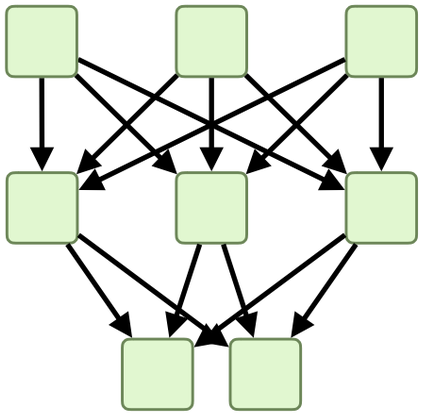
Sequence-to-sequence and neural attention

Recap

Feedforward neural networks

Multi-layer perceptron, convolutional NN

One-to-one input-output mapping

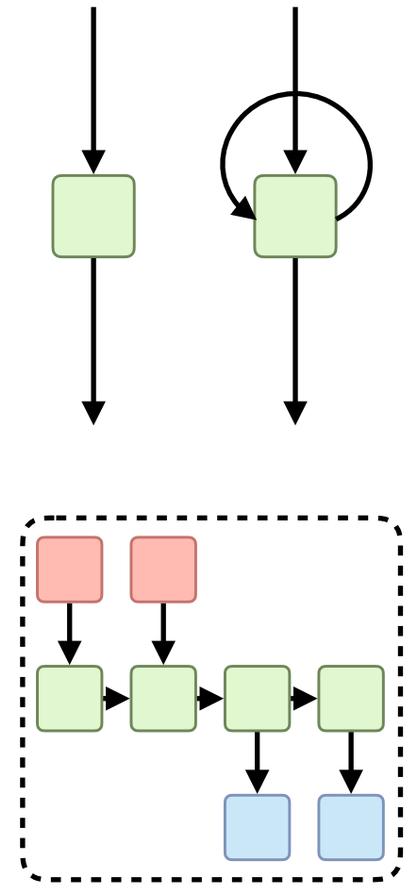


Recurrent neural networks

Cycles → state / memory of past inputs

Many-to-many input-output mapping

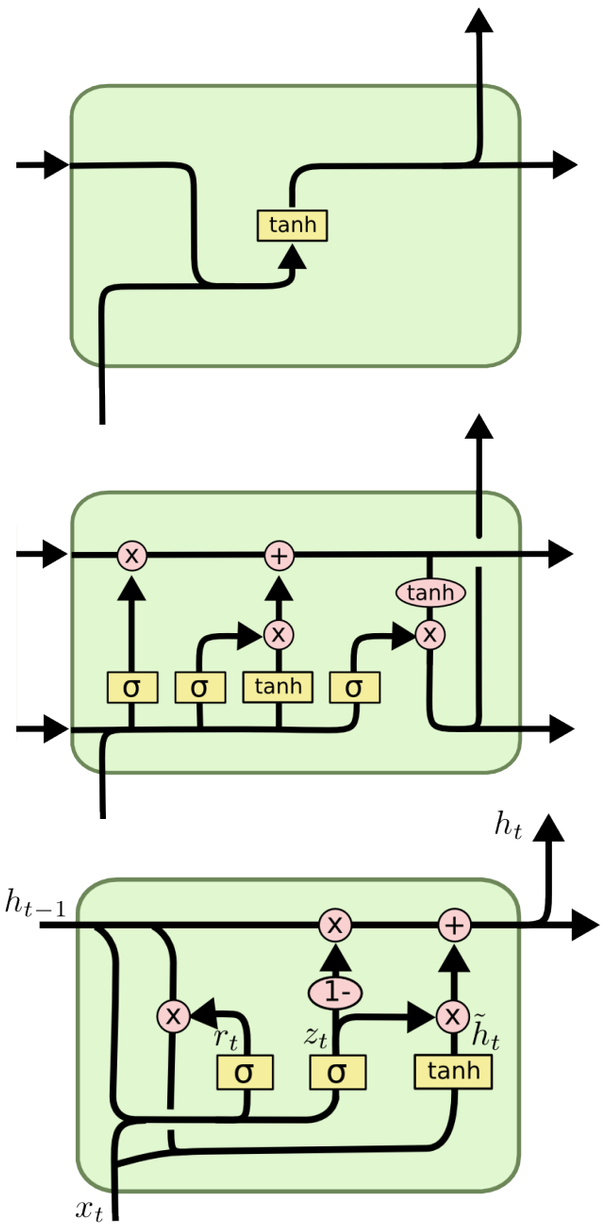
Variable-length inputs and outputs



Recap

Vanilla RNNs have poor short-term memory

Hard to train: exploding / vanishing gradients



LSTM/GRU RNN cell design alleviates short-term memory issues

Changes to cell state controlled by gates

Illustrations from Olah (2015) *Understanding LSTM Networks*

Bidirectional and deep RNNs

Bidirectional RNNs

Standard “forward” RNNs can remember **previous** inputs

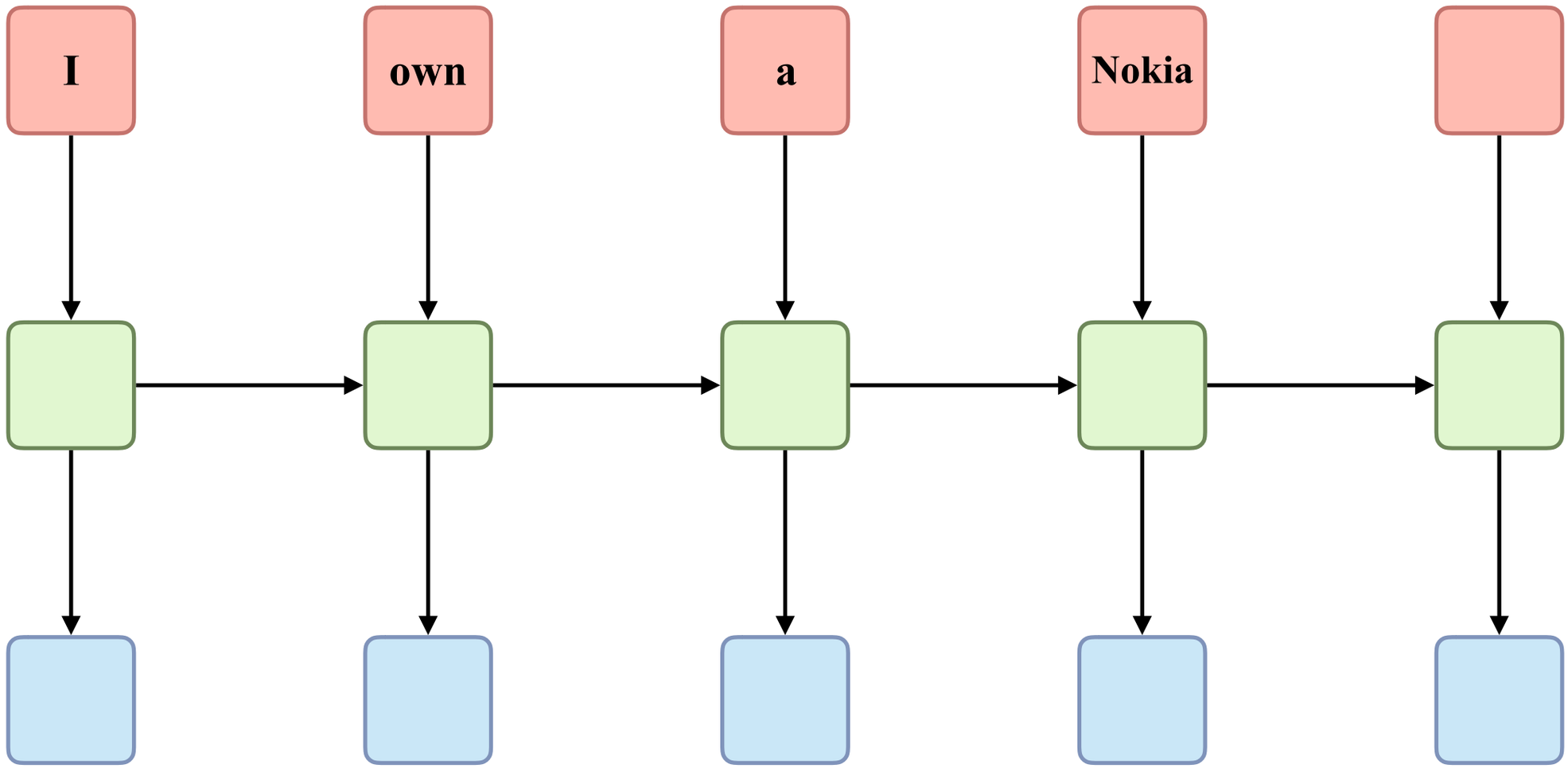
Key information may be found in “**future**” inputs (e.g. right text context) (obviously, these are inaccessible in *actual* time series)

Consider the task of distinguishing mentions of locations (`LOC`) from products (`PRO`) in the following examples:

- “I visited Nokia” (`LOC`) vs . “I own a Nokia” (`PRO`)
- “Nokia is a nice place” (`LOC`) vs. “Nokia made good phones” (`PRO`)

