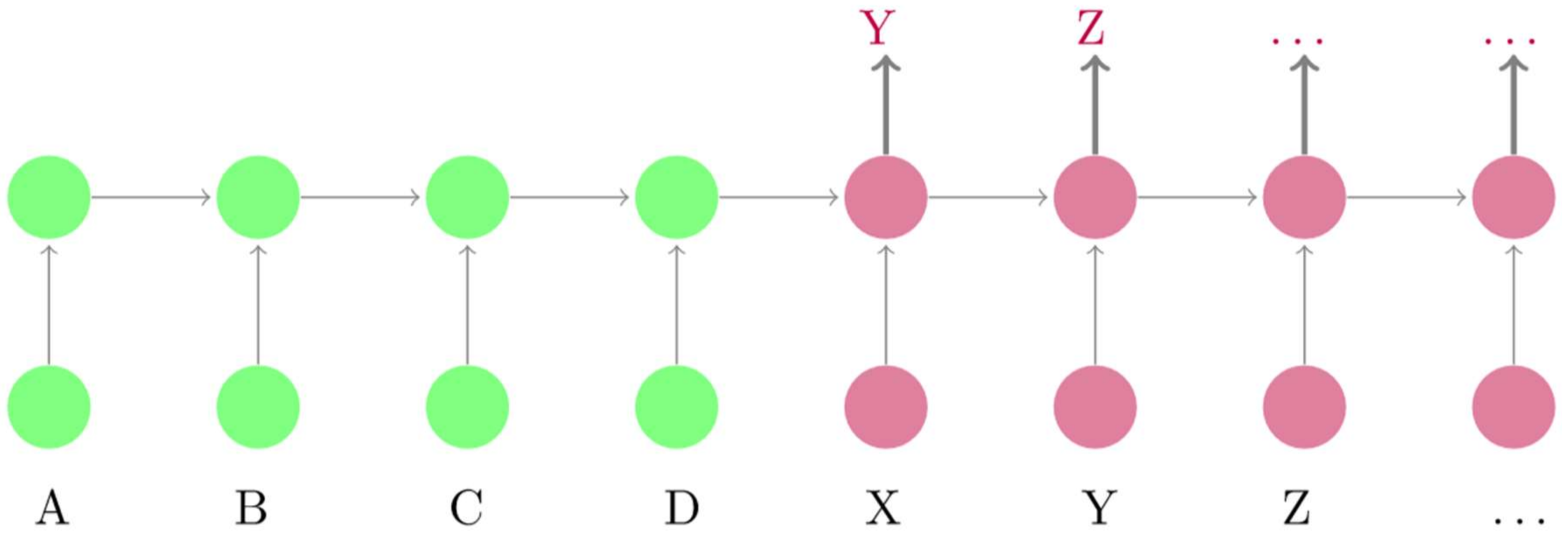
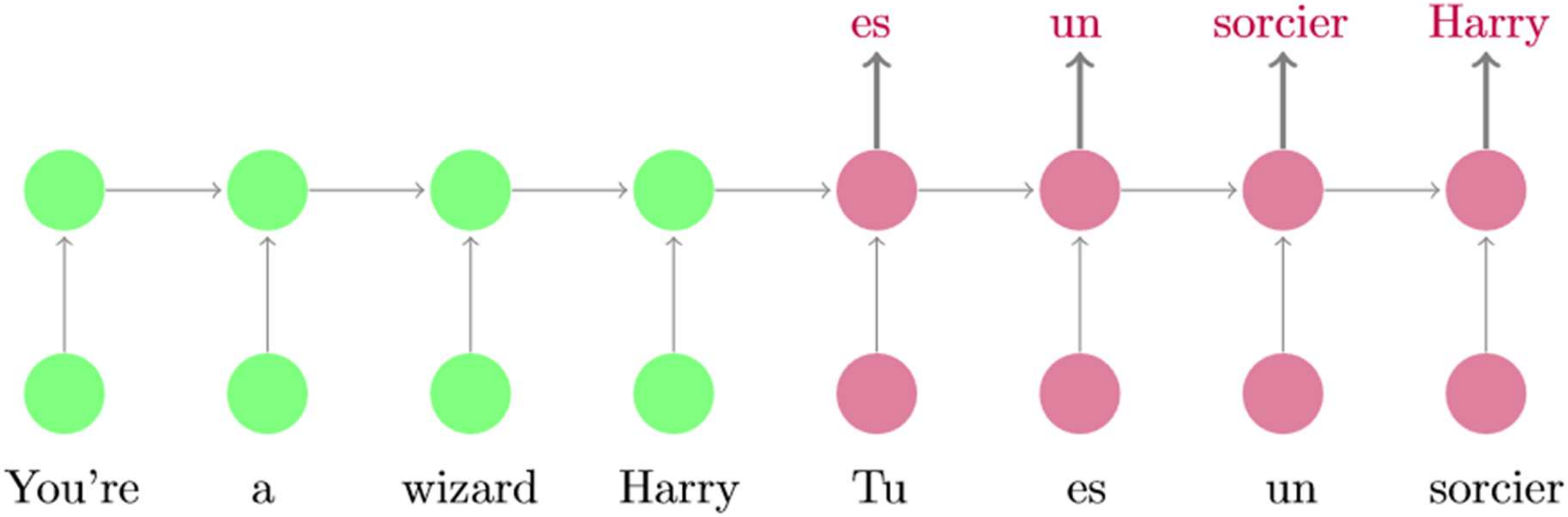
