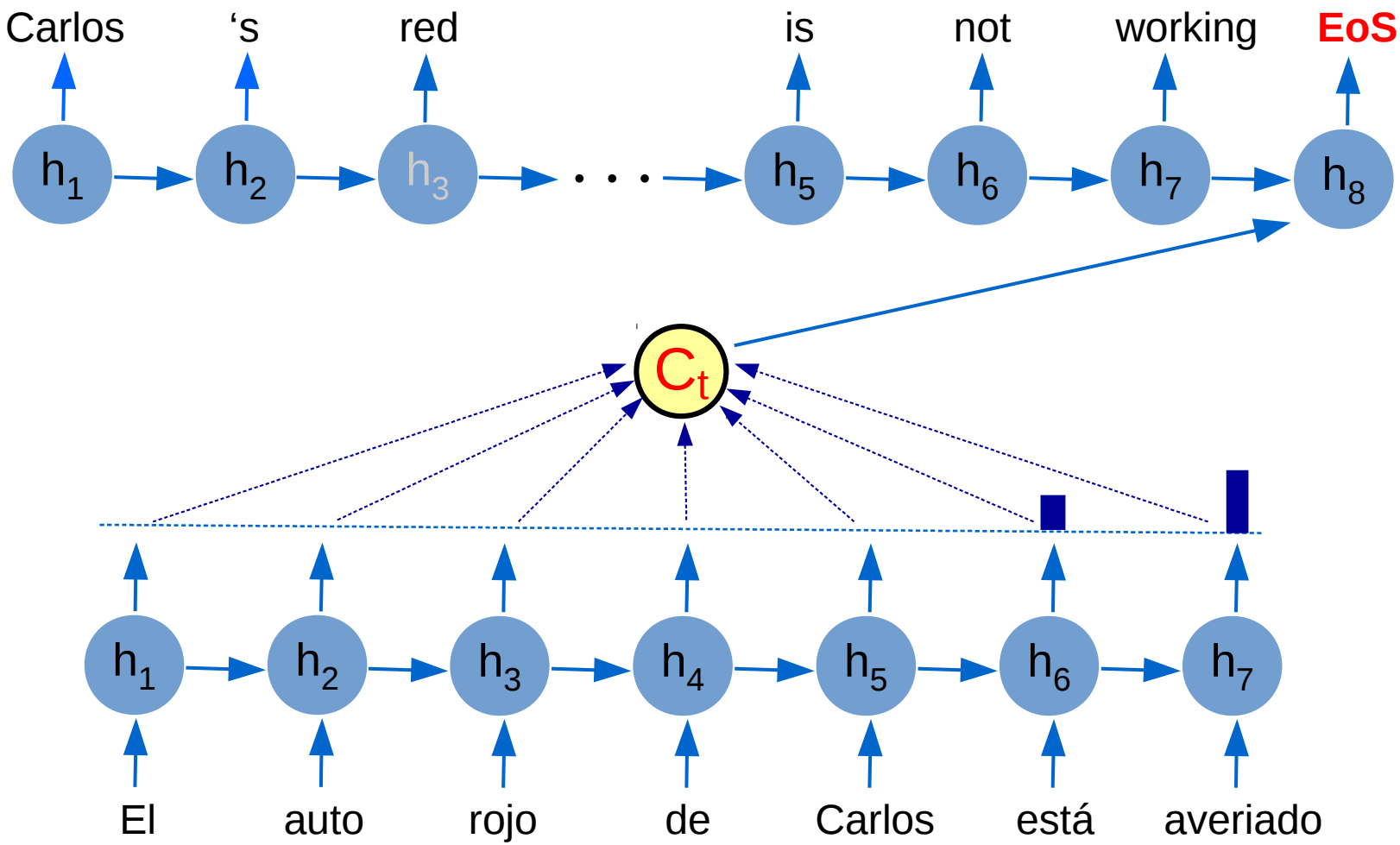

Carlos
?
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
What input words are more relevant at this decoding step?
El
auto
rojo
de
Carlos
está
averiado


Carlos
?
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
h1
h2
h3
h4
h5
h6
h7
El
auto
rojo
de
Carlos
está
averiado


Carlos
's
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
El
auto
rojo
de
Carlos
está
averiado