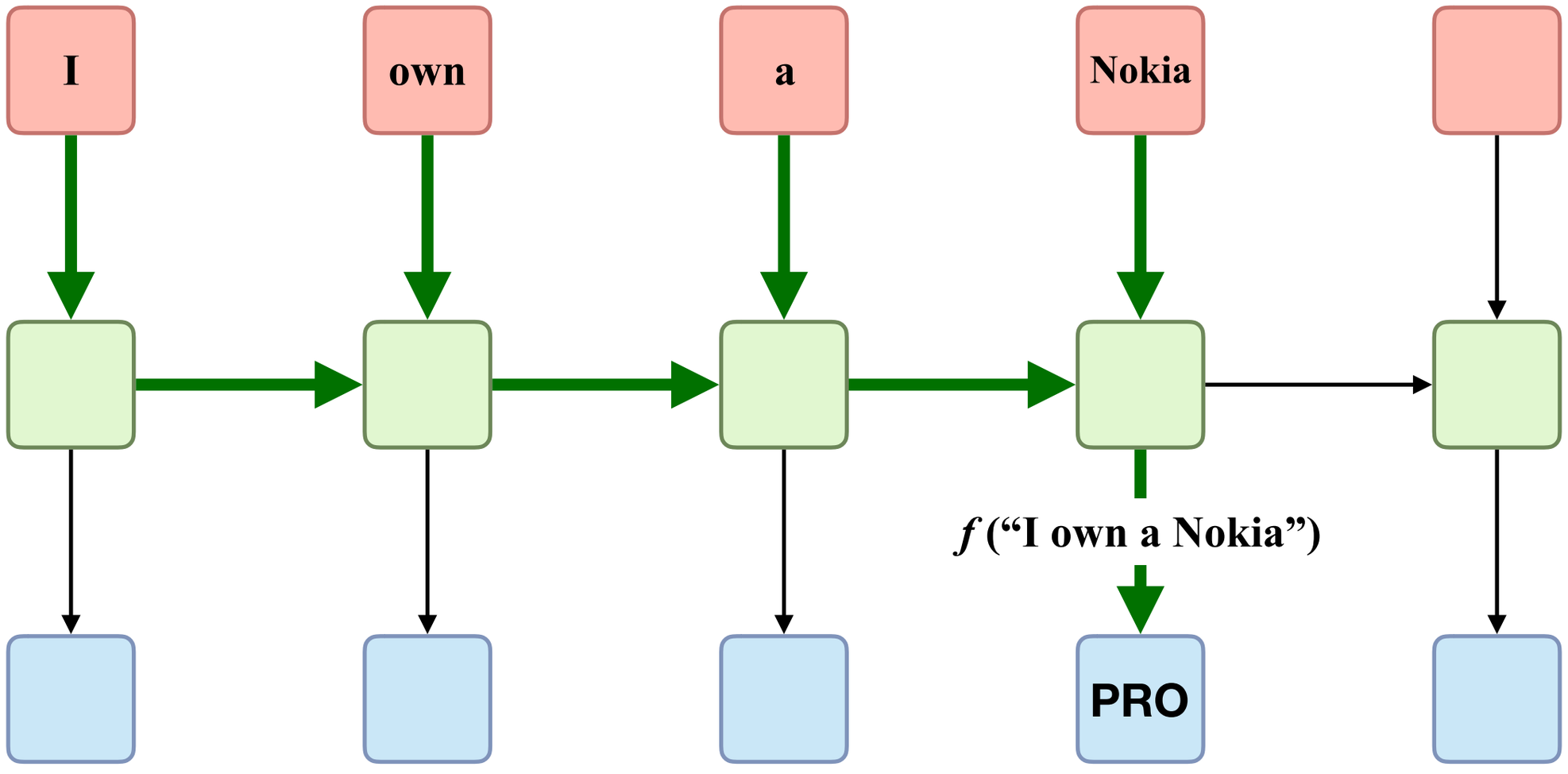
Bidirectional RNNs

Standard forward RNN has access to previous inputs



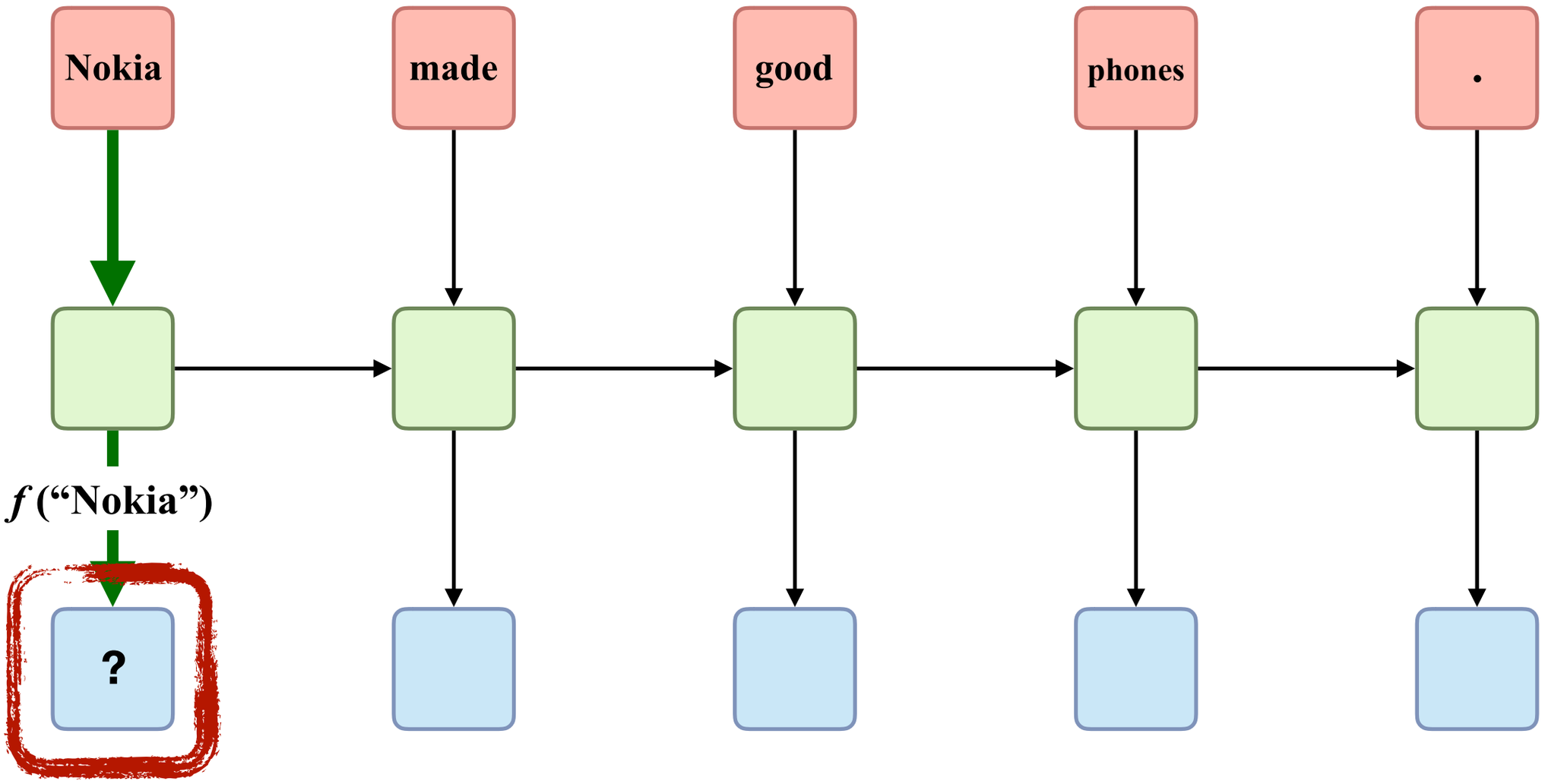
Bidirectional RNNs

Standard forward RNN has access to previous inputs



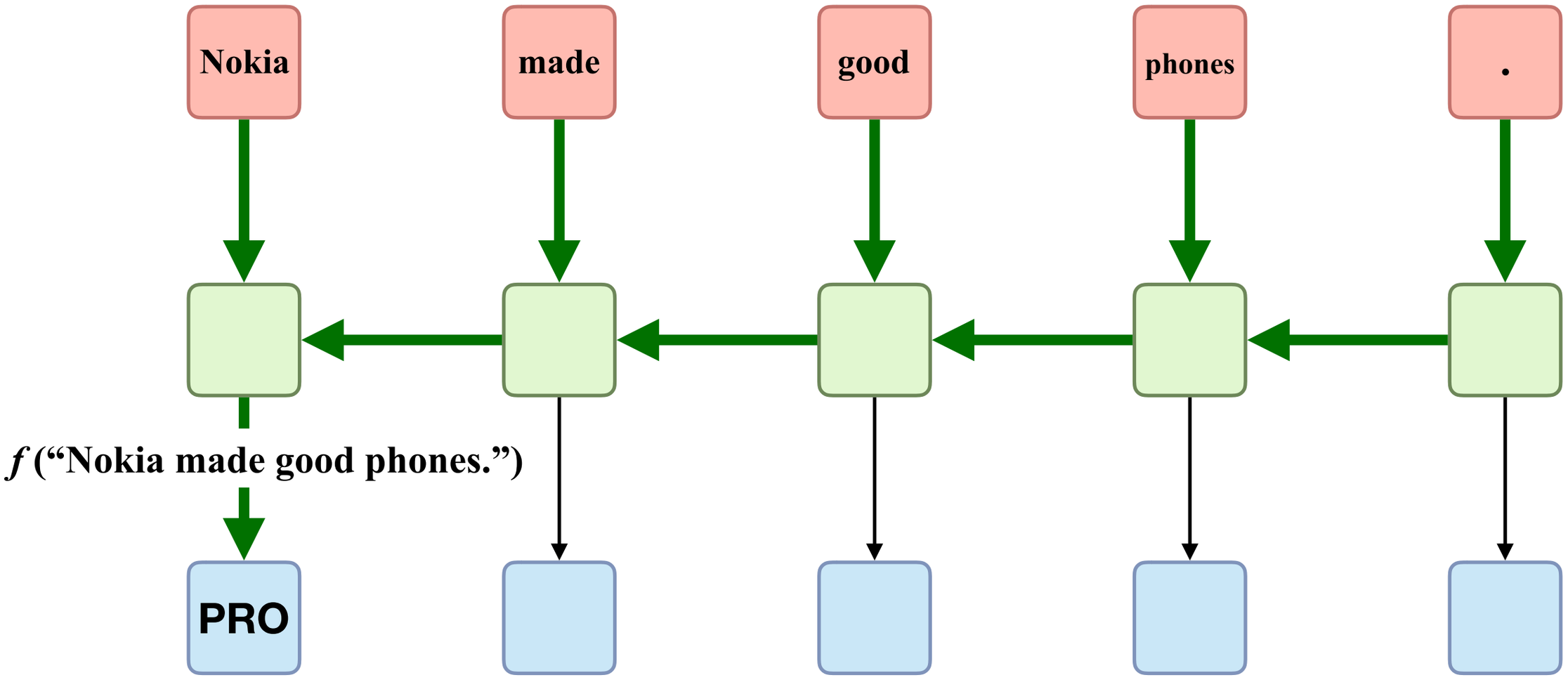
Bidirectional RNNs

… forward RNN cannot access “future” inputs



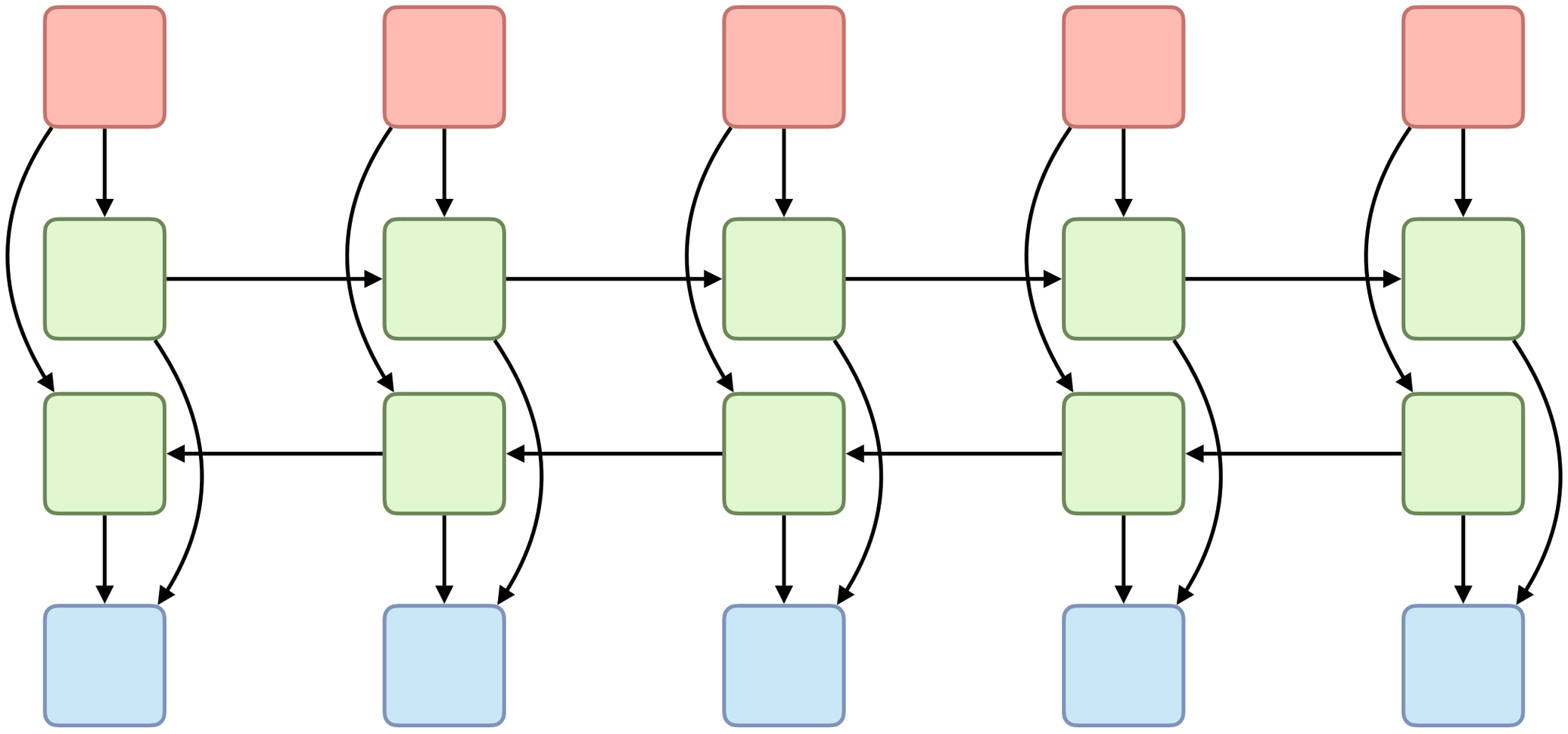
Bidirectional RNNs

Backward RNN has access to right text context



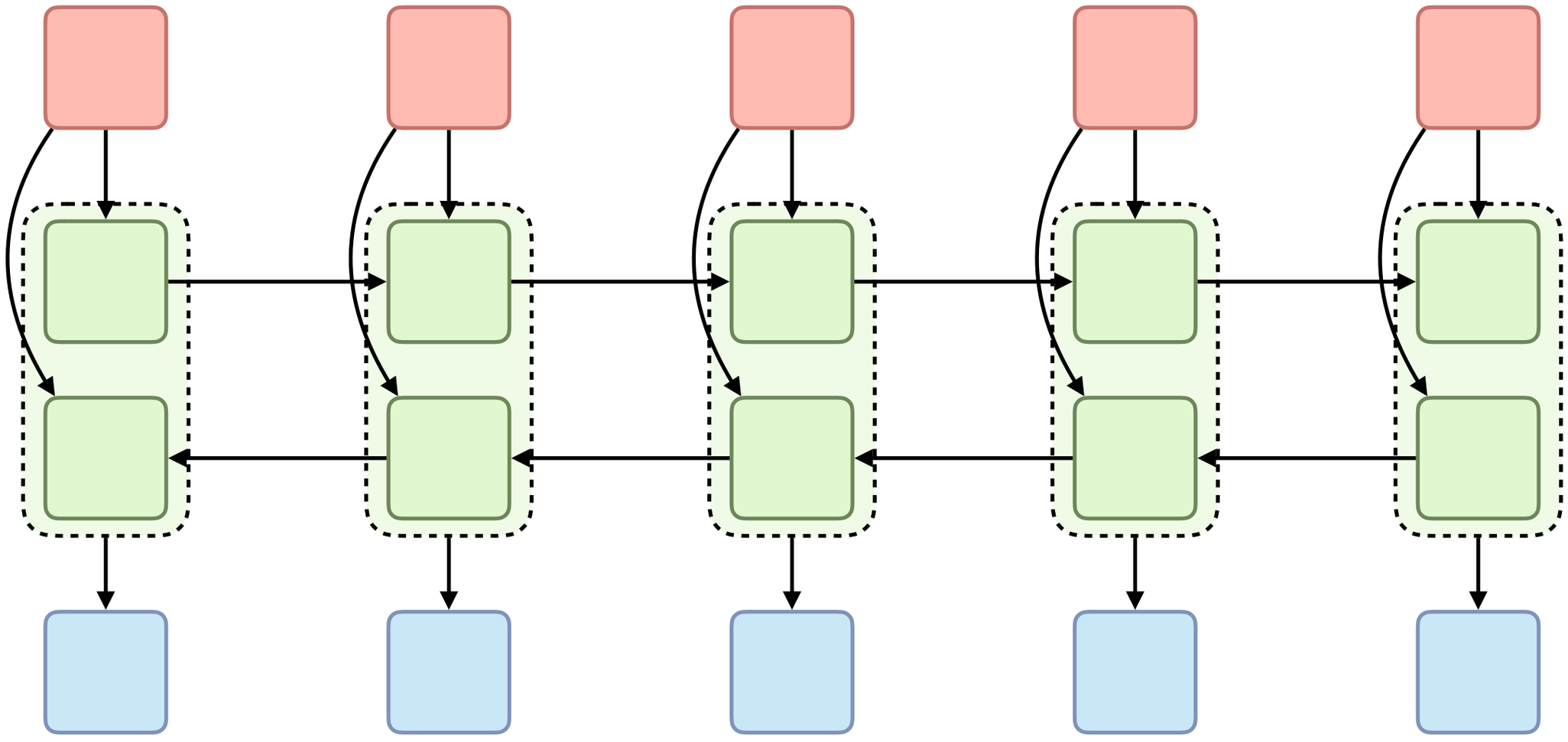
Bidirectional RNNs

Bidirectional RNNs combine forward and backward RNNs



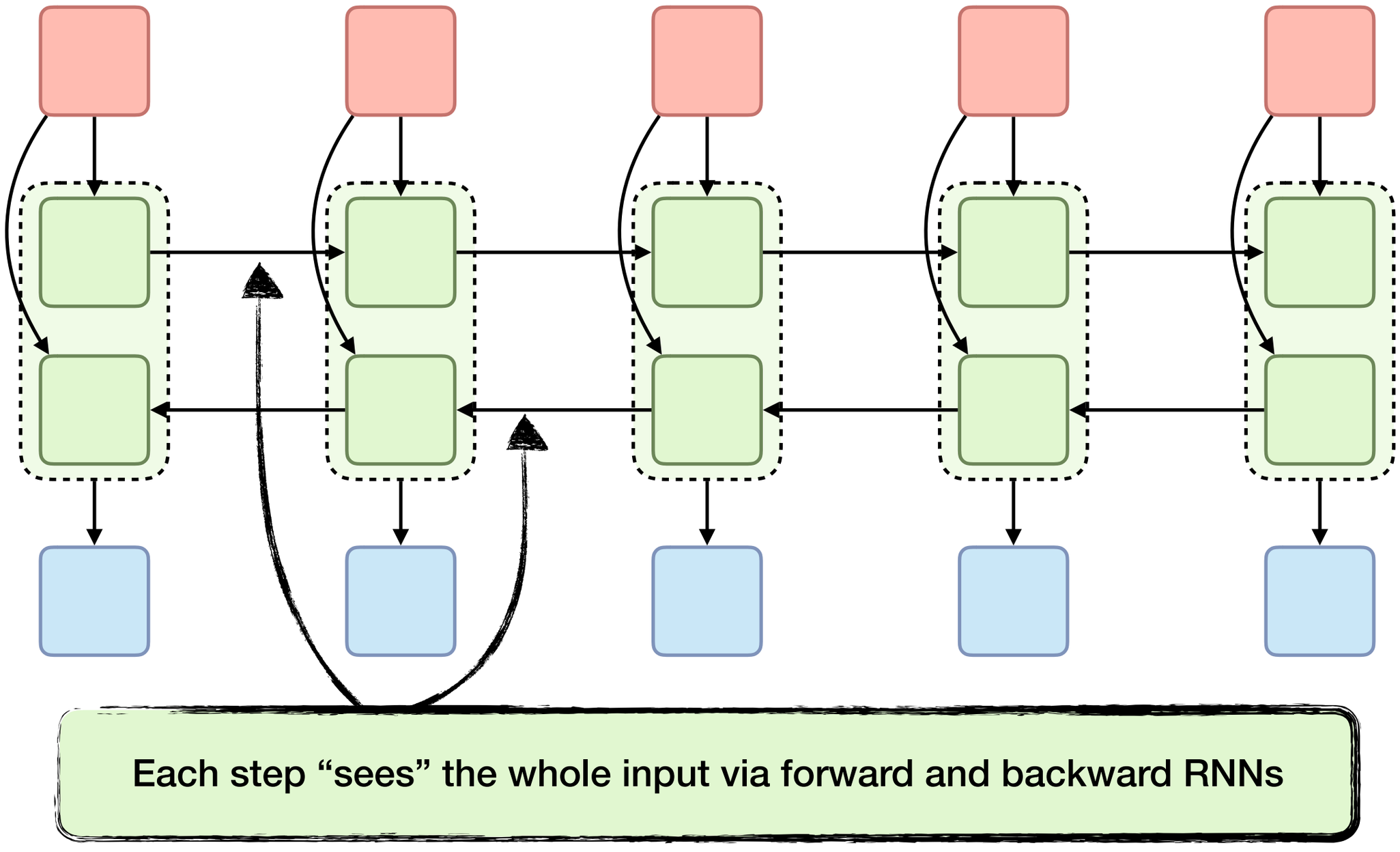
Bidirectional RNNs

Fw & bw RNNs form a unit, output can combine information from both



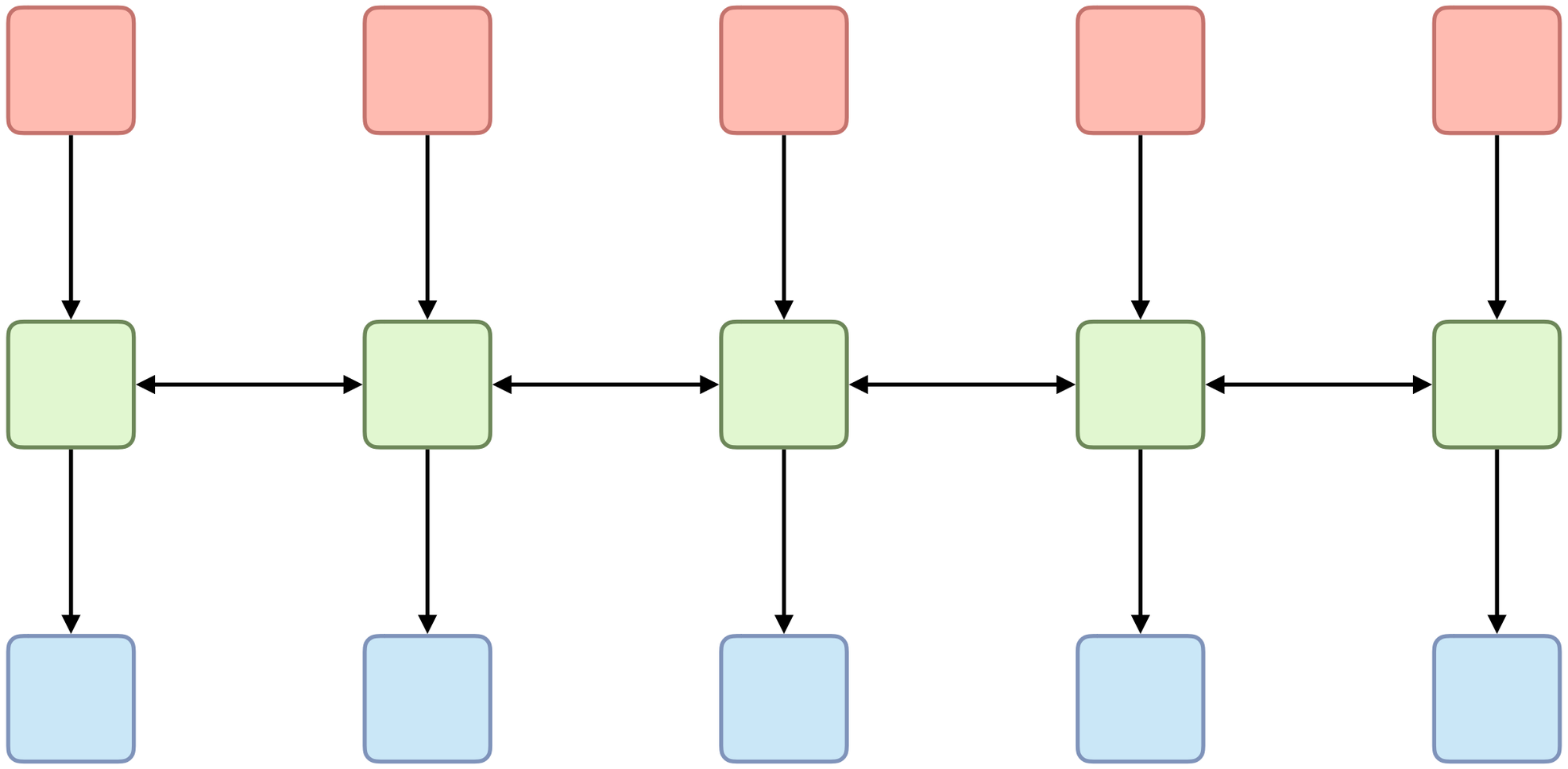
Bidirectional RNNs

Fw & bw RNNs form a unit, output can combine information from both



Bidirectional RNNs

The “interface” of bidirectional RNNs is identical to standard RNNs



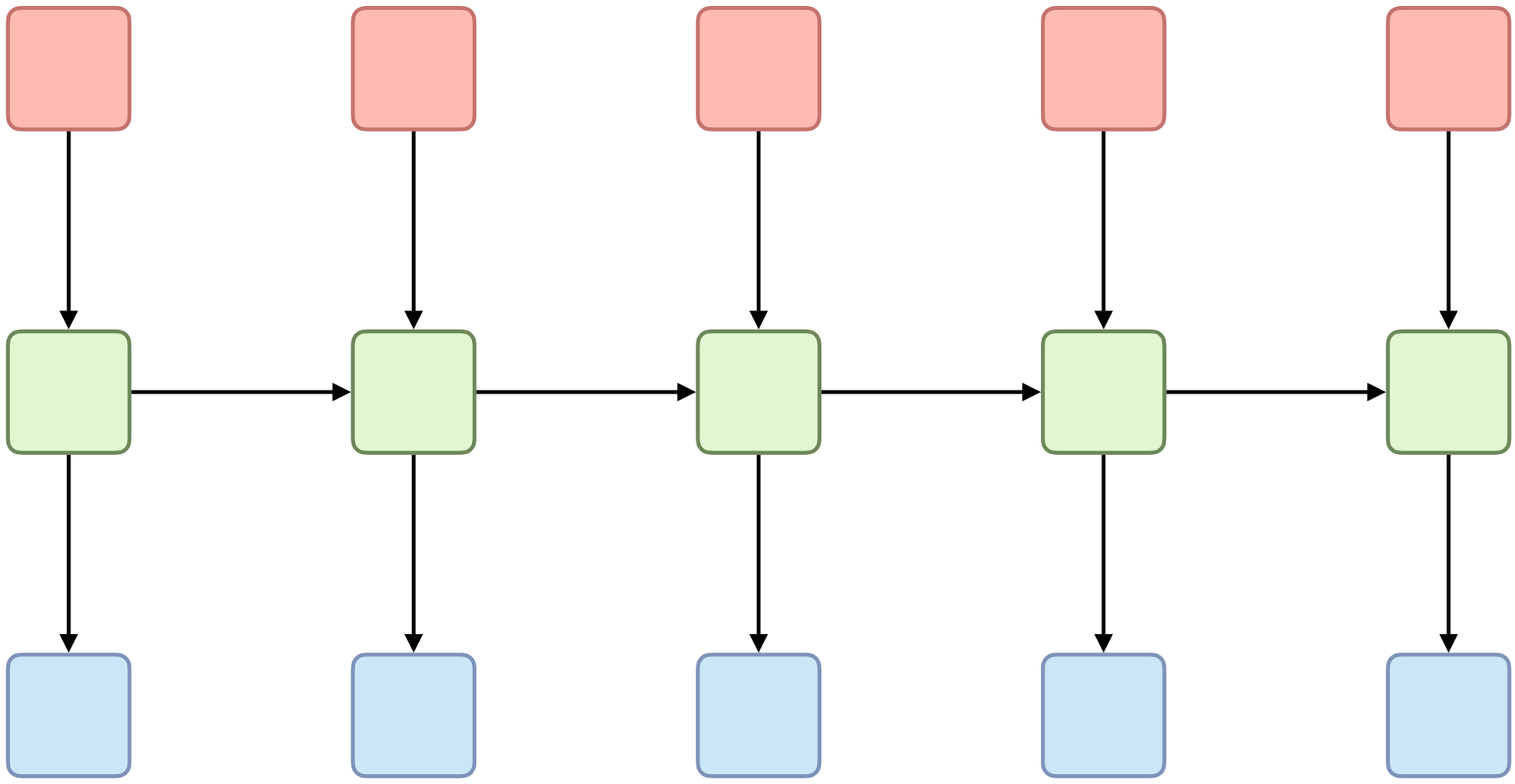
In code (keras)

```
…
emb = Embedding(…)(input_)
rnn = LSTM(…)(emb)
out = Dense(…)(rnn)
```

```
…
emb = Embedding(…)(input_)
rnn = Bidirectional(LSTM(…))(emb)
out = Dense(…)(rnn)
```