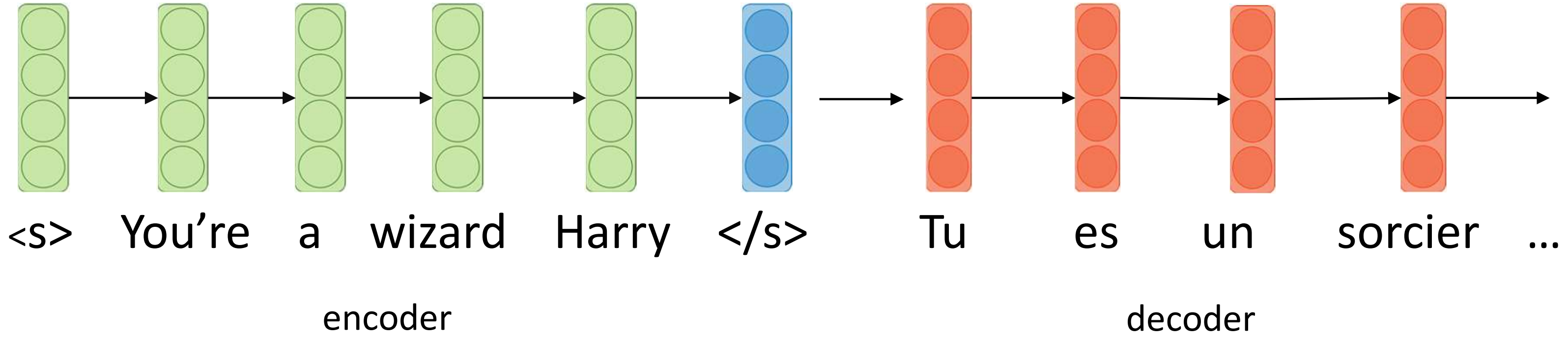