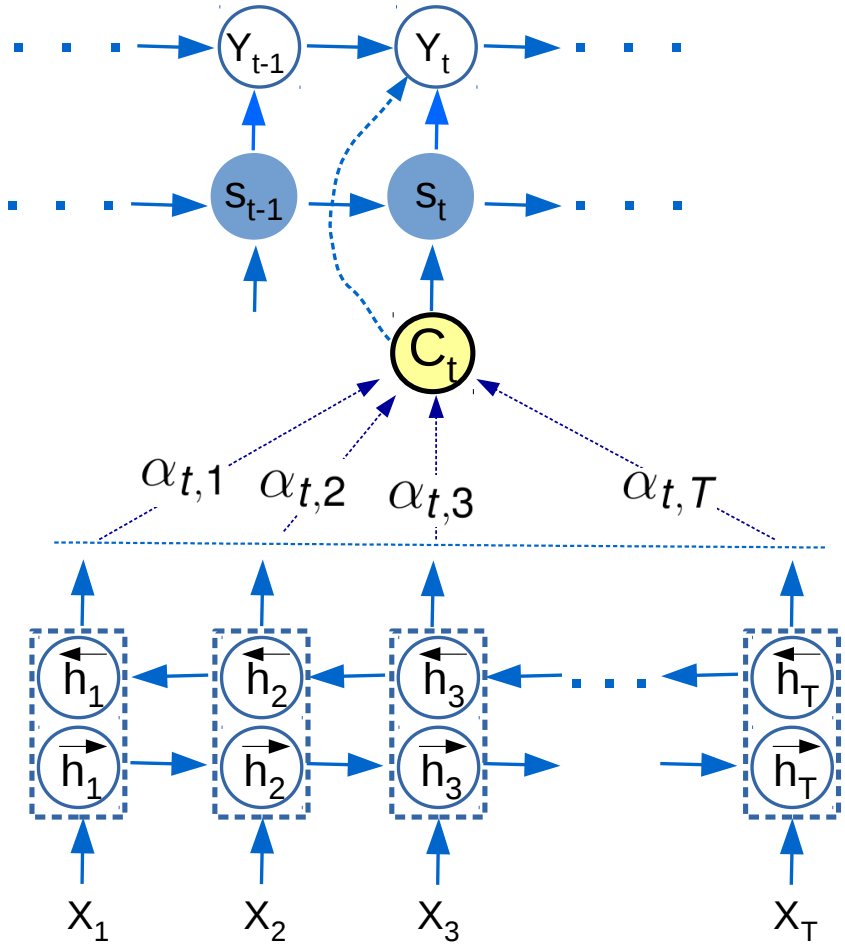

Carlos
‘s
?
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
h1
h2
h3
h4
h5
h6
h7
El
auto
rojo
de
Carlos
está
averiado


Carlos
‘s
red
car
is
not
working
EoS
h1
h2
h3
h4
h5
h6
h7
h8
Ct
h1
h2
h3
h4
h5
h6
h7
El
auto
rojo
de
Carlos
está
averiado


Carlos
‘s
red
is
not
working
EoS
h1
h2
h3
h5
h6
h7
h8
Ct
h1
h2
h3
h4
h5
h6
h7
El
auto
rojo
de
Carlos
está
averiado

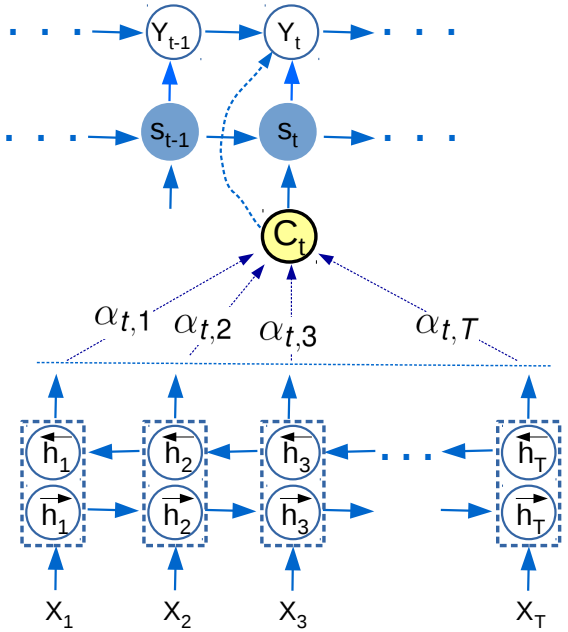
How can we estimate the attention weights?