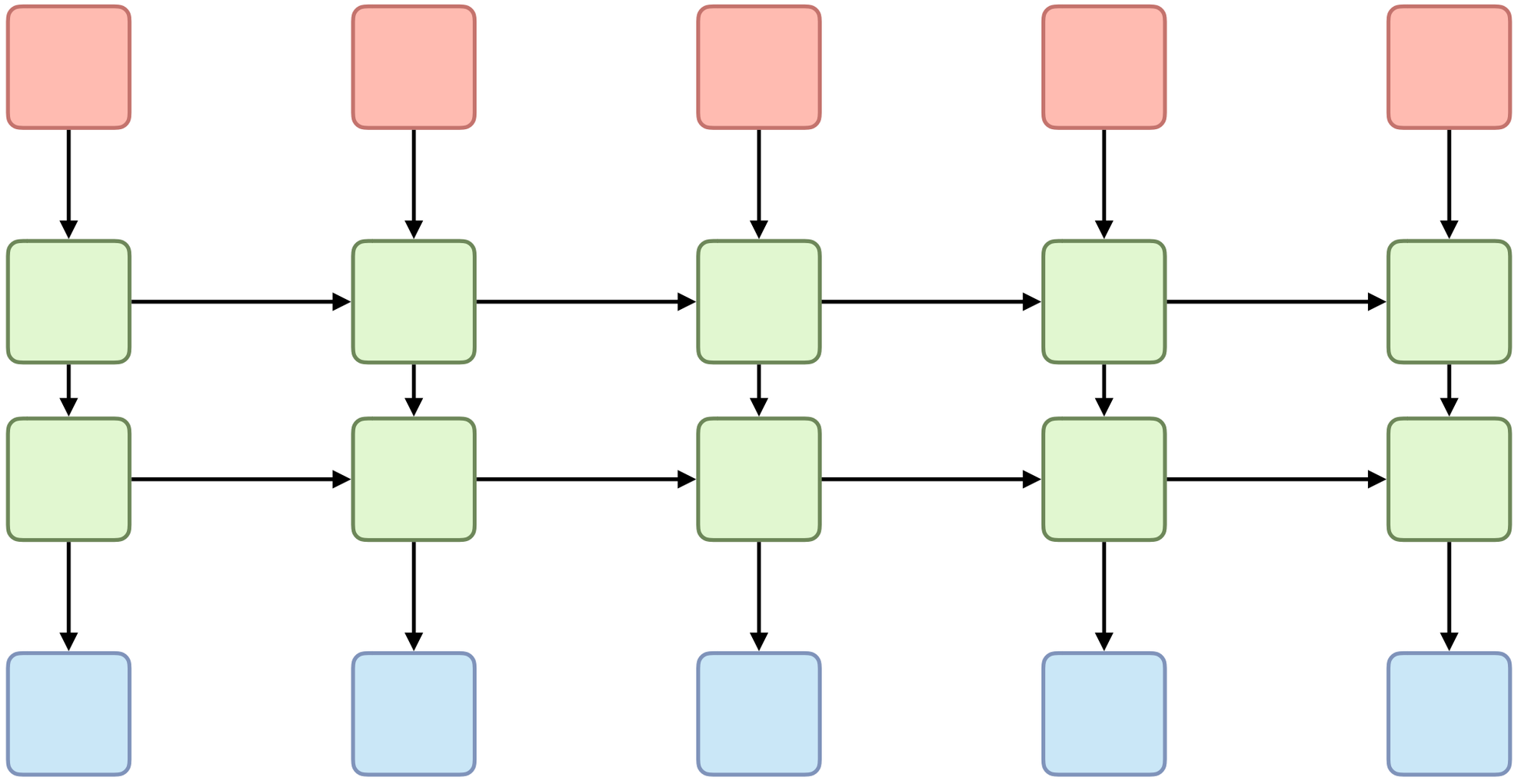
Deep RNNs

Standard single-layer forward RNN



Deep RNNs

Stacked architecture: output of one layer becomes input of the next

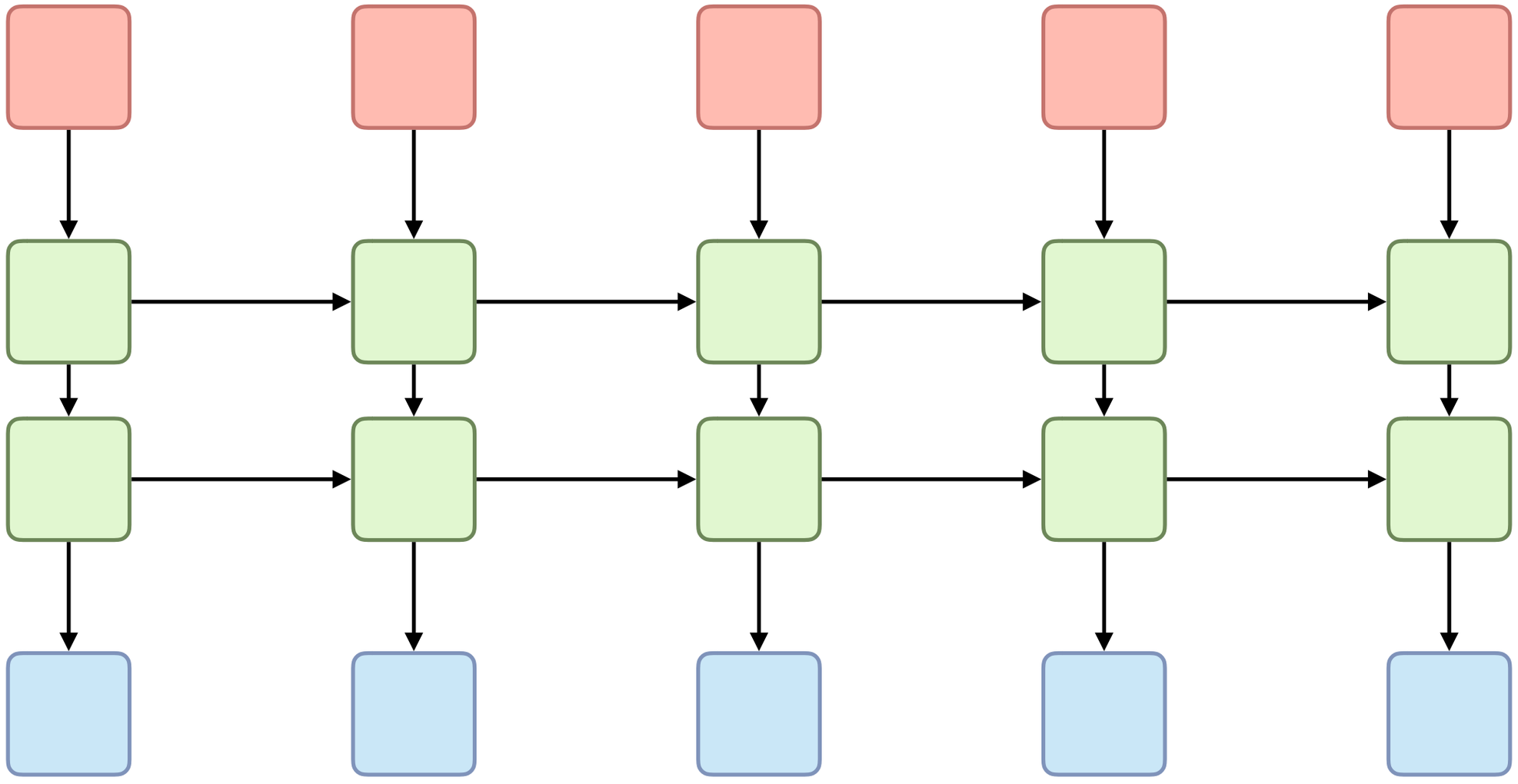


Graves *et al*. (2013) *Speech Recognition with Deep Recurrent Neural Networks*

Pascanu *et al*. (2014) *How to Construct Deep Recurrent Neural Networks*

Deep RNNs

Stacked architecture: output of one layer becomes input of the next



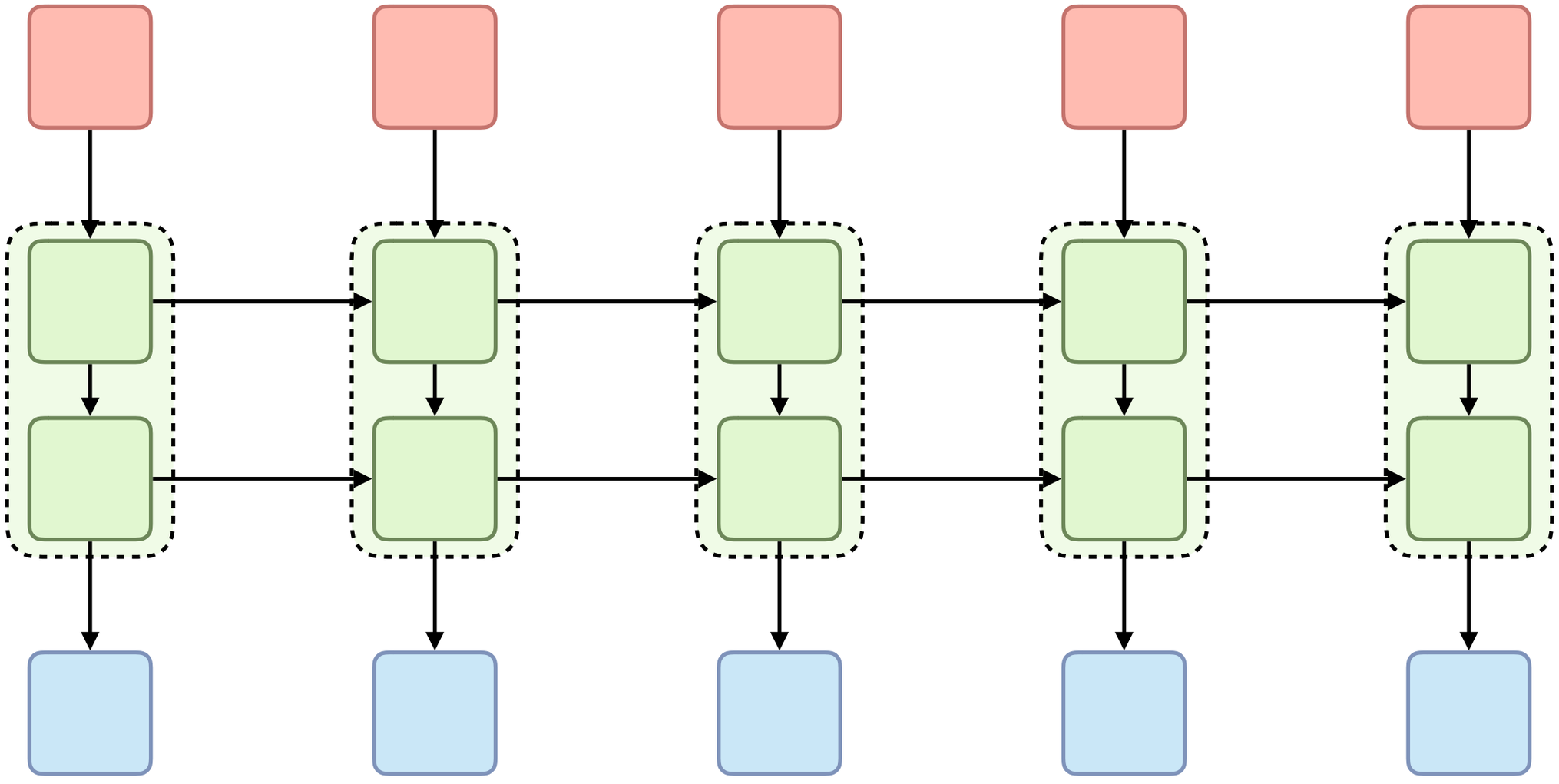
Graves *et al*. (2013) *Speech Recognition with Deep Recurrent Neural Networks*

Pascanu *et al*. (2014) *How to Construct Deep Recurrent Neural Networks*

Not the only option, see here

Deep RNNs

We can view the stack as a unit, the interface stays the same



Graves *et al*. (2013) *Speech Recognition with Deep Recurrent Neural Networks*

Pascanu *et al*. (2014) *How to Construct Deep Recurrent Neural Networks*

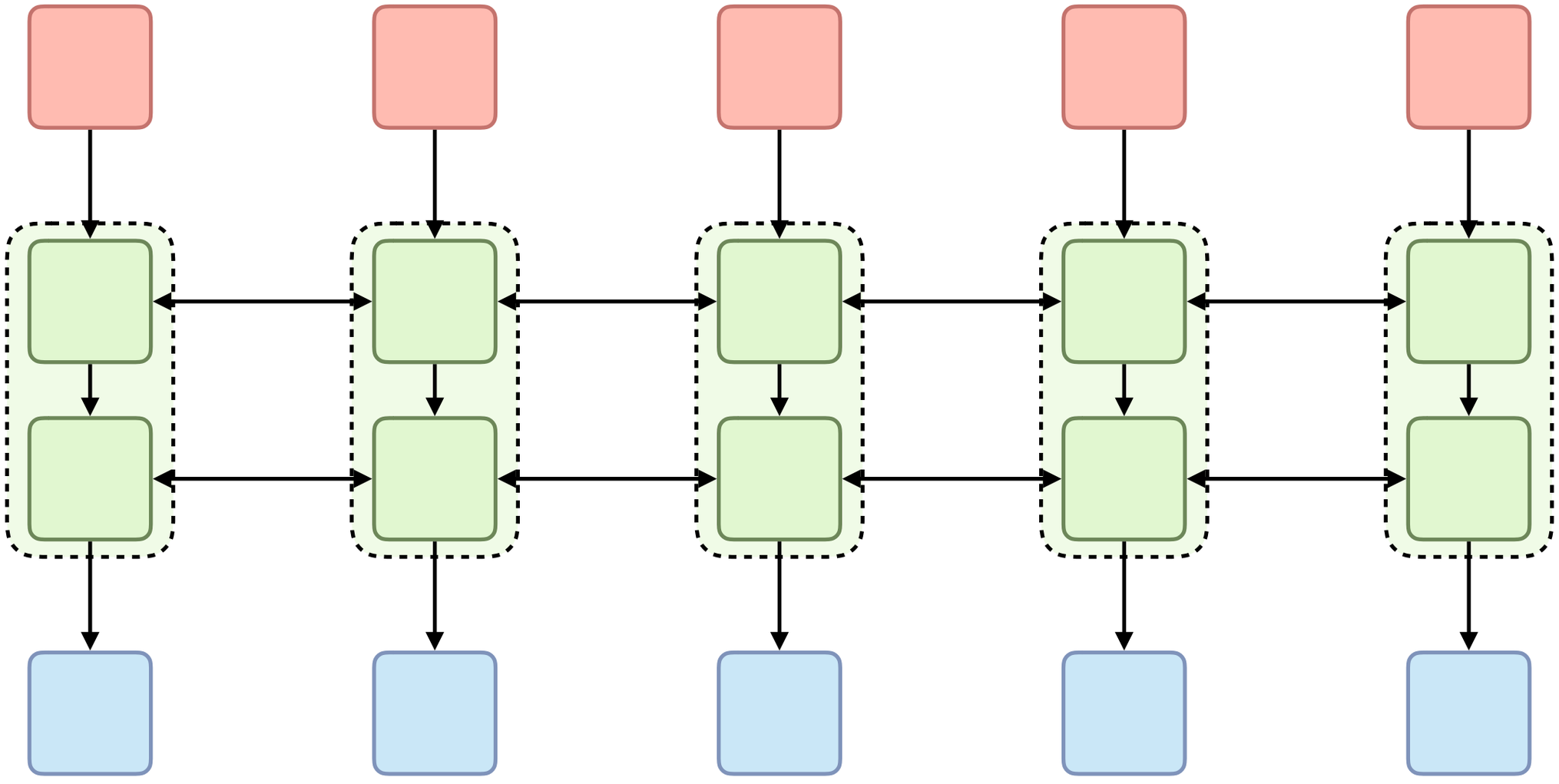
In code (keras)

```
…
emb = Embedding(…)(input_)
rnn = LSTM(…)(emb)
out = Dense(…)(rnn)
```

```
…
emb = Embedding(…)(input_)
sl1 = LSTM(…, return_sequences=True)(emb)
sl2 = LSTM(…)(sl1)
out = Dense(…)(sl2)
```

Deep RNNs

And we can naturally also stack bidirectional RNNs



Graves *et al*. (2013) *Speech Recognition with Deep Recurrent Neural Networks*

Pascanu *et al*. (2014) *How to Construct Deep Recurrent Neural Networks*

In code (keras)

```
from keras.layers import Bidirectional as Bi

…
emb = Embedding(…)(input_)
sl1 = Bi(LSTM(…, return_sequences=True))(emb)
sl2 = Bi(LSTM(…, return_sequences=True))(sl1)
sl3 = Bi(LSTM(…, return_sequences=True))(sl2)
sl4 = Bi(LSTM(…))(sl3)
out = Dense(…)(sl4)
```

Summary

Bidirectional RNNs have access to both previous and future inputs (i.e. in text, left and right context)

Deep RNN models can be created e.g. by stacking, allowing for benefits associated with deeper models (e.g. abstraction)

Deep bidirectional RNNs combine the two straightforwardly

These architectures can be used with **any RNN cell** (vanilla, LSTM, …)

The input/output interface of deep and/or bidirectional RNNs matches that of single-layer forward RNNs