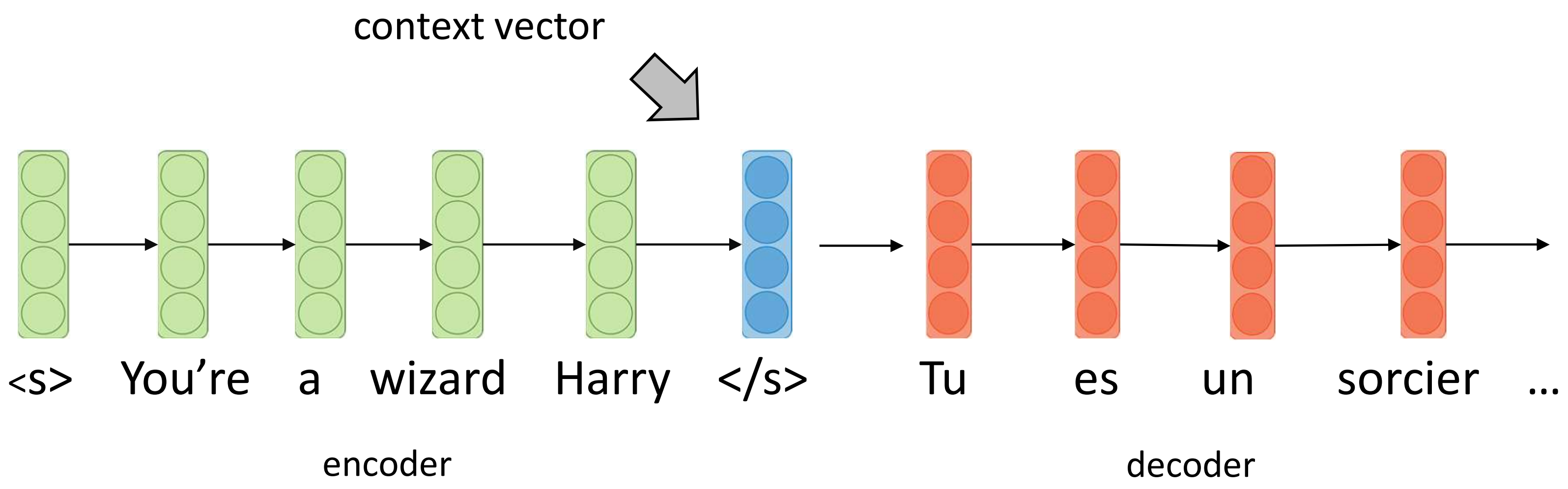
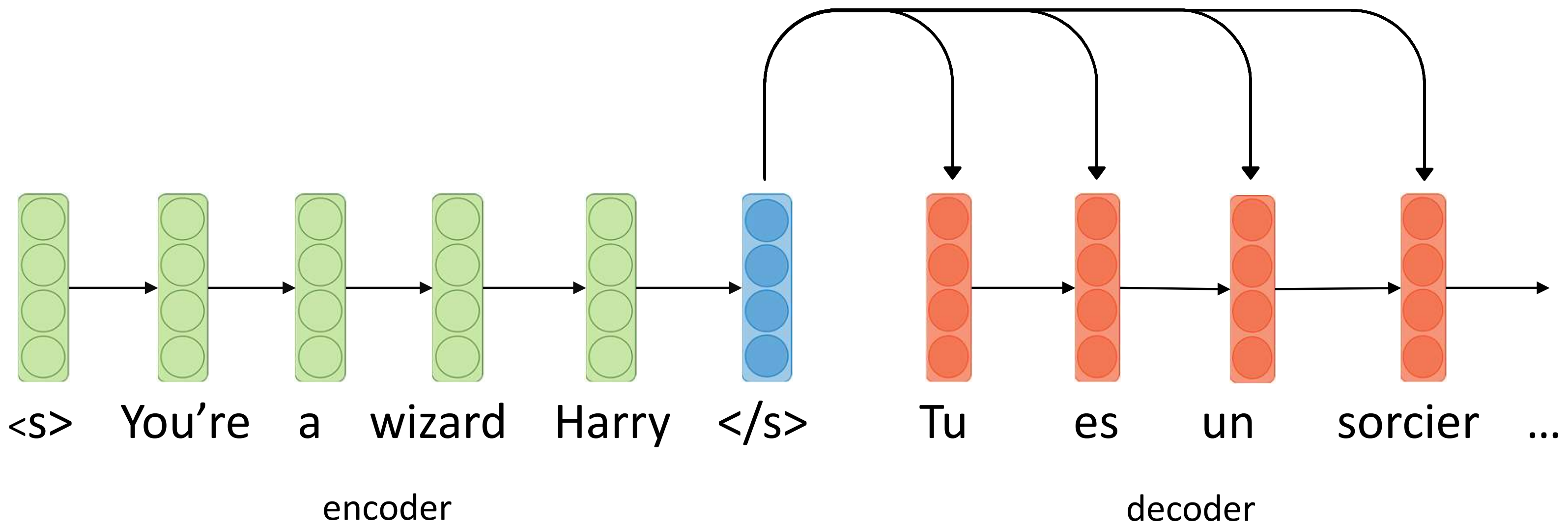