A popular approach was proposed by Bahdanau et al., ICLR 2015



Notice that now the context vector C_t changes at every step

Attention models (Bahdanau et al., ICLR 2015)



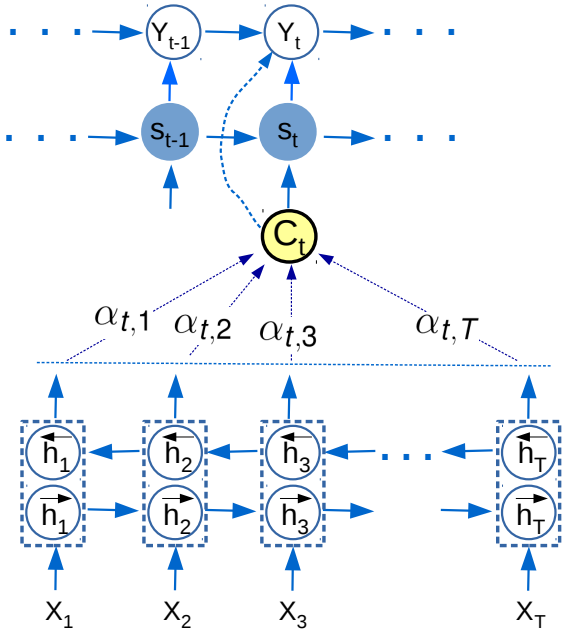
- Decoder generates output y_t according to an adaptive context C_t.
- The math follows the information flow:

$$y_t = \sigma(W_{yy} y_{t-1} + W_{sy} s_t + W_{cy} C_t)$$

$$s_t = \sigma(W_{ss} s_{t-1} + W_{cs} C_t)$$

$$C_t = \sum_{i=1}^{T} \alpha_{t,i} <\overrightarrow{h}_i, \overleftarrow{h}_i>$$

Attention models (Bahdanau et al., ICLR 2015)



- Decoder generates output y_t according to an adaptive context C_t.
- The math follows the information flow:

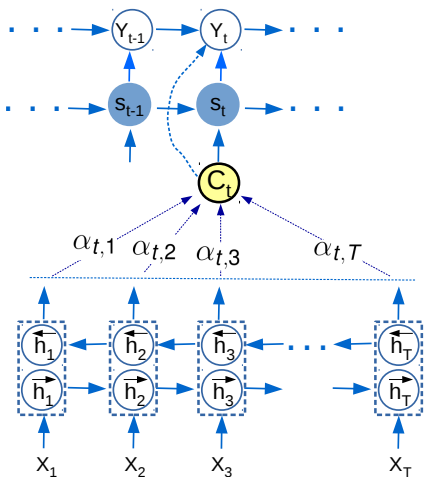
$$y_t = \sigma(W_{yy} y_{t-1} + W_{sy} s_t + W_{cy} C_t)$$

$$s_t = \sigma(W_{ss} s_{t-1} + W_{cs} C_t)$$

$$C_t = \sum_{i=1}^{T} \alpha_{t,i} <\overrightarrow{h}_i, \overleftarrow{h}_i>$$

- At each step t, context vector C_t is given by a weighted average of the hidden states $h_j = <\overrightarrow{h}_i, \overleftarrow{h}_i>$ of the bidirectional LSTM run by the encoder.
- α_{tj} encodes relevance of hidden state h_j to generate output y_t at step t.
- This weighting mechanism is called soft-attention. It has the advantage to be differentiable (good for SGD).
- Using this mechanism, the model is able to focus on different parts of the input sequence when generating each translated word.
- The use of a bidirectional RNN allows the context vector C_t to consider dependencies in both ways.

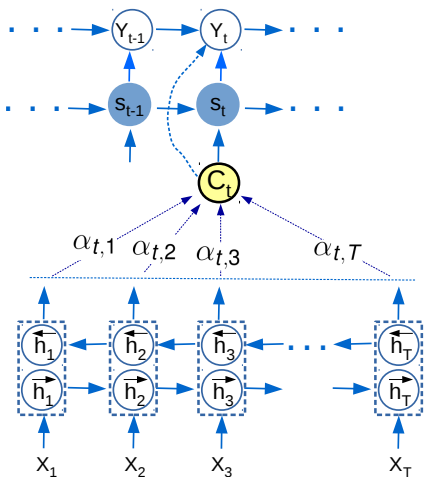
Attention models (Bahdanau et al., ICLR 2015)



- How do we estimate attention coefficients α_{tj} ?.
- α_{tj} estimates how previous output context s_{t-1} and hidden state h_j affects the context s_t used to generate output y_t.
- We can write this mathematically:

$$\hat{\alpha}_{tj} = V_c^T \sigma(W_c s_{t-1} + U_c h_j),$$
$$V_c \in \mathbb{R}^n, W_c \in \mathbb{R}^{nxn}, U_c \in \mathbb{R}^{nx2n}$$

Attention models (Bahdanau et al., ICLR 2015)



- How do we estimate attention coefficients α_{tj} ?.
- α_{tj} estimates how previous output context s_{t-1} and hidden state h_j affects the context s_t used to generate output y_t.
- We can write this mathematically:

$$\hat{\alpha}_{tj} = V_c^T \sigma(W_c s_{t-1} + U_c h_j),$$
$$V_c \in \mathbb{R}^n, W_c \in \mathbb{R}^{nxn}, U_c \in \mathbb{R}^{nx2n}$$