→ In upcoming material, we'll (mostly) treat these interchangeably

Encoder-decoder architectures and sequence-to-sequence

Encoder-decoder architecture

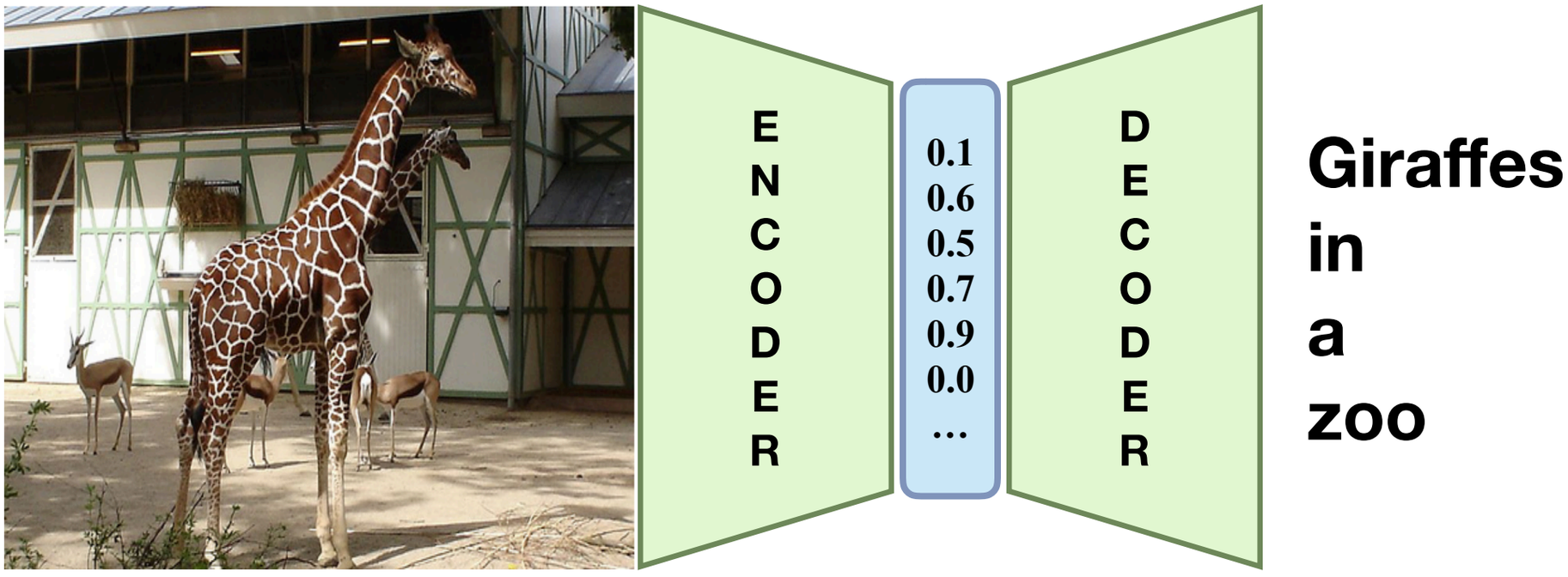
Consider the task of learning a function for image captioning

$f($  $) =$ **“Giraffes in a zoo”**

Example from Wang *et al*. (2016) *Image Captioning with Deep Bidirectional LSTMs*

Encoder-decoder architecture

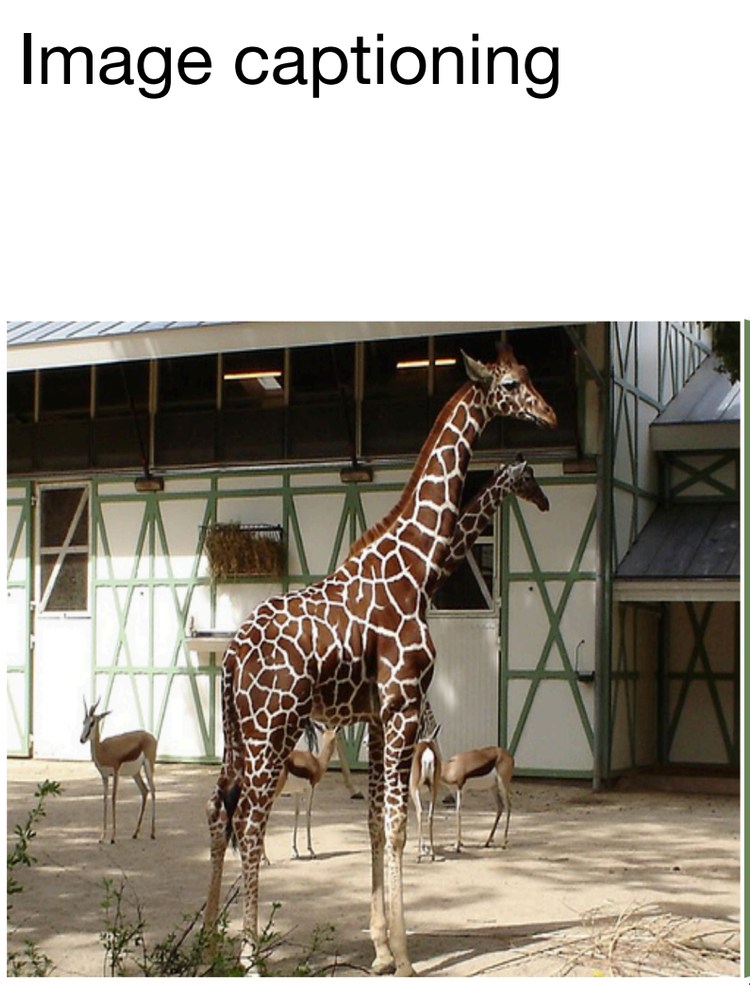
Image captioning

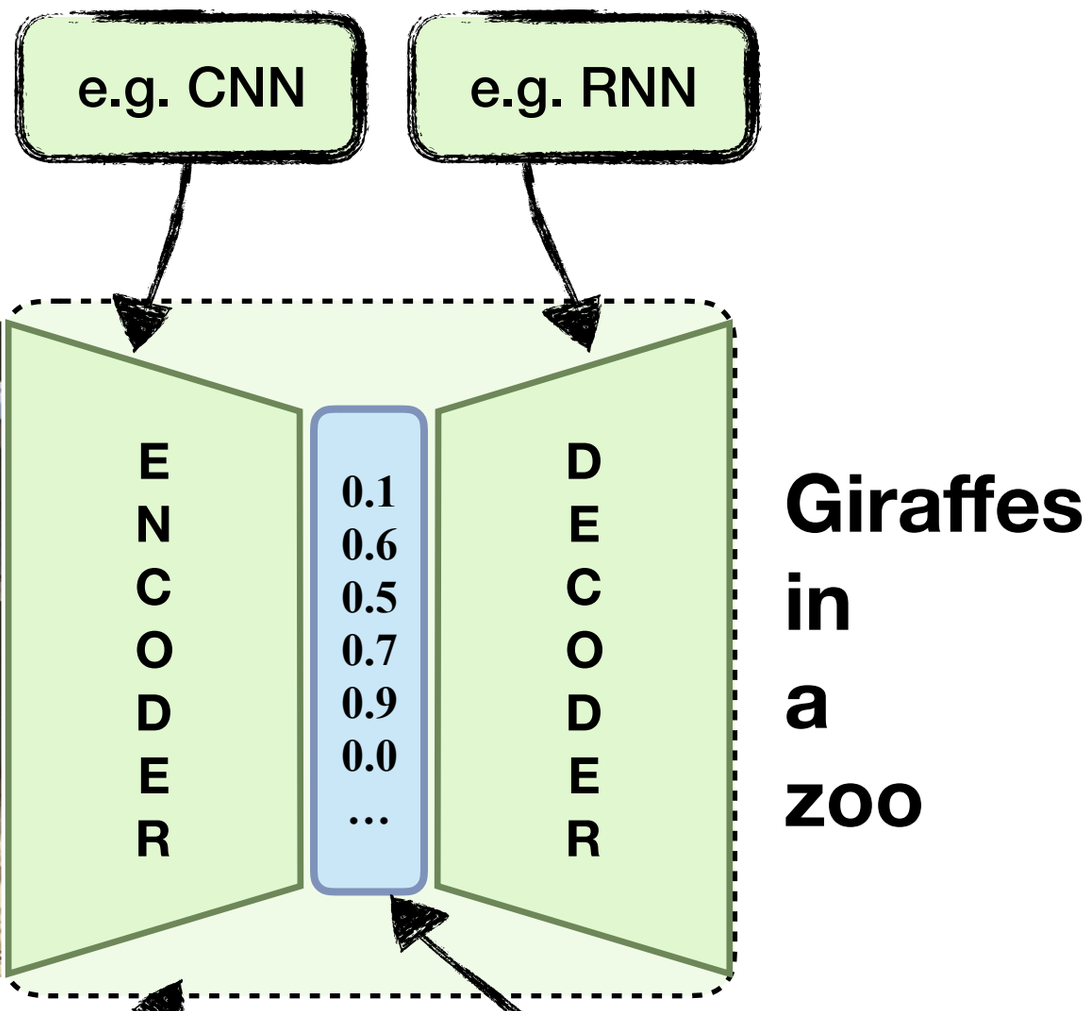


Example from Wang *et al*. (2016) *Image Captioning with Deep Bidirectional LSTMs*

Encoder-decoder architecture

Image captioning

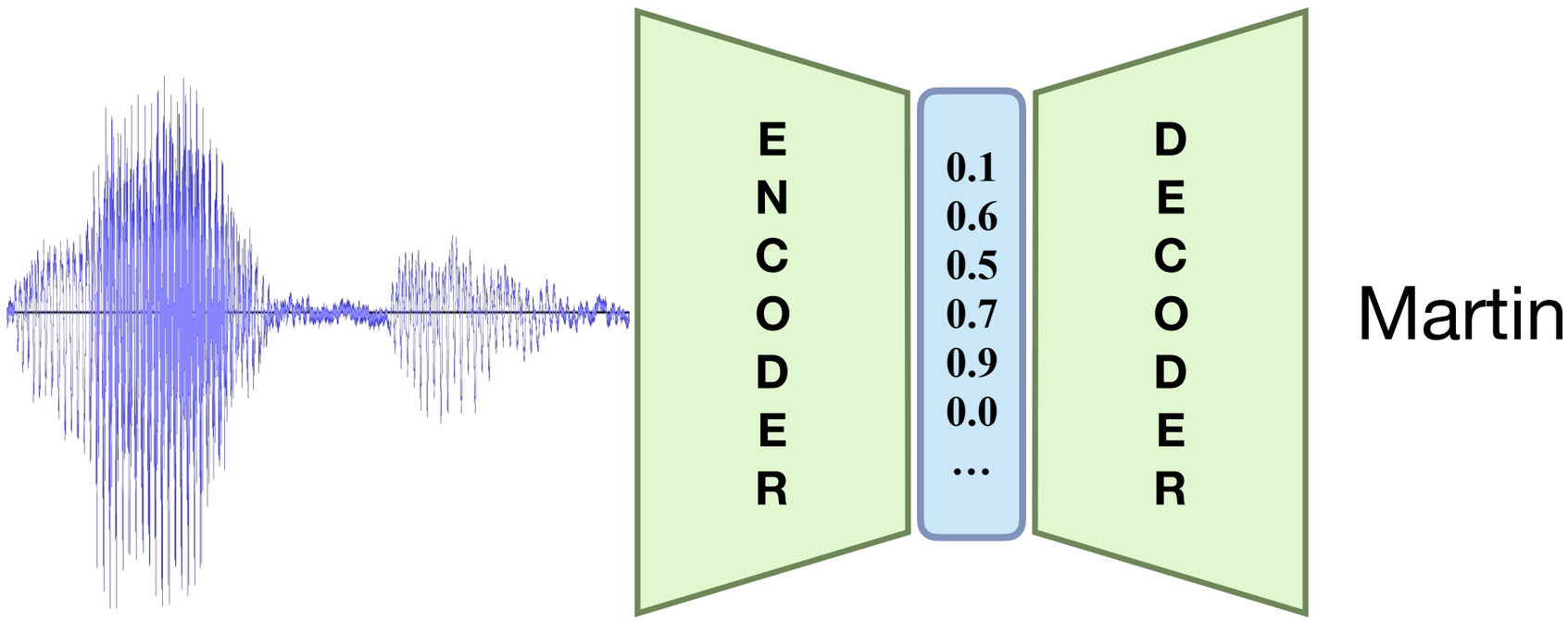




Example from Wang *et al*. (2016) *Image Captioning with Deep Bidirectional LSTMs*

Encoder-decoder architecture

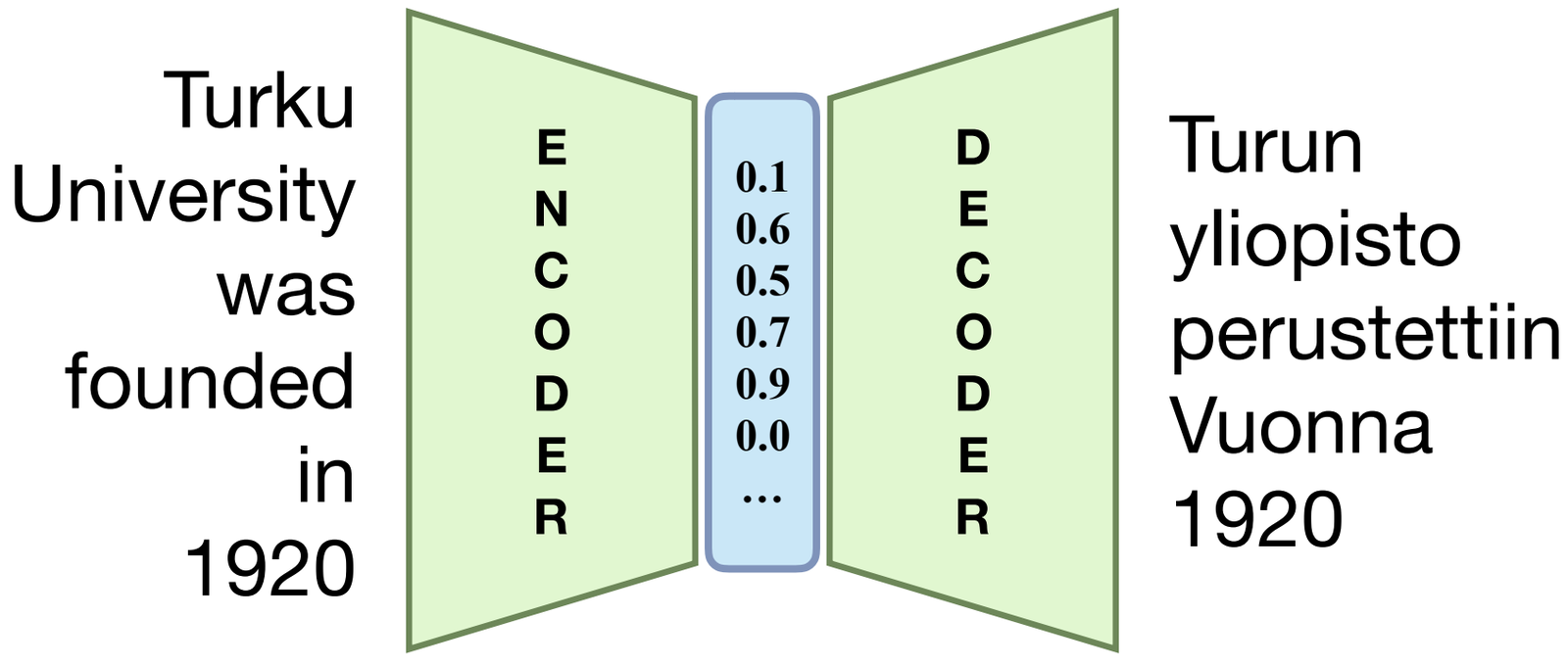
Speech recognition (or, the other way around, text-to-speech)