Y
Z
...
...
A
B
C
D
X
Y
Z
...


es
un
sorcier
Harry
You're
a
wizard
Harry
Tu
es
un
sorcier


<s>
You’re
a
wizard
Harry
</s>
Tu
es
un
sorcier
…
encoder
decoder


context vector
<s>
You’re
a
wizard
Harry
</s>
Tu
es
un
sorcier
…
encoder
decoder

Cho et al., (2014): connect to every state in decoder



Attention

- Soft version of alignment
- Represents how important each word in input is to predicting a word in output
- We’ll talk about how much the network “attends” to each word.
- First used in MT, improves BLEU score by 10 pts

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. (2014) **Neural machine translation by jointly learning to align and translate**.

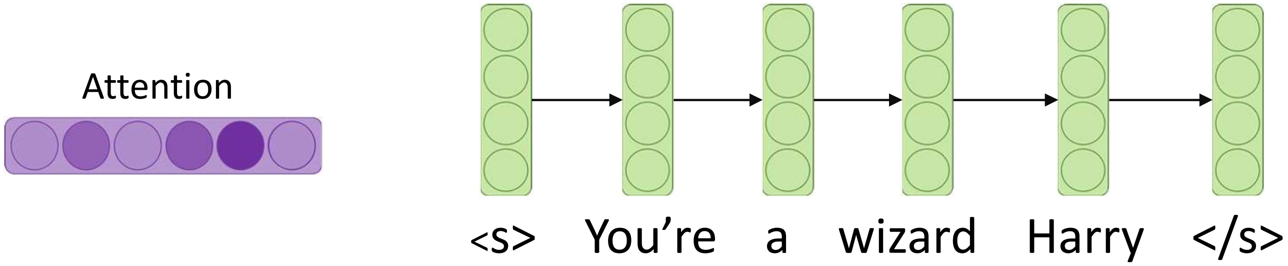
Activity

Question

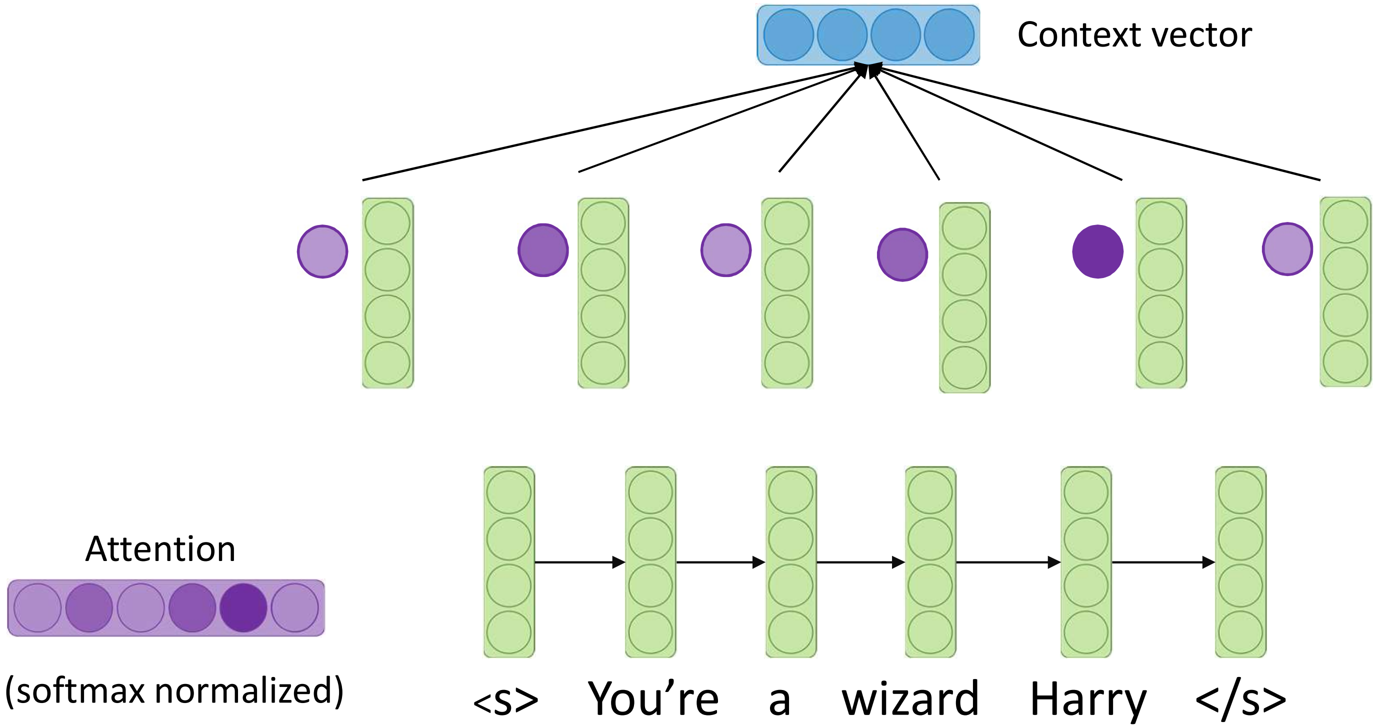
Answer

What	is	the	preferred	weapon	of	the	Jedi	?	A	light	saber
PRON	AUX	DET	ADJ	NOUN	ADP	DET	PROPN		1	2	3