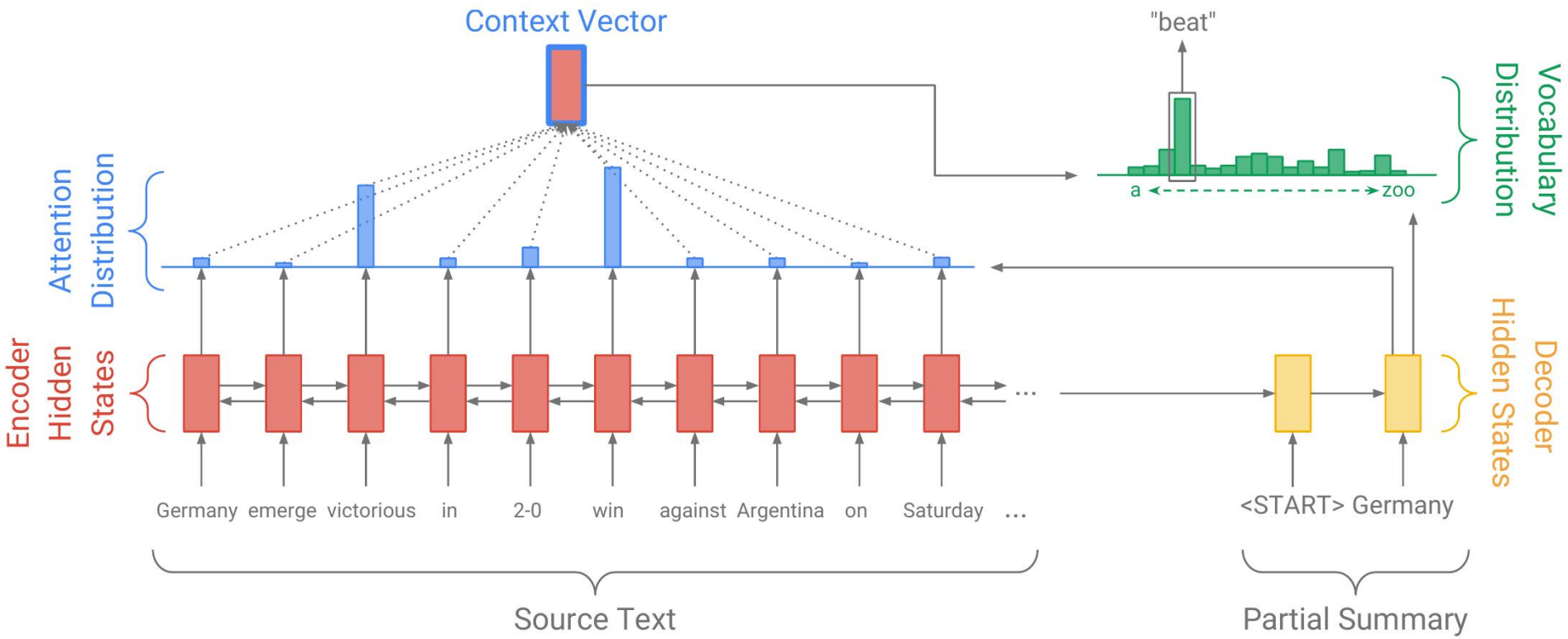
Normalizing the coefficients: $\alpha_{t,j} = \frac{\hat{\alpha}_{t,j}}{\sum_k \hat{\alpha}_{tk}}$

- Notice that decoder hidden state s_{t-1} is used to calculate the attention weights, why?
- By using s_{t-1} to generate the attention coefficients, the model tracks the decoder's progress (memory of the output so far!).
- In summary, each α_{tj} is estimated using an additive model that combines embedding of previous output context s_{t-1} and its corresponding input source h_j from the decoder.

Attention model as a selective memory access process

- We can consider the attention mechanism as a memory access process.
- Specifically, a memory access step consists of a triplet:<query, keys, values>, where the query matches the keys to access the values.
- For the previous attention mechanism we can consider:
  - query: previous hidden state of the decoder S_{t-1}.
  - keys: encoder hidden states h_t, ($t = [1 \dots T]$).
  - values: encoder hidden states h_t, ($t = [1 \dots T]$).
- This abstraction, that connects an attention mechanism to a memory access process, is a key idea to create intelligent machines.
- In general, keys, values, and queries could be anything. So the relevant process is to: i) Ask the right question (query), ii) Consider the right information (keys), and iii) Provide a suitable answer (values or a mix of them).