Martin

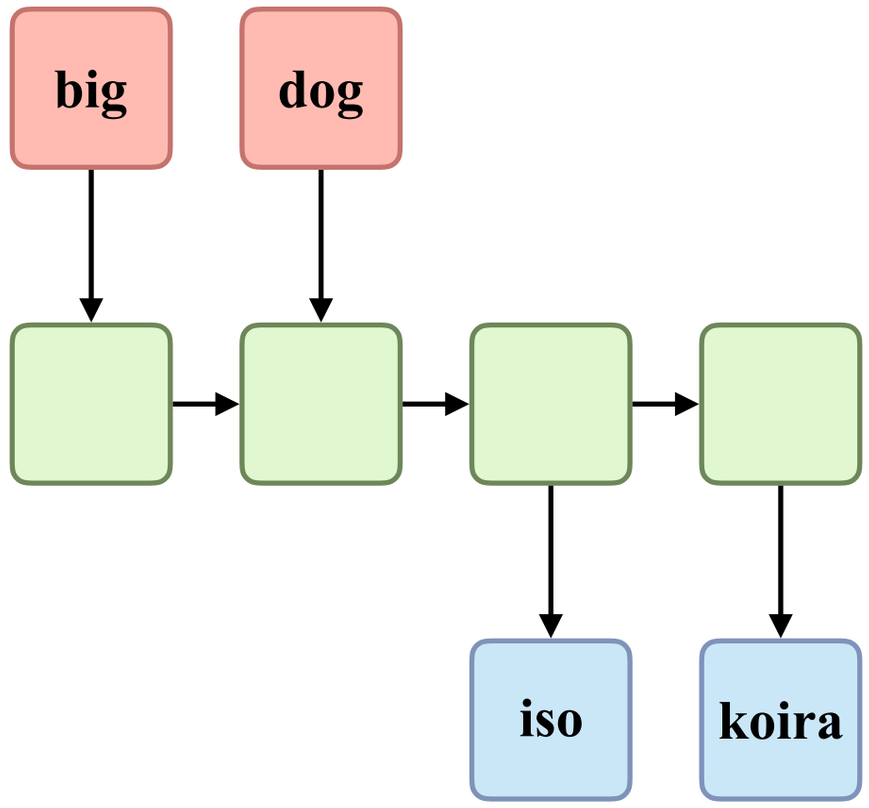
Encoder-decoder architecture

Machine translation, text summarization, dialogue systems, …



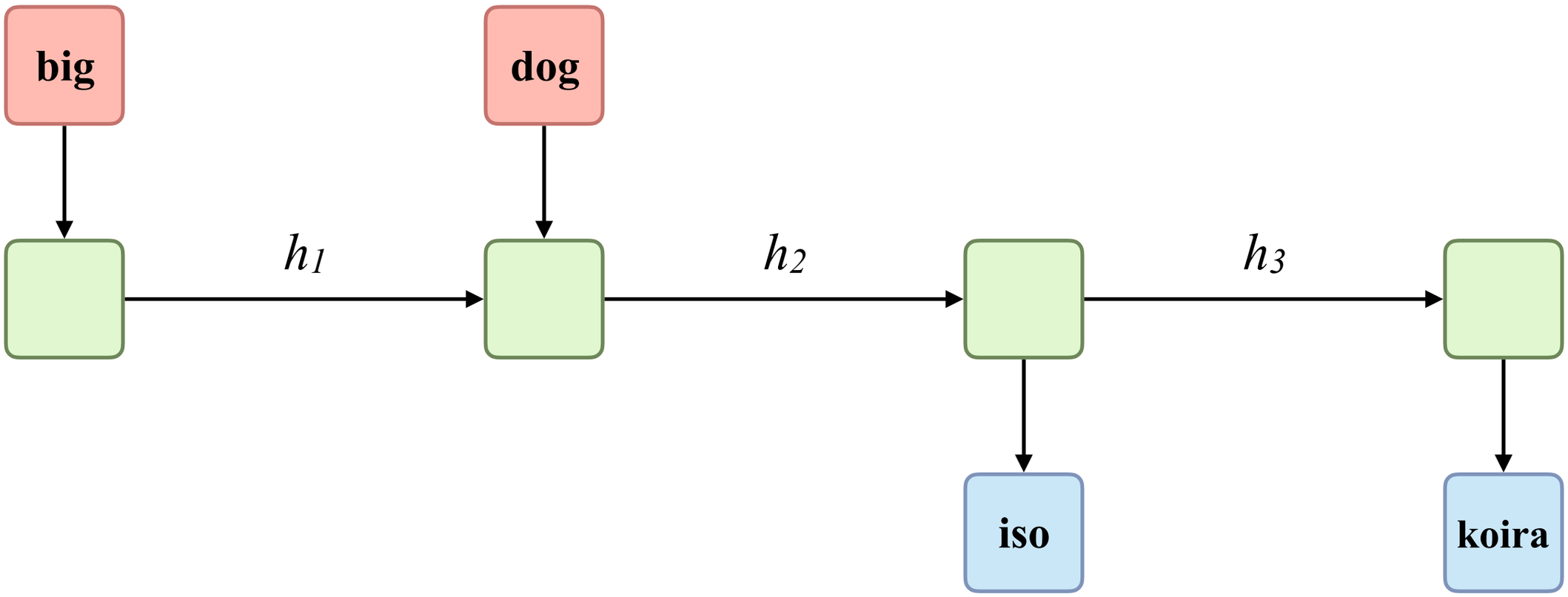
Sequence to sequence RNN

Recall many-to-many RNN example



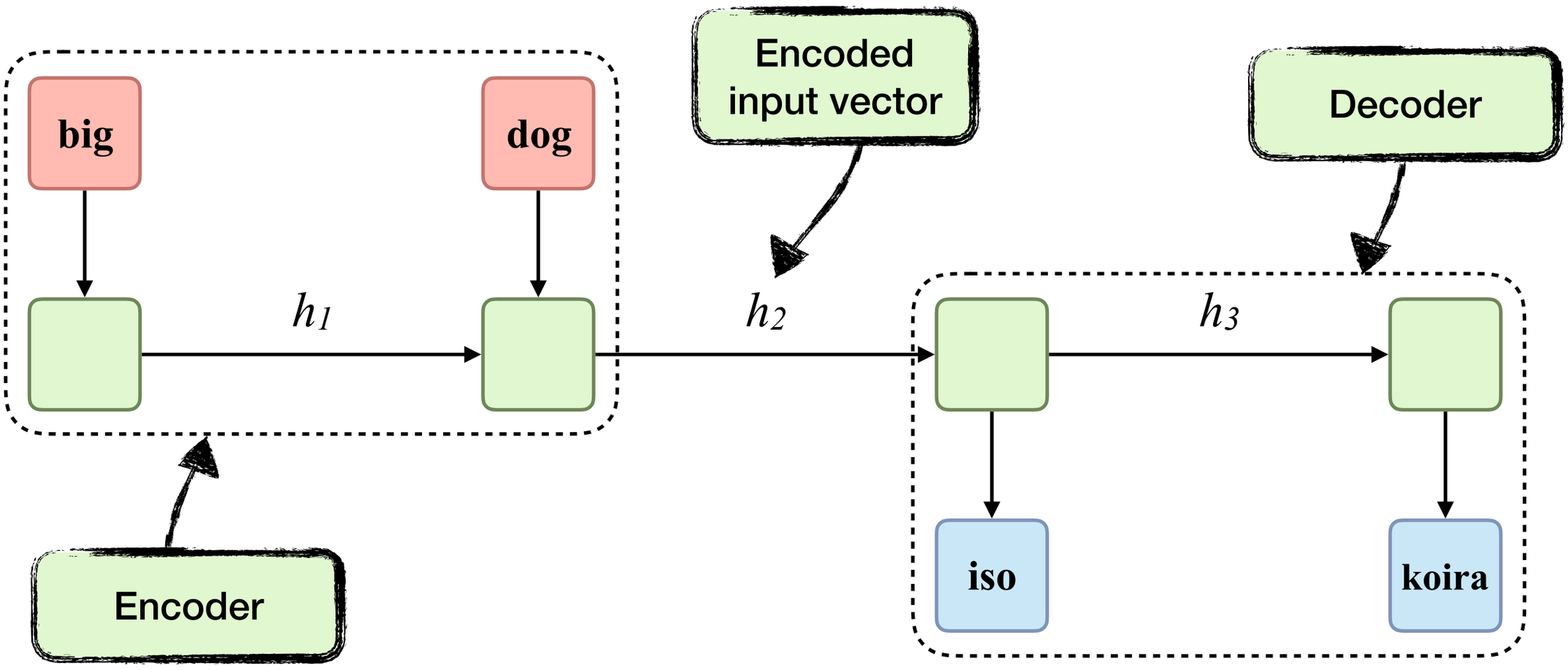
Sequence to sequence RNN

Recall many-to-many RNN example



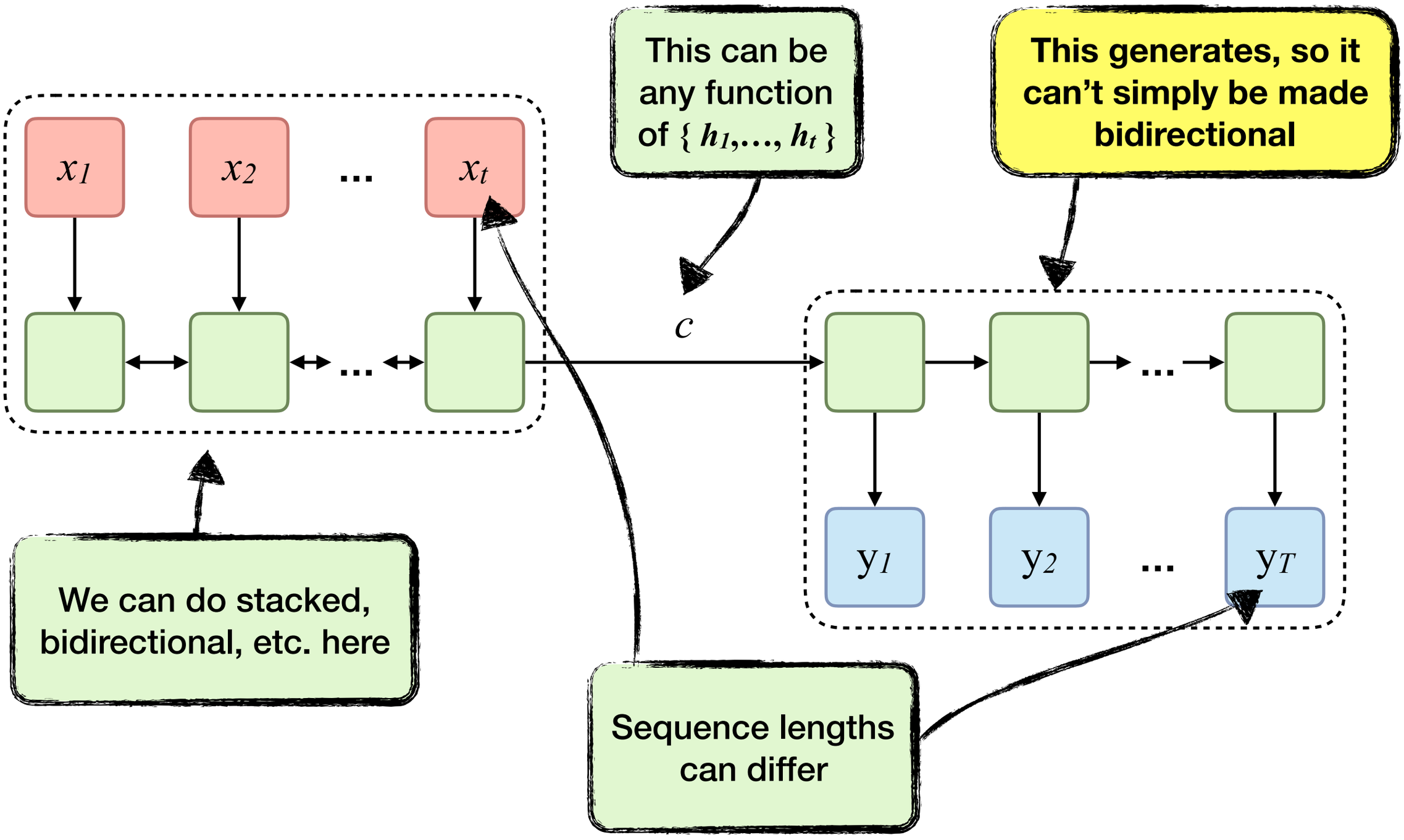
Sequence to sequence RNN

Many-to-many RNN example as encoder-decoder model



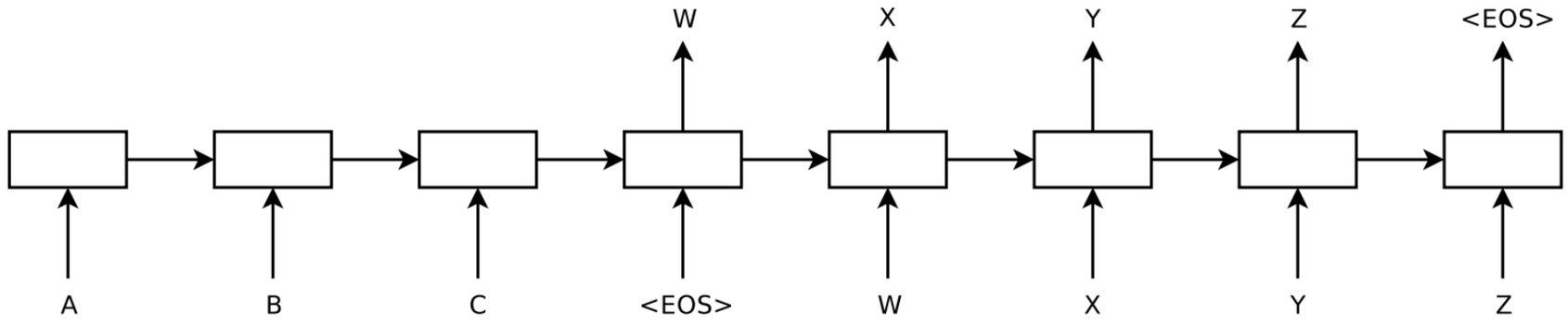
Sequence to sequence RNN

More generally:



Sequence to sequence RNN

This class of model is remarkably capable in machine translation:

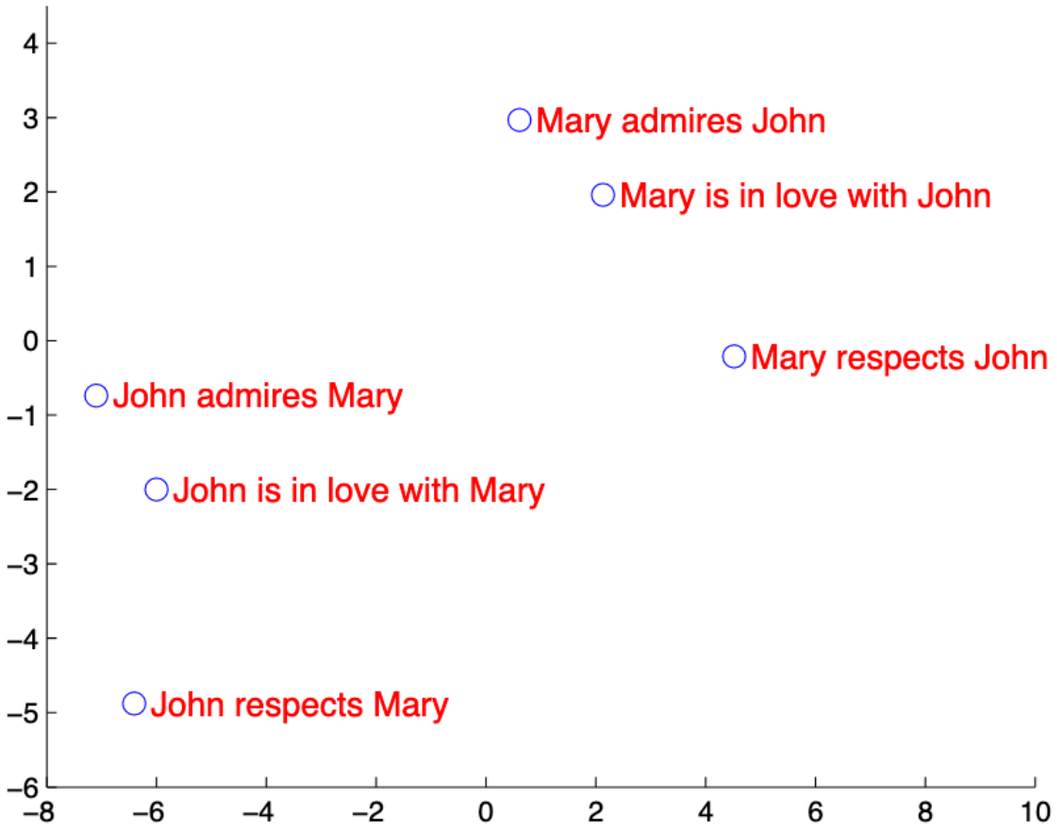


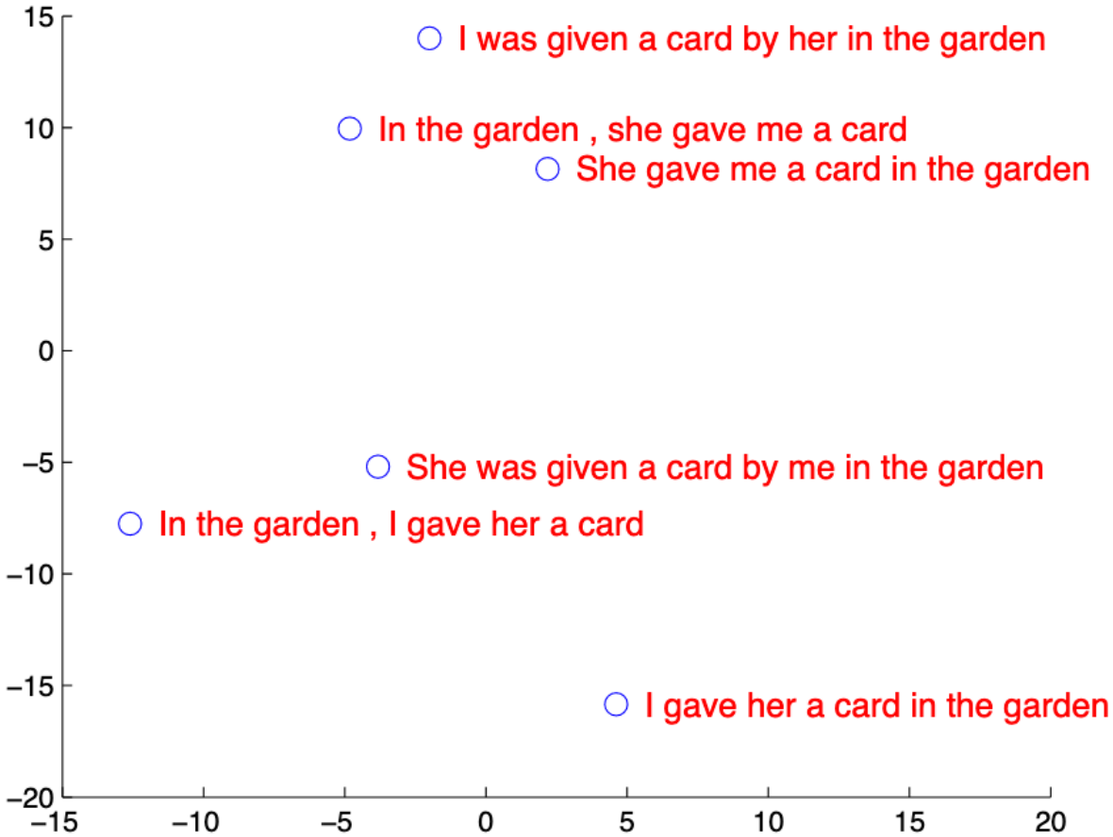
Our model	“ Les téléphones cellulaires , qui sont vraiment une question , non seulement parce qu’ ils pourraient potentiellement causer des interférences avec les appareils de navigation , mais nous savons , selon la FCC , qu’ ils pourraient interférer avec les tours de téléphone cellulaire lorsqu’ ils sont dans l’ air ” , dit UNK .
Truth	“ Les téléphones portables sont véritablement un problème , non seulement parce qu’ ils pourraient éventuellement créer des interférences avec les instruments de navigation , mais parce que nous savons , d’ après la FCC , qu’ ils pourraient perturber les antennes-relais de téléphonie mobile s’ ils sont utilisés à bord ” , a déclaré Rosenker .
Our model	Avec la crémation , il y a un “ sentiment de violence contre le corps d’ un être cher ” , qui sera “ réduit à une pile de cendres ” en très peu de temps au lieu d’ un processus de décomposition “ qui accompagnera les étapes du deuil ” .
Truth	Il y a , avec la crémation , “ une violence faite au corps aimé ” , qui va être “ réduit à un tas de cendres ” en très peu de temps , et non après un processus de décomposition , qui “ accompagnerait les phases du deuil ” .

Figures from Sutskever *et al*. (2014) *Sequence to Sequence Learning with Neural Networks*

Sequence to sequence RNN

Illustration of encoded vector representations (sentence embeddings)





1000-dim vectors

Figure from Sutskever *et al*. (2014) *Sequence to Sequence Learning with Neural Networks*

Limitations

Consider the task of translating the following:

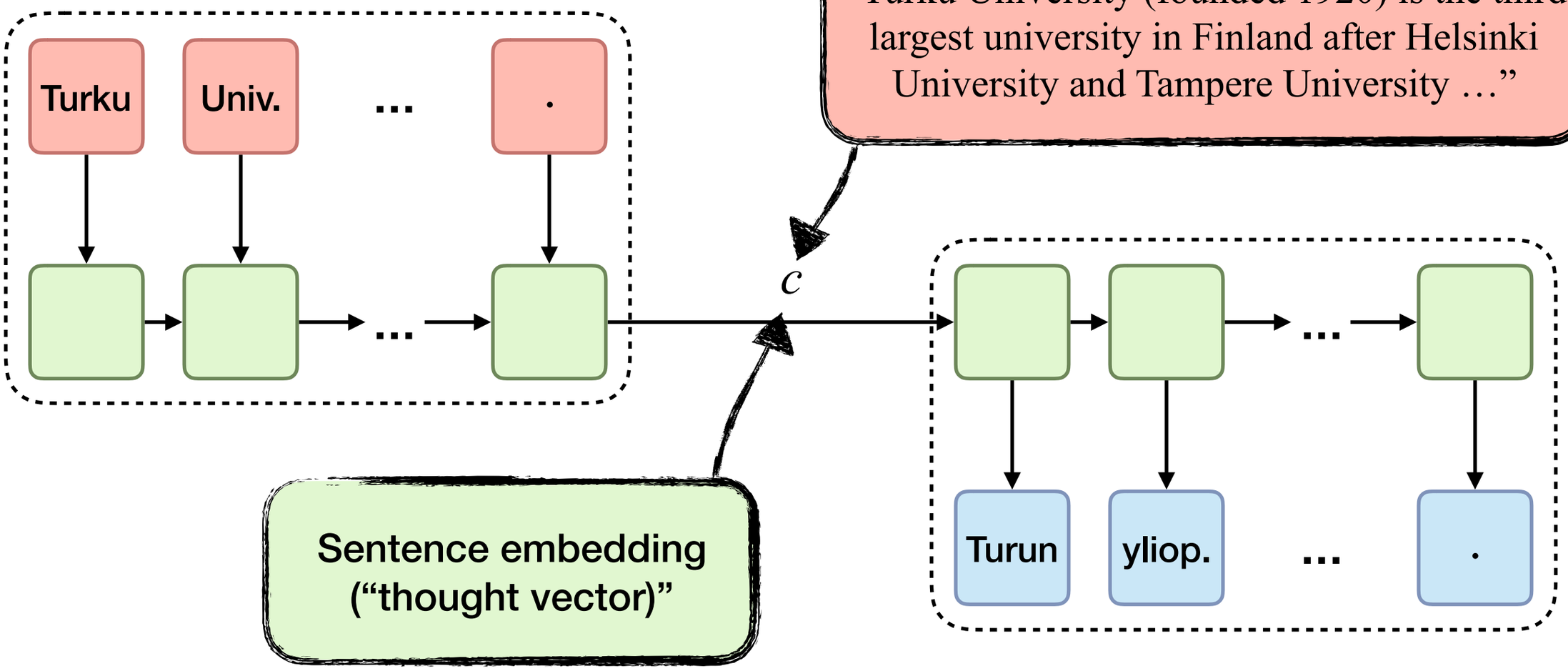
Turku University (founded 1920) is the third largest university in Finland after Helsinki University and Tampere University. The University has approximately 20,000 students, of which 5,000 are postgraduate students, and it operates in Rauma and Pori in addition to Turku.