Attention



Attention



Context vector

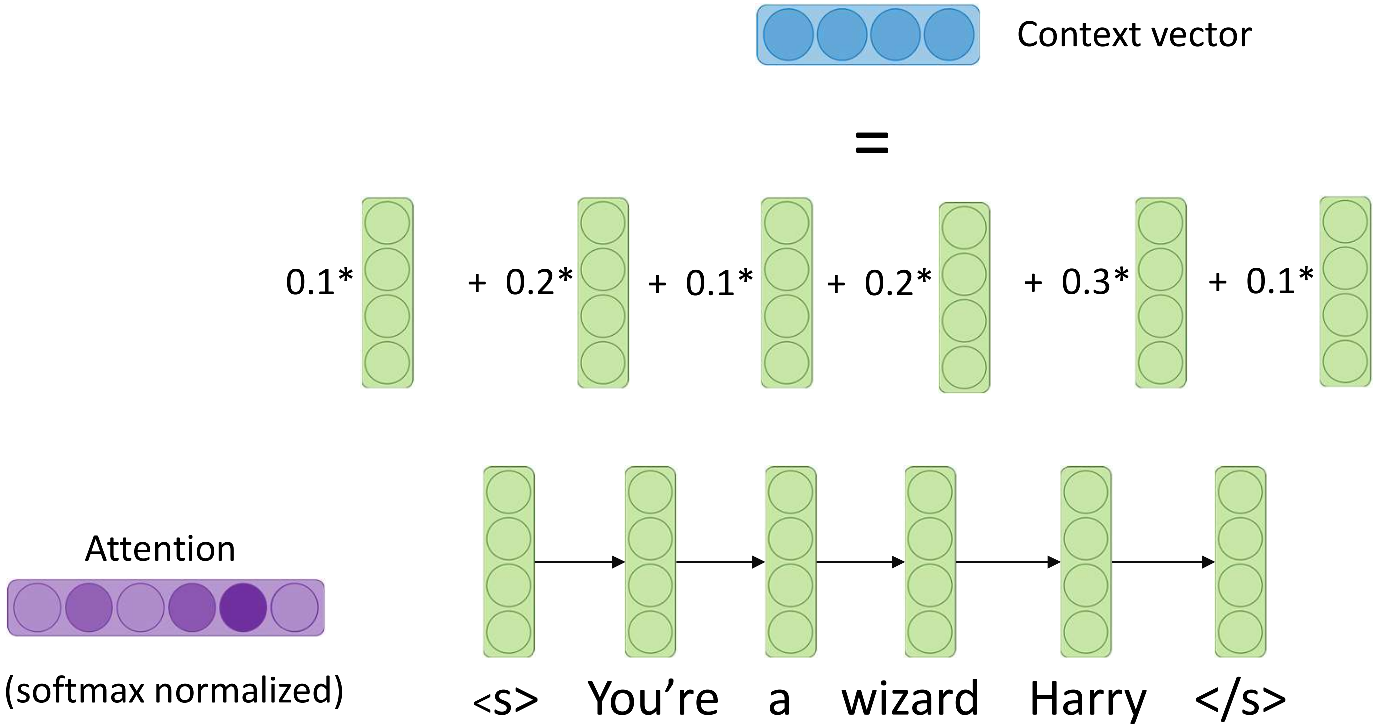
Attention

(softmax normalized)