Ex. Attention model: Summarizing Documents See and Manning, ACL 2017



Ex. Attention model: Summarizing Documents See and Manning, ACL 2017

Article: smugglers lure arab and african migrants by offering discounts to get onto overcrowded ships if people bring more potential passengers, a cnn investigation has revealed. (...)
Summary: cnn investigation **uncovers** the **business inside** a **human smuggling ring**.

Article: eyewitness video showing white north charleston police officer michael slager shooting to death an unarmed black man has exposed discrepancies in the reports of the first officers on the scene. (...)
Summary: more **questions than answers emerge** in **controversial s.c.** police shooting.

- Abstractive vs extractive summarization.
- These are examples of abstractive summarization, bold denotes novel words.

Article (truncated): andy murray came close to giving himself some extra preparation time for his wedding next week before ensuring that he still has unfinished tennis business to attend to . the world no 4 is into the semi-finals of the miami open , but not before getting a scare from 21 year-old austrian dominic *thiem* , who pushed him to 4-4 in the second set before going down 3-6 6-4 , 6-1 in an hour and three quarters . murray was awaiting the winner from the last eight match between tomas berdych and argentina 's juan monaco . prior to this tournament *thiem* lost in the second round of a challenger event to soon-to-be new brit *aljaz* bedene . andy murray pumps his first after defeating dominic *thiem* to reach the miami open semi finals . *muray* throws his *sweatband* into the crowd after completing a 3-6 , 6-4 , 6-1 victory in florida . murray shakes hands with *thiem* who he described as a ' strong guy ' after the game . and murray has a fairly simple message for any of his fellow british tennis players who might be agitated about his imminent arrival into the home ranks : do n't complain . instead the british no 1 believes his colleagues should use the assimilation of the world number 83 , originally from slovenia , as motivation to better themselves .

andy murray defeated dominic *thiem* 3-6 6-4 , 6-1 in an hour and three quarters .
murray was awaiting the winner from the last eight match between tomas berdych and argentina 's juan monaco .
prior to this tournament thiem lost in the second round of a challenger event to soon-to-be new brit *aljaz* bedene .

- Italics denote out-of-vocabulary words
- Green shading represents value of generation probability.
- Yellow shading represents a coverage vector (see paper for details).

Article (truncated): munster have signed new zealand international francis *saili* on a two-year deal . utility back *saili* , who made his all blacks debut against argentina in 2013 , will move to the province later this year after the completion of his 2015 contractual commitments . the 24-year-old currently plays for *auckland-based* super rugby side the blues and was part of the new zealand under-20 side that won the junior world championship in italy in 2011 . *saili* 's signature is something of a coup for munster and head coach anthony foley believes he will be a great addition to their backline . francis *saili* has signed a two-year deal to join munster and will link up with them later this year . ' we are really pleased that francis has committed his future to the province , ' foley told munster 's official website . ' he is a talented centre with an impressive *skill-set* and he possesses the physical attributes to excel in the northern hemisphere . ' i believe he will be a great addition to our backline and we look forward to welcoming him to munster . ' *saili* has been capped twice by new zealand and was part of the under 20 side that won the junior championship in 2011 .

francis *saili* has signed a two-year deal to join munster later this year .
the 24-year-old was part of the new zealand under-20 side that won the junior world championship in italy in 2011 .
saili 's signature is something of a coup for munster and head coach anthony foley .

- Italics denote out-of-vocabulary words
- Green shading represents value of generation probability.
- Yellow shading represents a coverage vector (see paper for details).

Ex. Attention models: Image captioning

A woman with a little girl in a park, the woman is throwing a fresbee.

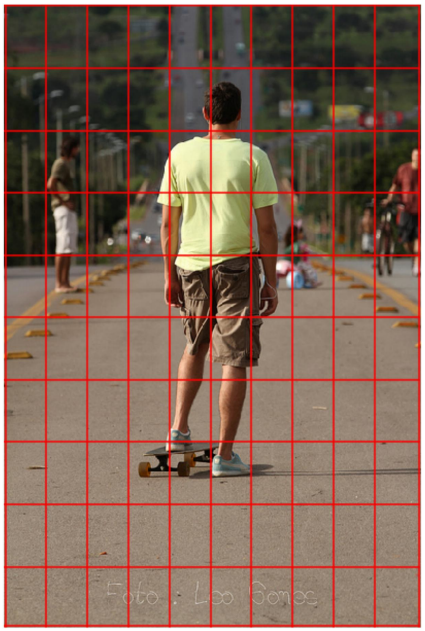
- Image captioning can be considered as a translation problem between image and language modalities.
- In contrast to textual inputs, the visual world poses two complications:
  1. Spatial location of relevant structures (word order) is not predefined.
  2. List of valid words (visual dictionary) is unknown.