(Take a moment to read that.)

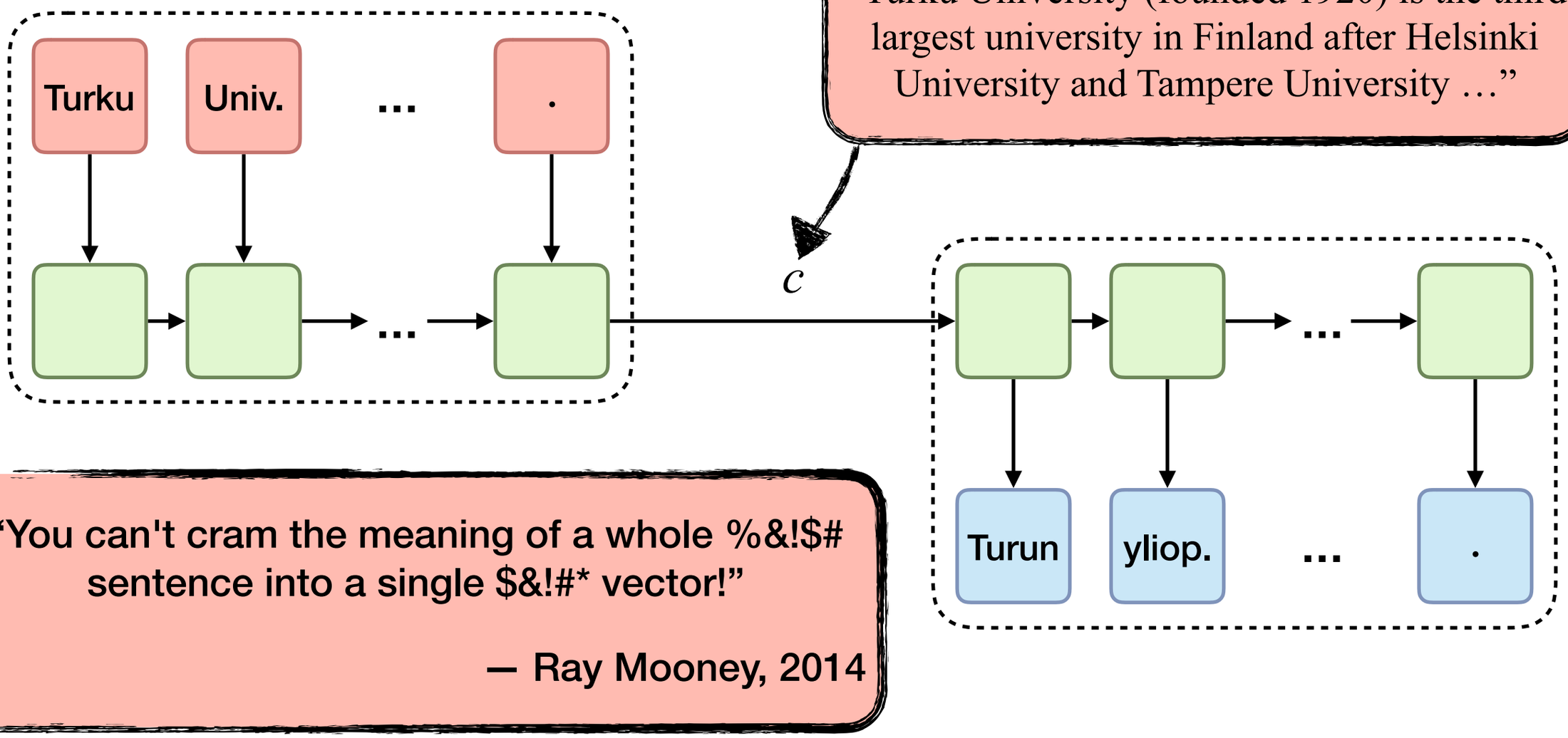
Limitations

(Now figure out the translation.)

Limitations



Limitations



“You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector!”

— Ray Mooney, 2014

Conneau *et al*. (2018) *What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties*

Limitations

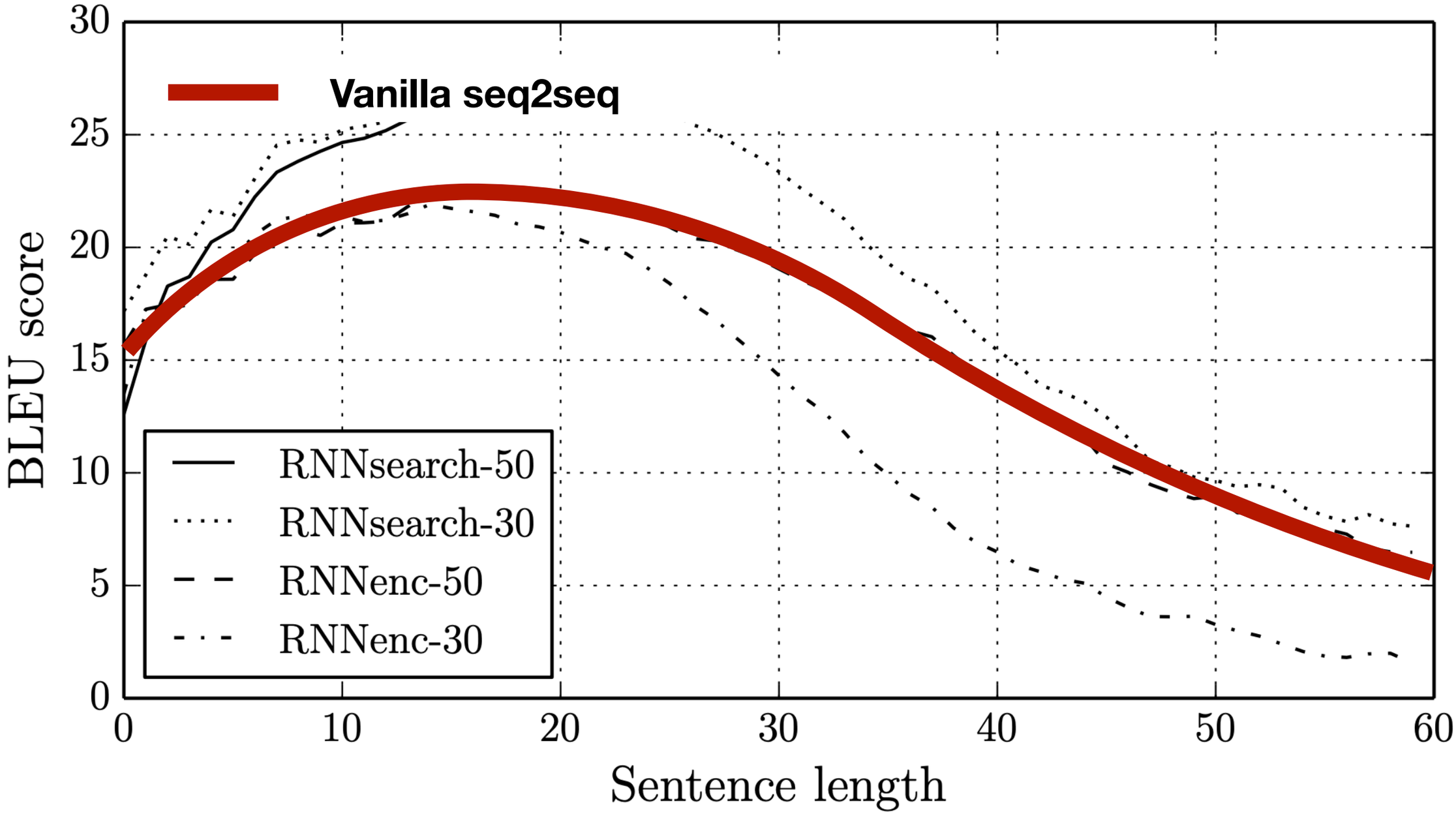
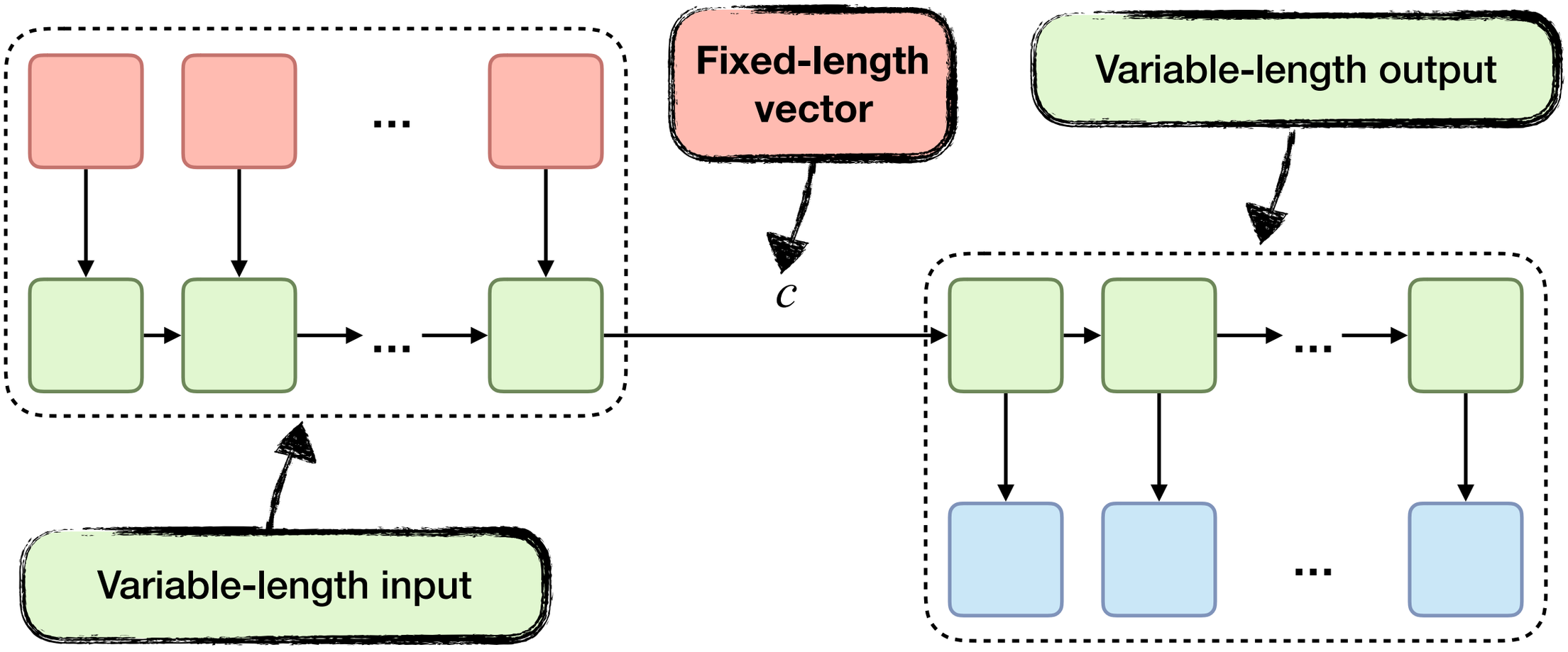


Figure from Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

Limitations

The vector representing the encoded input is a bottleneck



Summary

Encoder-decoder architectures can be applied to train end-to-end models for a broad range of tasks, including cross-modal (image/audio/text)

RNNs can be used straightforwardly to implement **sequence-to-sequence** encoder-decoder models applicable to e.g. machine translation, summarization, and dialogue systems

While powerful, the **fixed-length encoded vector representation** of the input represents a bottleneck for these models

Neural attention

Attention: background

Focus on one thing while ignoring others

Rough intuition: **make whole input available**, allow model to choose what to pay attention to

First NN models proposed in computer vision (Larochelle and Hinton 2010)

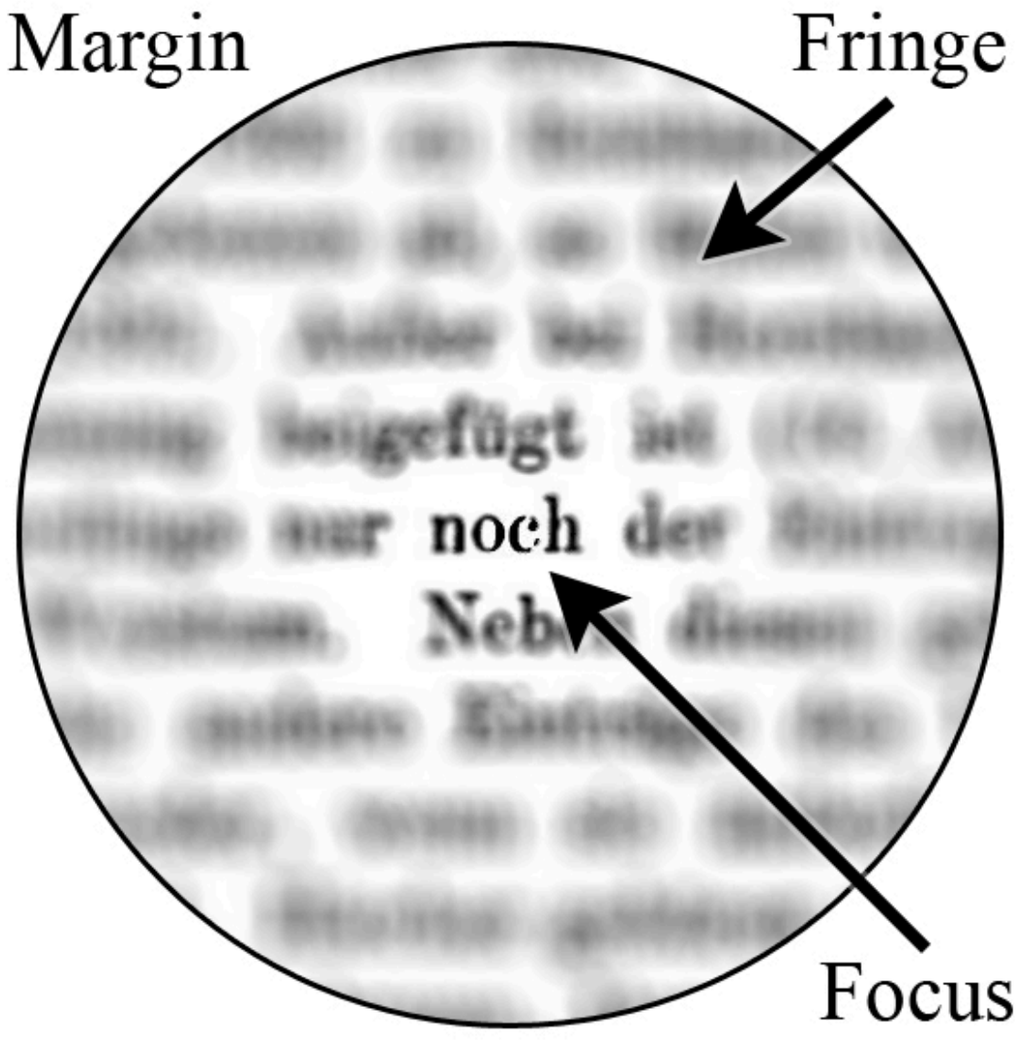


Illustration of “spotlight” model of visual attention from https://en.wikipedia.org/wiki/Attention

Larochelle and Hinton (2010) *Learning to combine foveal glimpses with a third-order Boltzmann machine*

Attention: intuition

Consider (again) the task of translating the following:

Turku University (founded 1920) is the third largest university in Finland after Helsinki University and Tampere University. The University has approximately 20,000 students, of which 5,000 are postgraduate students, and it operates in Rauma and Pori in addition to Turku.

(No need to try to memorize it this time)

Attention: intuition

Translating piece by piece:

Turku University (founded 1920) is the third largest university in Finland after Helsinki University and Tampere University. The University has approximately 20,000 students, of which 5,000 are postgraduate students, and it operates in Rauma and Pori in addition to Turku.

Attention: intuition

Translating piece by piece:

Turku University (founded 1920) is the third largest university in Finland after Helsinki University and Tampere University. The University has approximately 20,000 students, of which 5,000 are postgraduate students, and it operates in Rauma and Pori in addition to Turku.

Turun yliopisto (perustettu 1920)

Attention: intuition

Translating piece by piece:

Turku University (founded 1920) is the third largest university in Finland after Helsinki University and Tampere University. The University has approximately 20,000 students, of which 5,000 are postgraduate students, and it operates in Rauma and Pori in addition to Turku.

Turun yliopisto (perustettu 1920) on Suomen kolmanneksi suurin yliopisto

Attention: intuition

Translating piece by piece:

Turku University (founded 1920) is the third largest university in Finland after Helsinki University and Tampere University. The University has approximately 20,000 students, of which 5,000 are postgraduate students, and it operates in Rauma and Pori in addition to Turku.

Turun yliopisto (perustettu 1920) on Suomen kolmanneksi suurin yliopisto Helsingin yliopiston ja Tampereen yliopiston jälkeen.

Attention: intuition

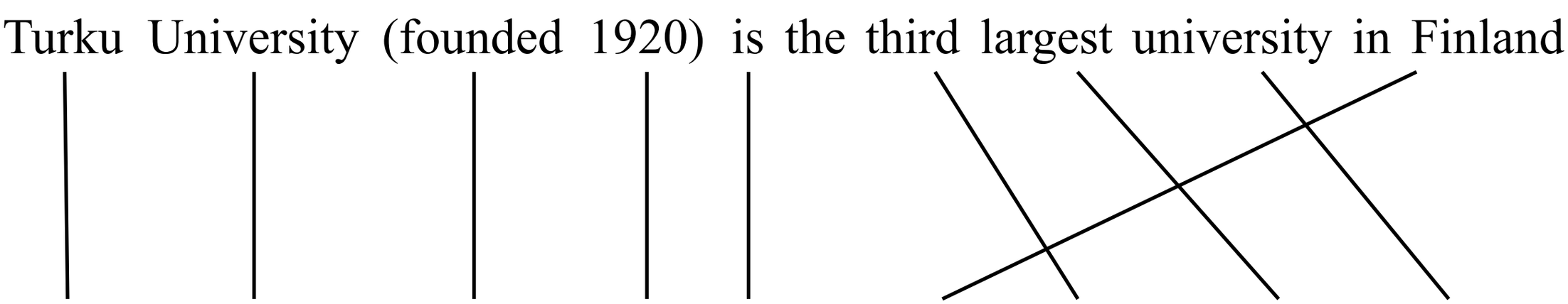
What to pay attention to when translating?