<s> You’re a wizard Harry </s>

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

Attention



Context vector

=

0.1* + 0.2* + 0.1* + 0.2* + 0.3* + 0.1*

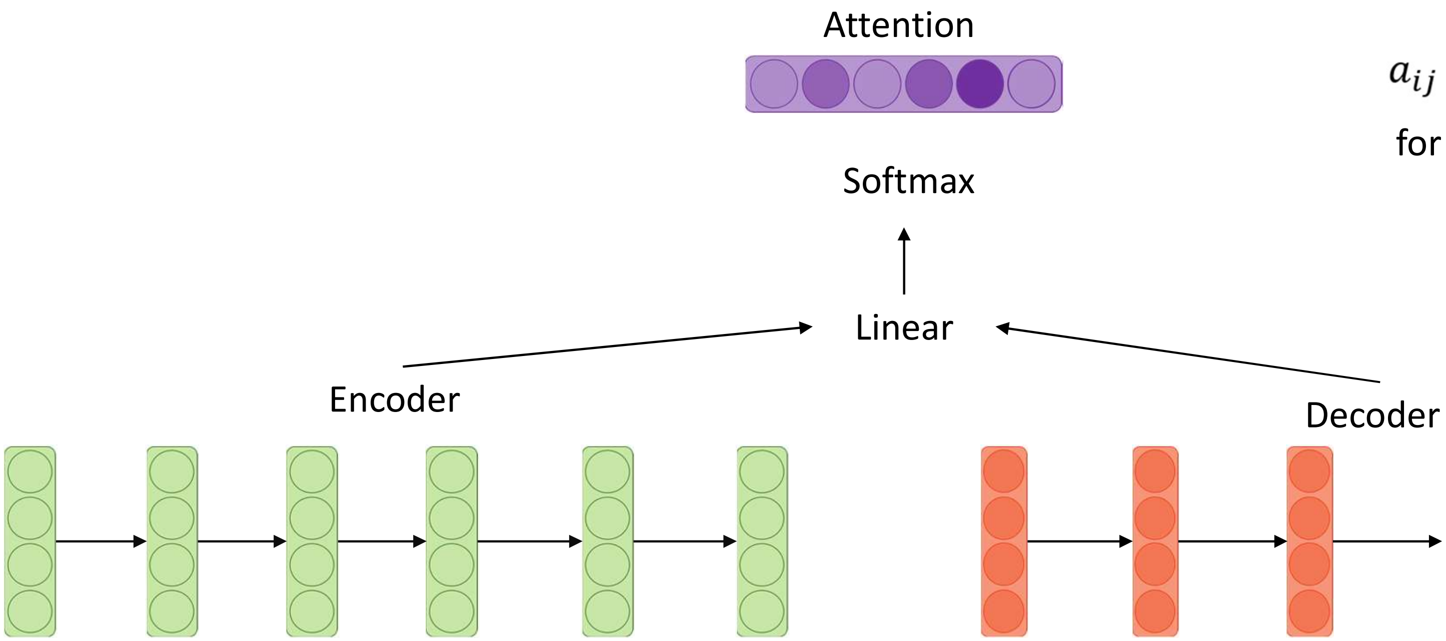
Attention

(softmax normalized)

<s> You’re a wizard Harry </s>

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

Attention

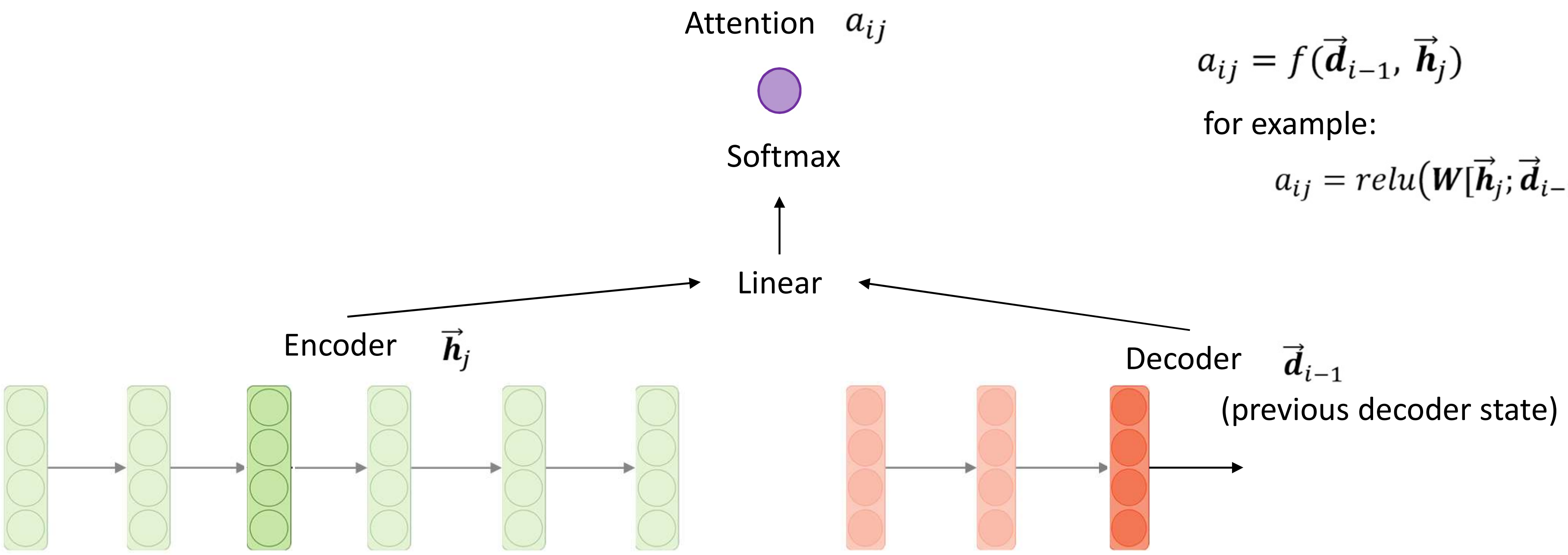


$$a_{ij} = f(\vec{\boldsymbol{d}}_{i-1}, \vec{\boldsymbol{h}}_j)$$

for example:

$$a_{ij} = relu\big(\boldsymbol{W}[\vec{\boldsymbol{h}}_j; \vec{\boldsymbol{d}}_{i-1}]\big)$$