1. Spatial location of relevant structures (word order) is not predefined.

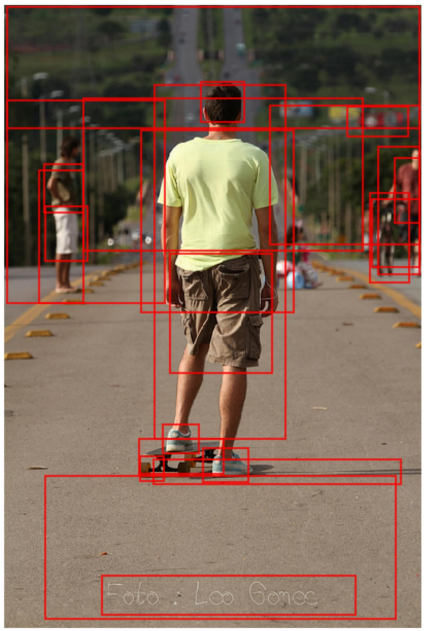
Spatial locations can be identified using different schemes, ex:

- Regular grid of patches (Xu, ICML-15).
- Objectness detector (Shih et al., CVPR-16).
- Bank of attribute detectors (You et al., CVPR 2016).
- Among others.

Ex. Regular grid

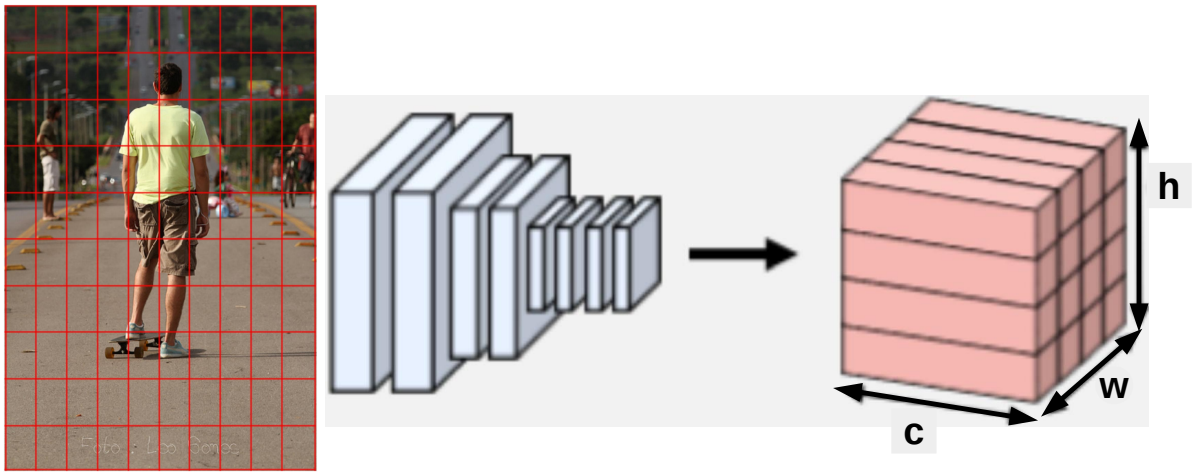


Ex. Objectness



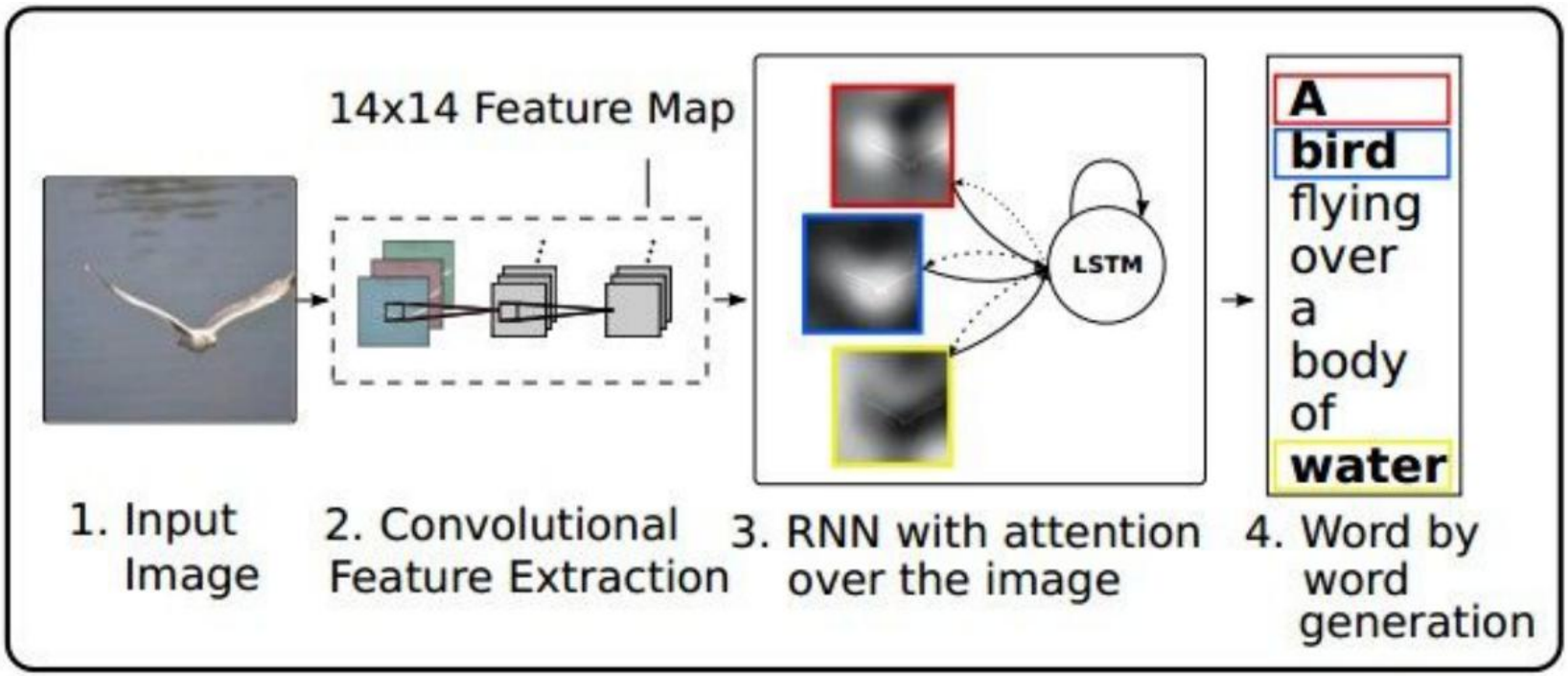
2. List of valid words (visual dictionary) is unknown.

- Idea: Use as input embedding (visual dictionary) the outputs of a pre-trained CNN (ex. last layers of Alex or VGG).
- To accomodate for region level encoding, use convolutional layers instead of fully connected layers.