Turku University (founded 1920) is the third largest university in Finland

Turun yliopisto (perustettu 1920) on Suomen kolmanneksi suurin yliopisto

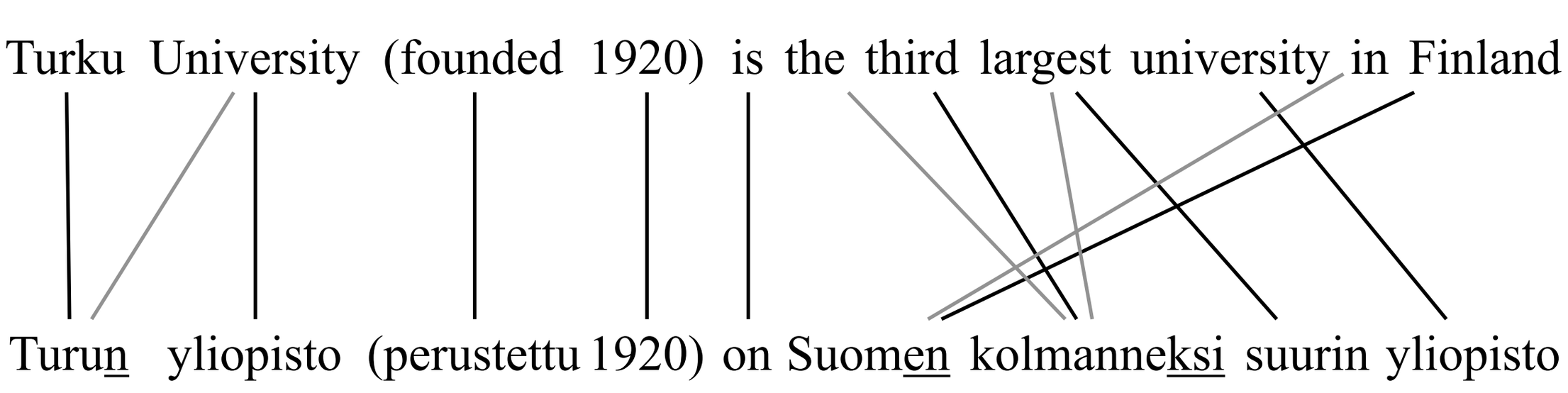
Attention: intuition

What to pay attention to when translating?



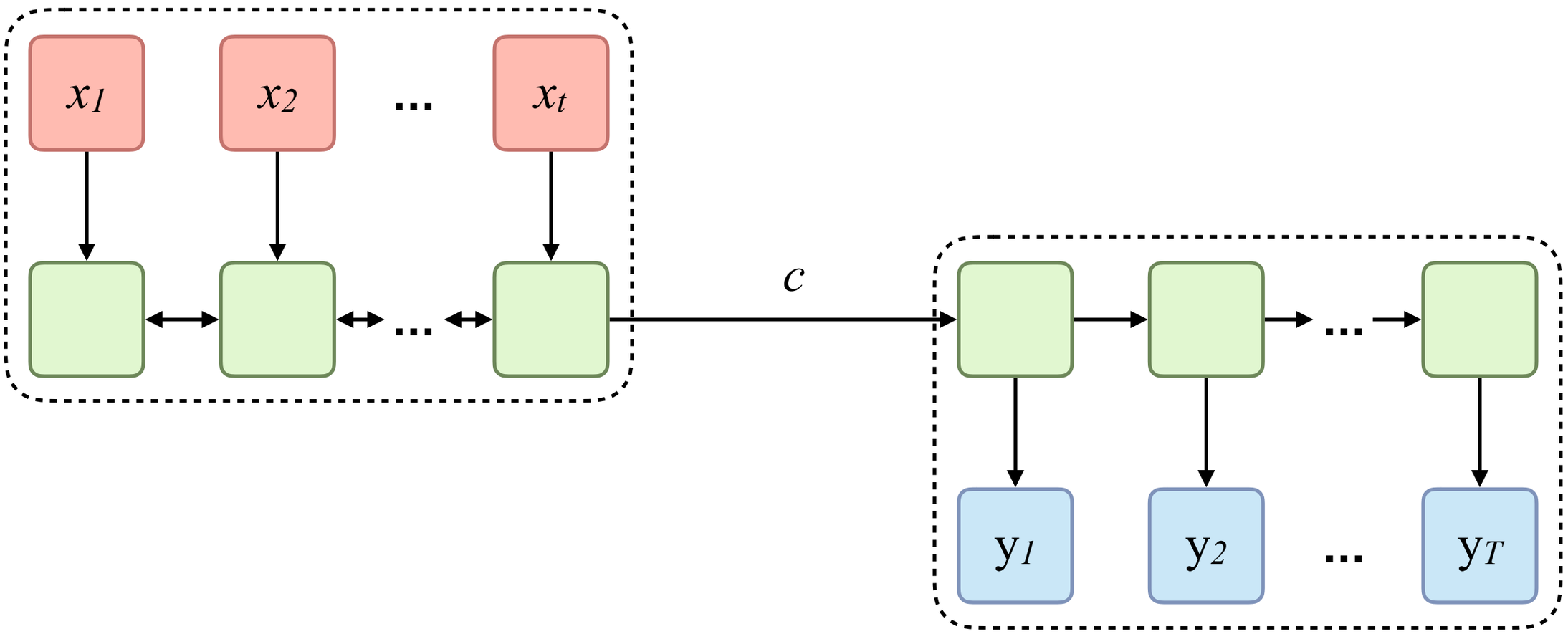
Attention: intuition

What to pay attention to when translating? (Not only single words)



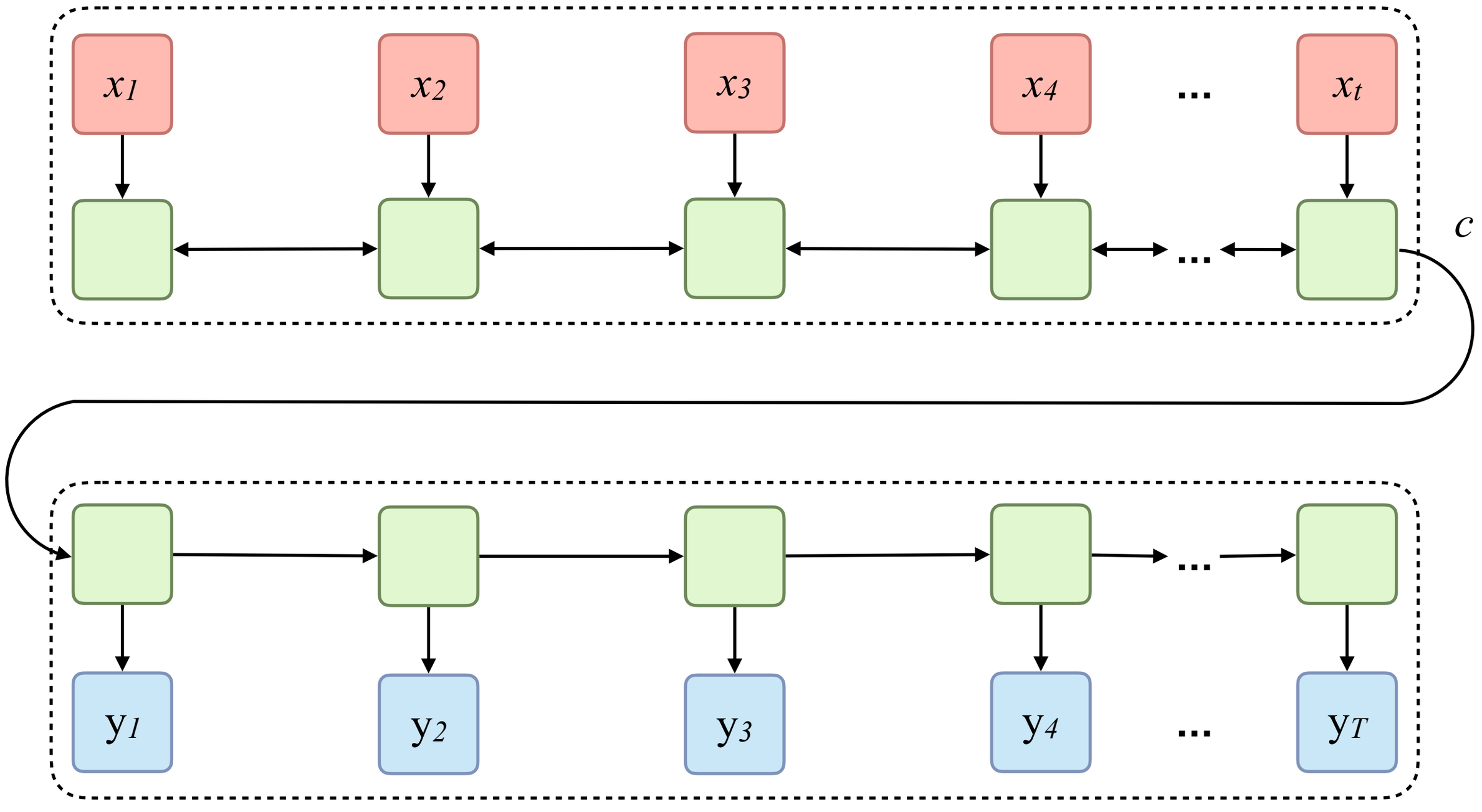
Neural attention

Conventional sequence to sequence



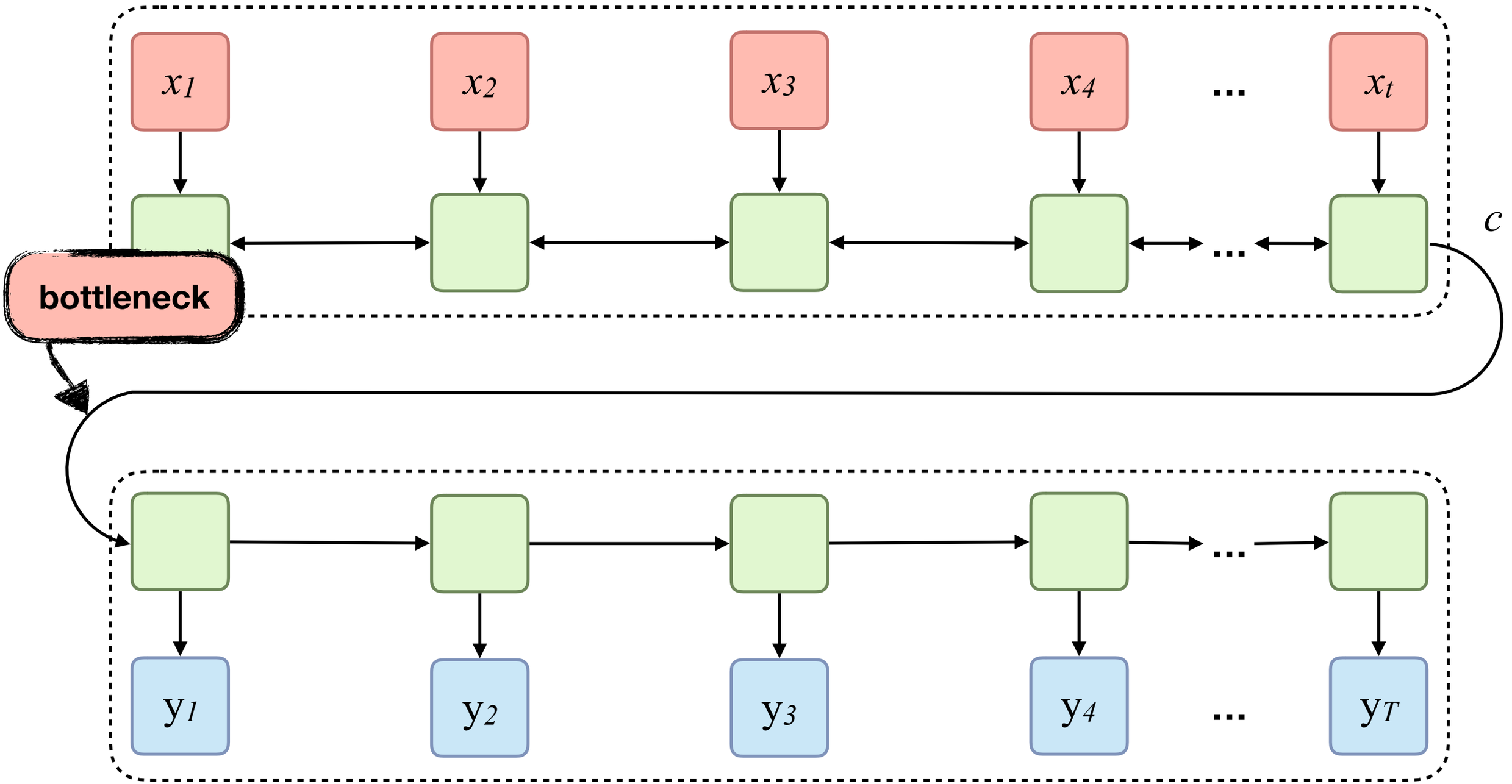
Neural attention

Conventional sequence to sequence (same thing, different layout)



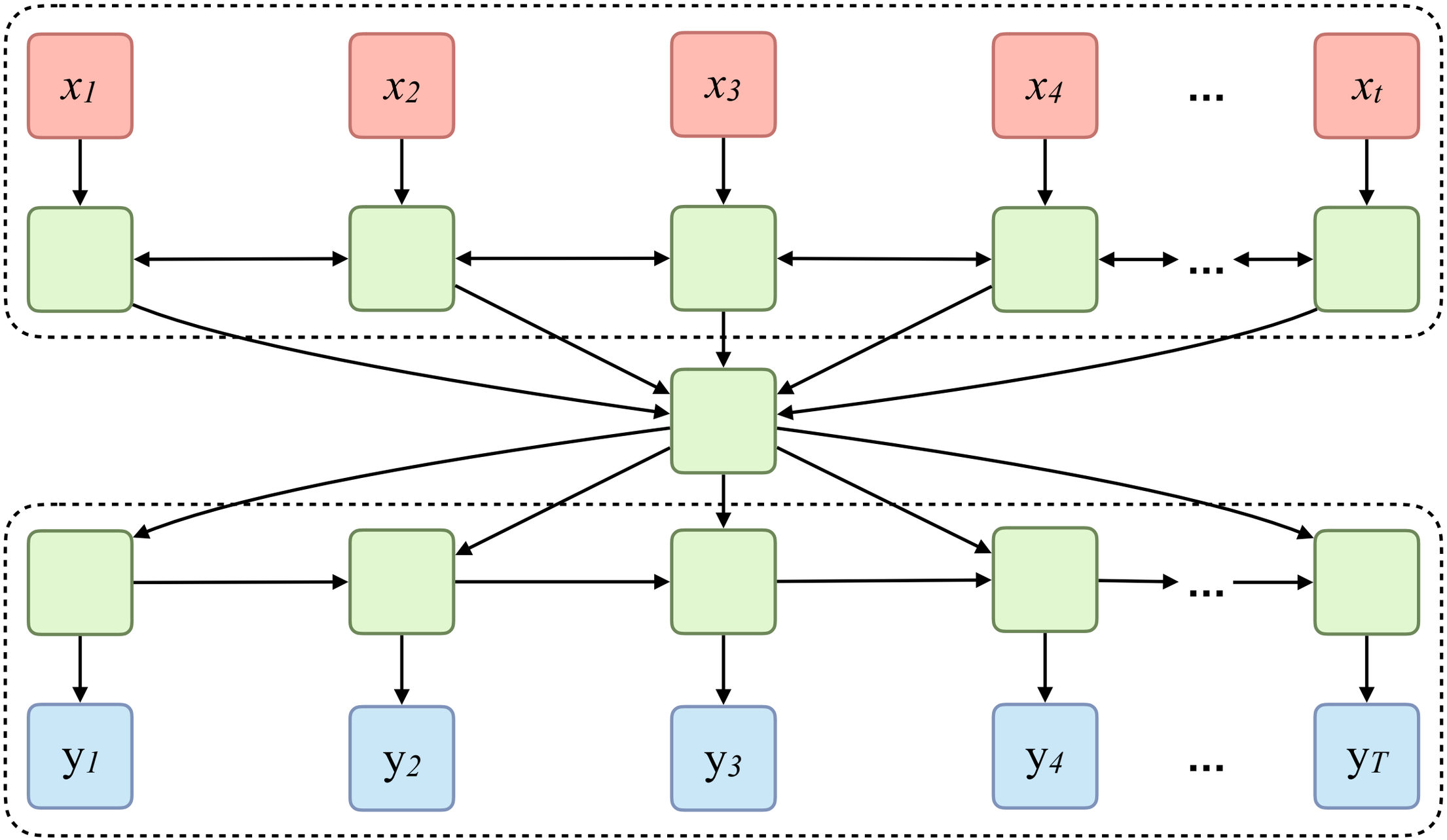
Neural attention

Conventional sequence to sequence (same thing, different layout)



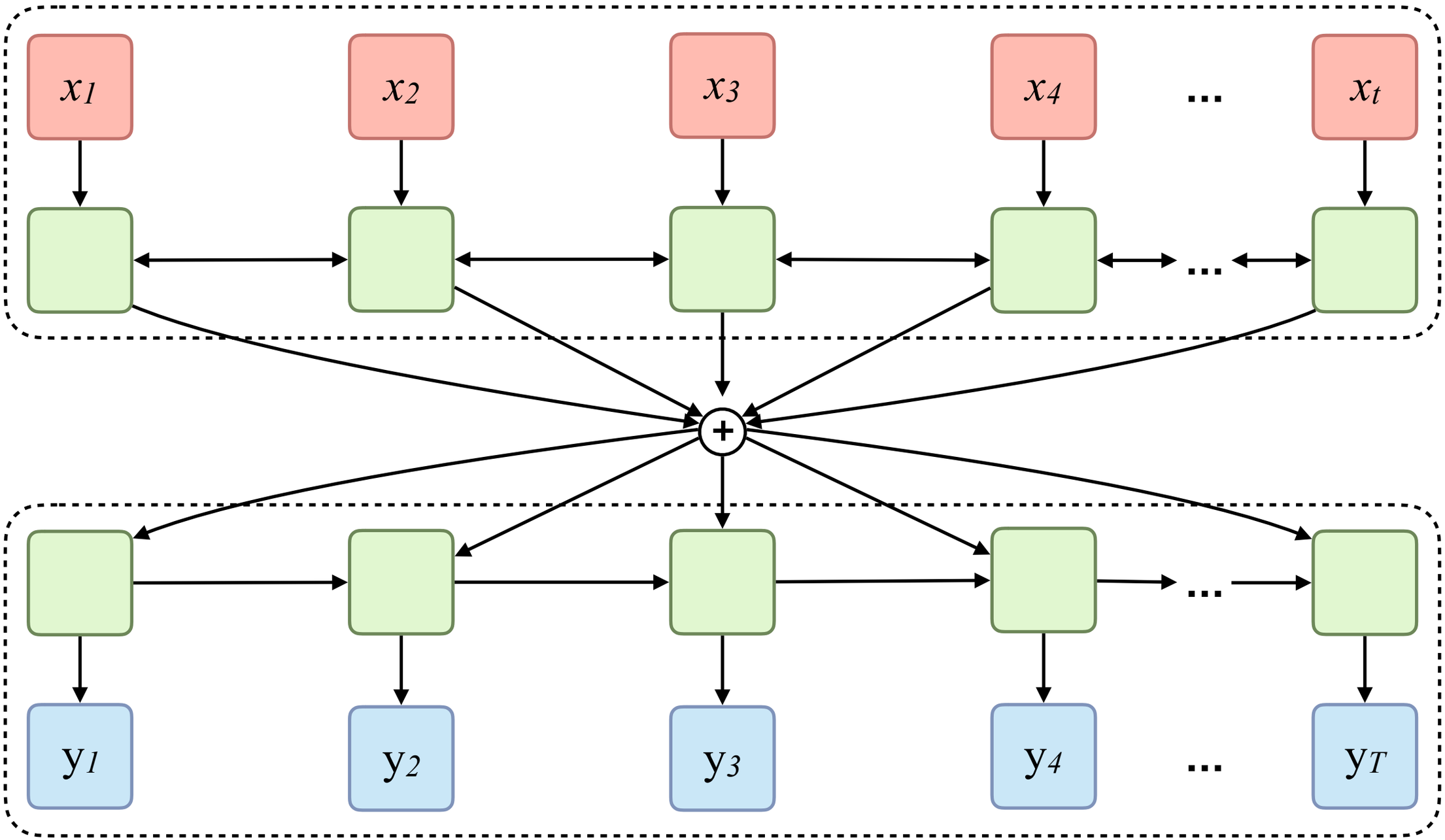
Neural attention

Rough idea: what if we just combined all input states (somehow)?