Attention



Attention a_{ij}

Softmax

Linear

Encoder $\vec{\boldsymbol{h}}_j$

Decoder $\vec{\boldsymbol{d}}_{i-1}$

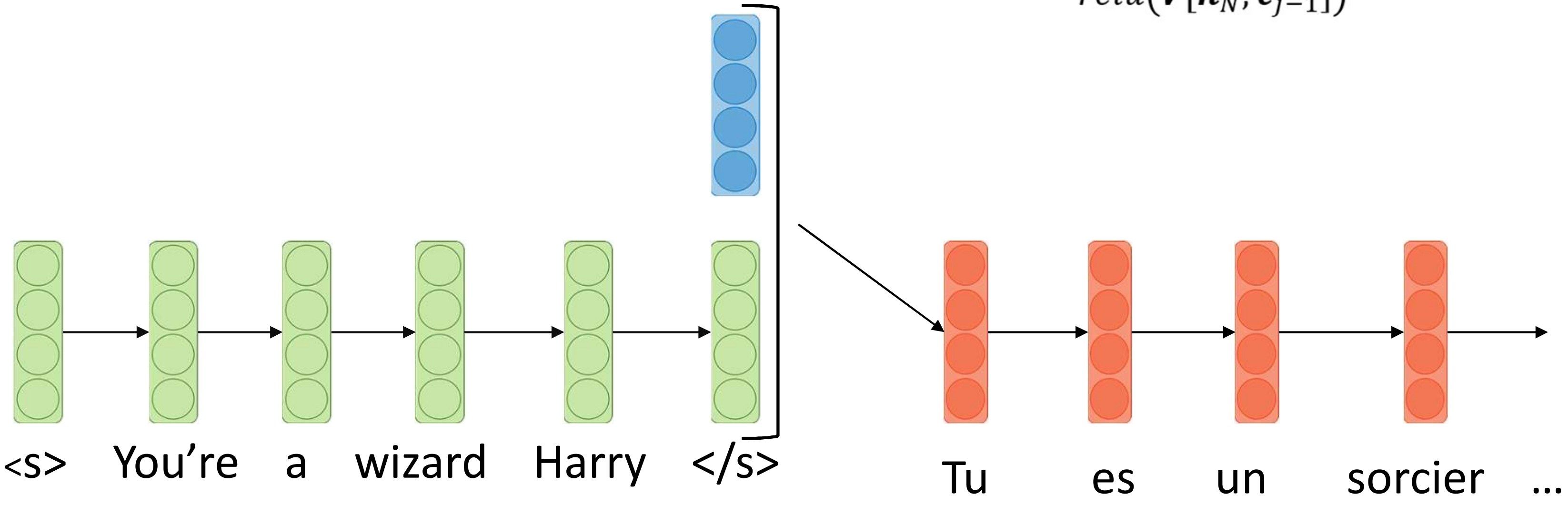
(previous decoder state)

$$a_{ij} = f(\vec{\boldsymbol{d}}_{i-1}, \vec{\boldsymbol{h}}_j)$$

for example:

$$a_{ij} = relu(\boldsymbol{W}[\vec{\boldsymbol{h}}_j; \vec{\boldsymbol{d}}_{i-1}])$$