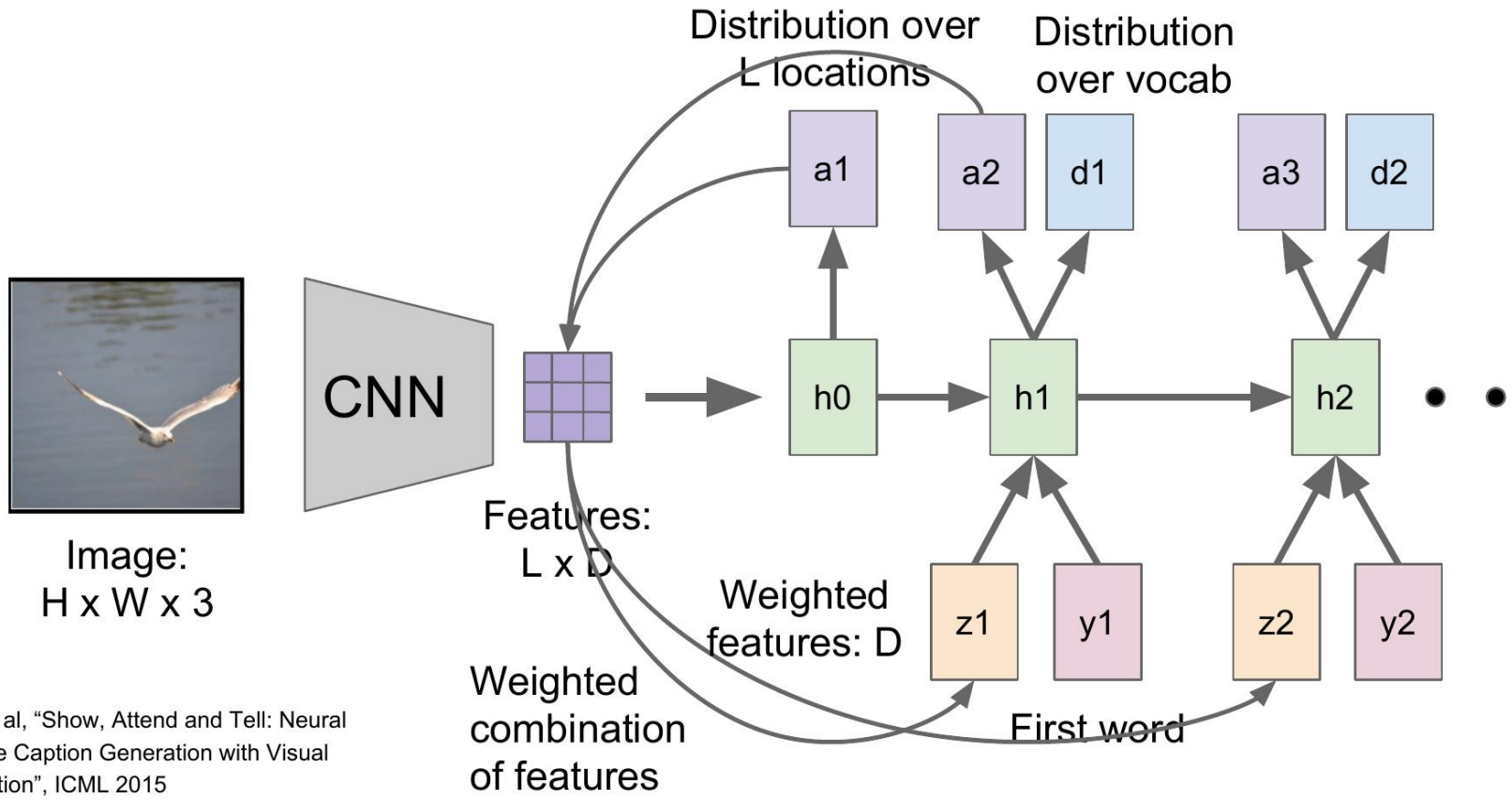


- Input image is embedded as a sequence of wxh visual words of c dimensions.
- Please keep in mind that cells in output grid encode information coming from most parts of the input image.

Attention model: Image captioning
Xu et al., ICML 2015



Attention model: Image captioning
Xu et al., ICML 2015



Xu et al, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", ICML 2015

Attention model: Image captioning
Xu et al., ICML 2015

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

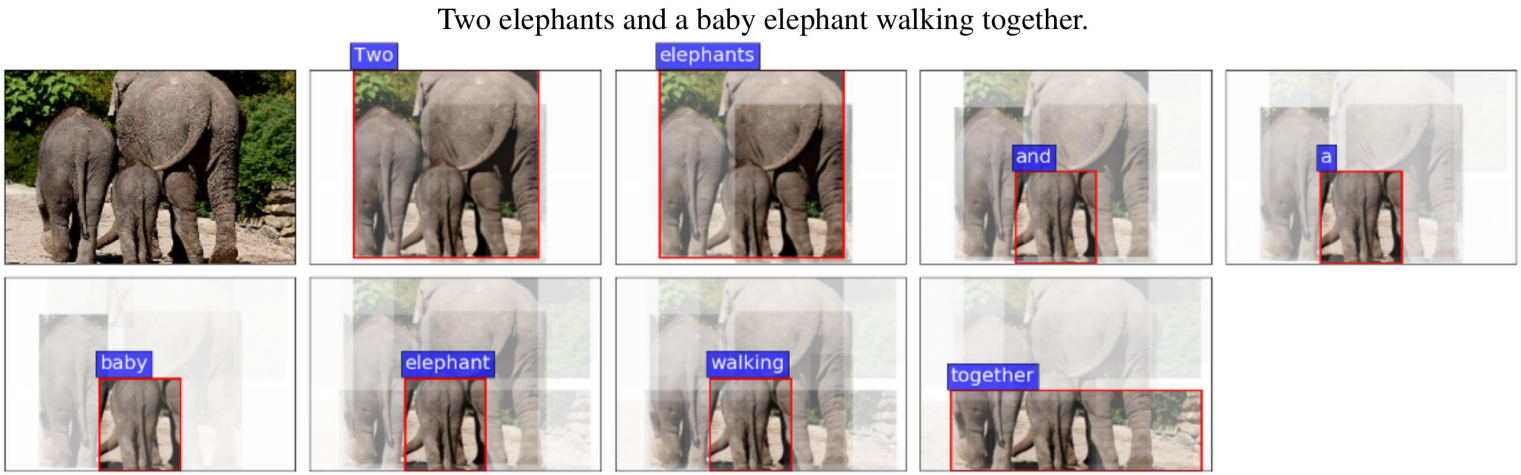
A stop sign is on a road with a mountain in the background.

A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Ex. Attention model based on objectness
Anderson et al., CVPR 2018



Ex. Attention model based on objectness Anderson et al., CVPR 2018