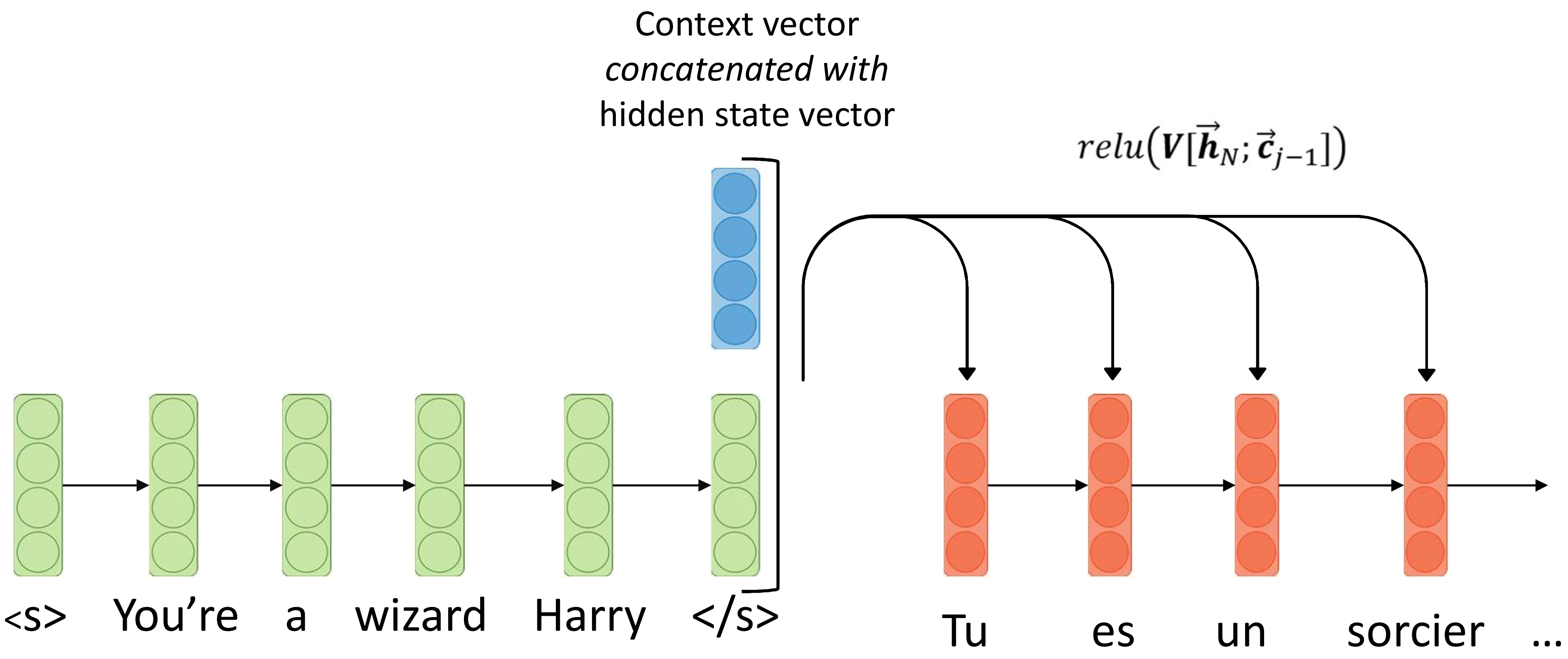
Attention



Context vector *concatenated with* hidden state vector

$relu\left(\boldsymbol{V}[\vec{\boldsymbol{h}}_N; \vec{\boldsymbol{c}}_{j-1}]\right)$

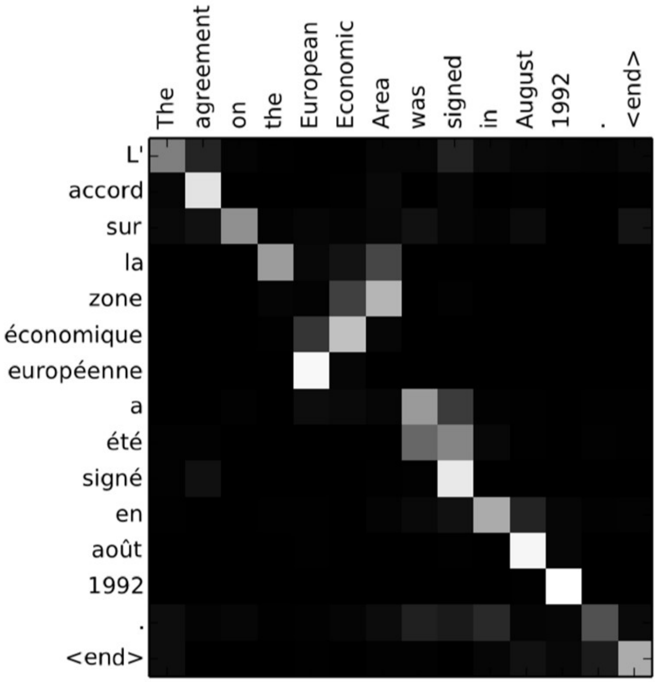
Attention


Context vector
concatenated with
hidden state vector
$relu(\boldsymbol{V}[\vec{\boldsymbol{h}}_N; \vec{\boldsymbol{c}}_{j-1}])$
<s>
You're
a
wizard
Harry
</s>
Tu
es
un
sorcier
...

Attention

Every attention value depends on one word in the source and one in the target.

Attention matrix tells us how “important” a source word is for each target word (much like alignment).


The
agreement
on
the
European
Economic
Area
was
signed
in
August
<end>
L'
accord
sur
la
zone
économique
européenne
a
été
signé
en
août
<end>