Neural attention

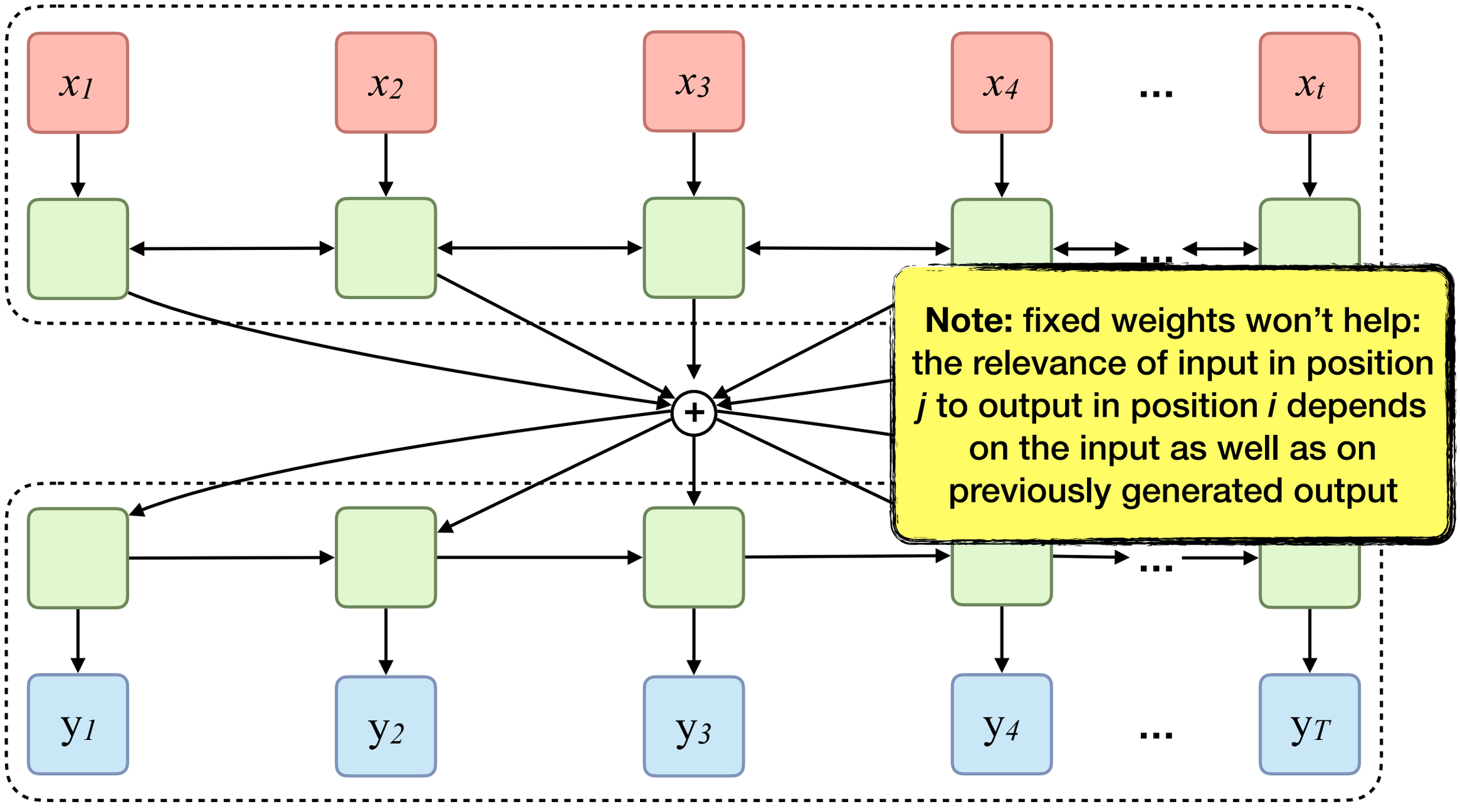
Attention model: weighted sum of input states



Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

Neural attention

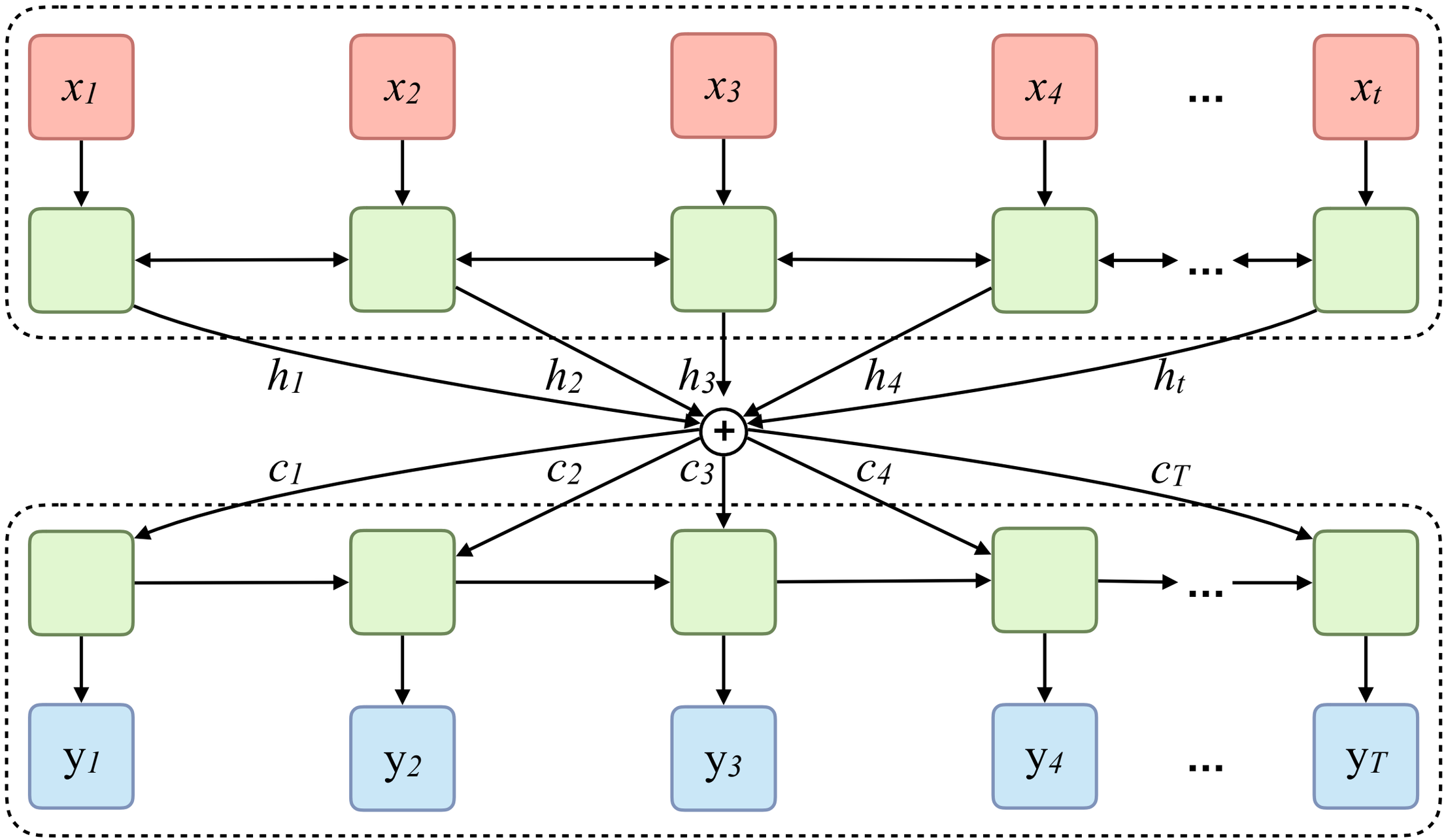
Attention model: weighted sum of input states



Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

Neural attention

Attention model



Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

Neural attention

Instead of a fixed-length vector representing encoded input, decoder has access to any part of the encoder state

Separate *context vector* c_i computed for each decoder state as a weighted sum of encoder states $h_1, \ldots, h_t$

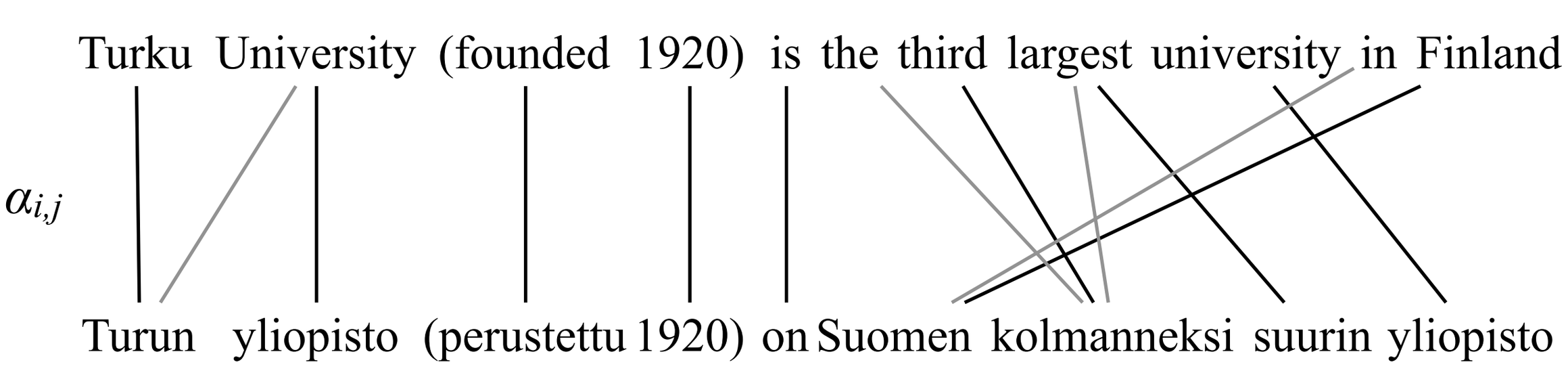
$$c_i = \sum_j \alpha_{i,j} h_j$$

where the *attention weights* $\alpha_{i,j}$ are computed as a function aiming to reflect the relevance of input position j to output position i

→ decoder steps can “pay attention” to different parts of the input

Attention: intuition

For MT, attention weights $\alpha_{i,j}$ ~ word alignment of input j and output i



Attention: intuition

For MT, attention weights $\alpha_{i,j}$ ~ word alignment of input j and output i

Diagonal shows word-to-word mappings (e.g. “in August 1992” - “en août 1992”)

Note inversion for “European Economic Area” vs. “zone économique european”

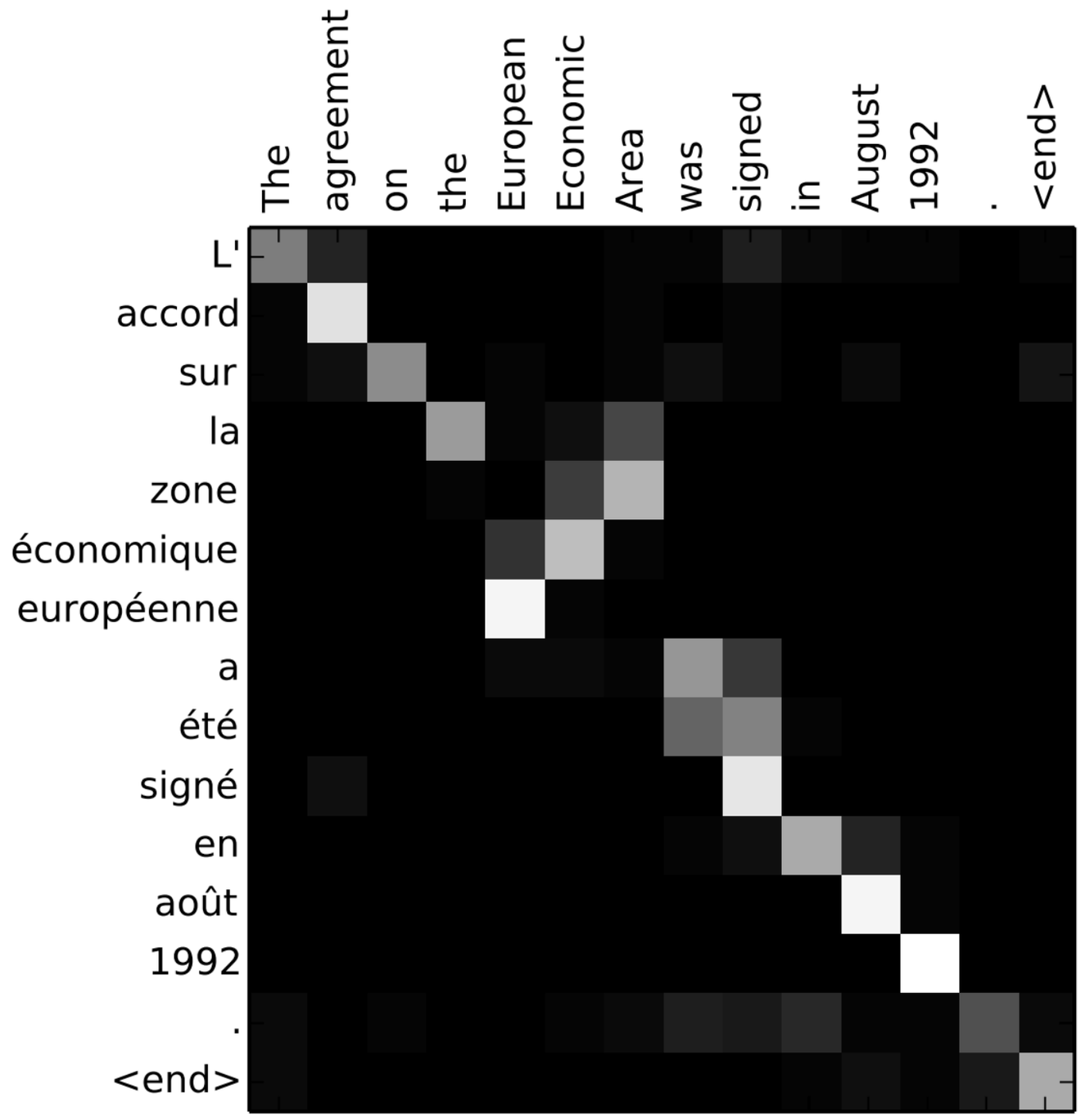
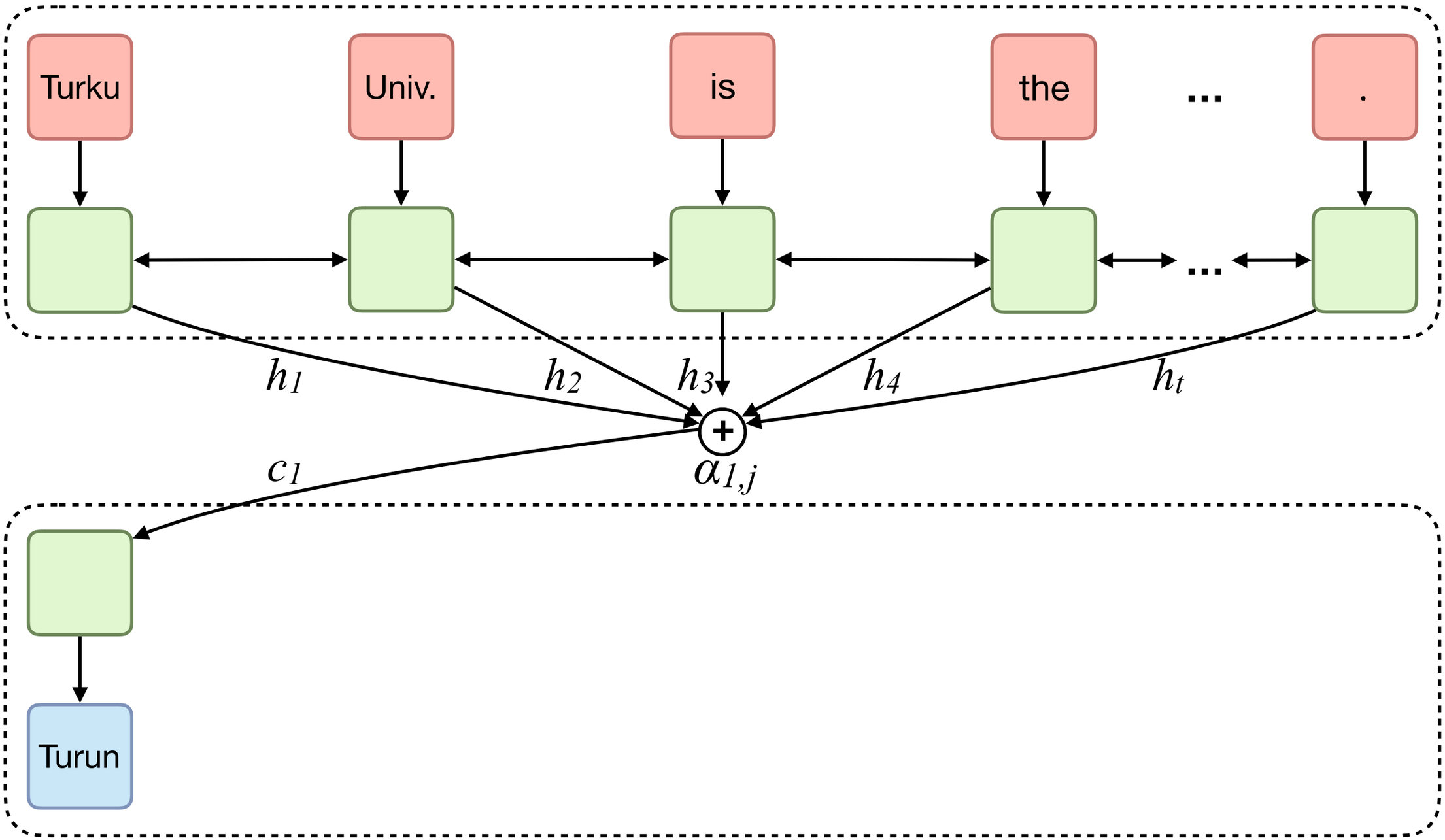
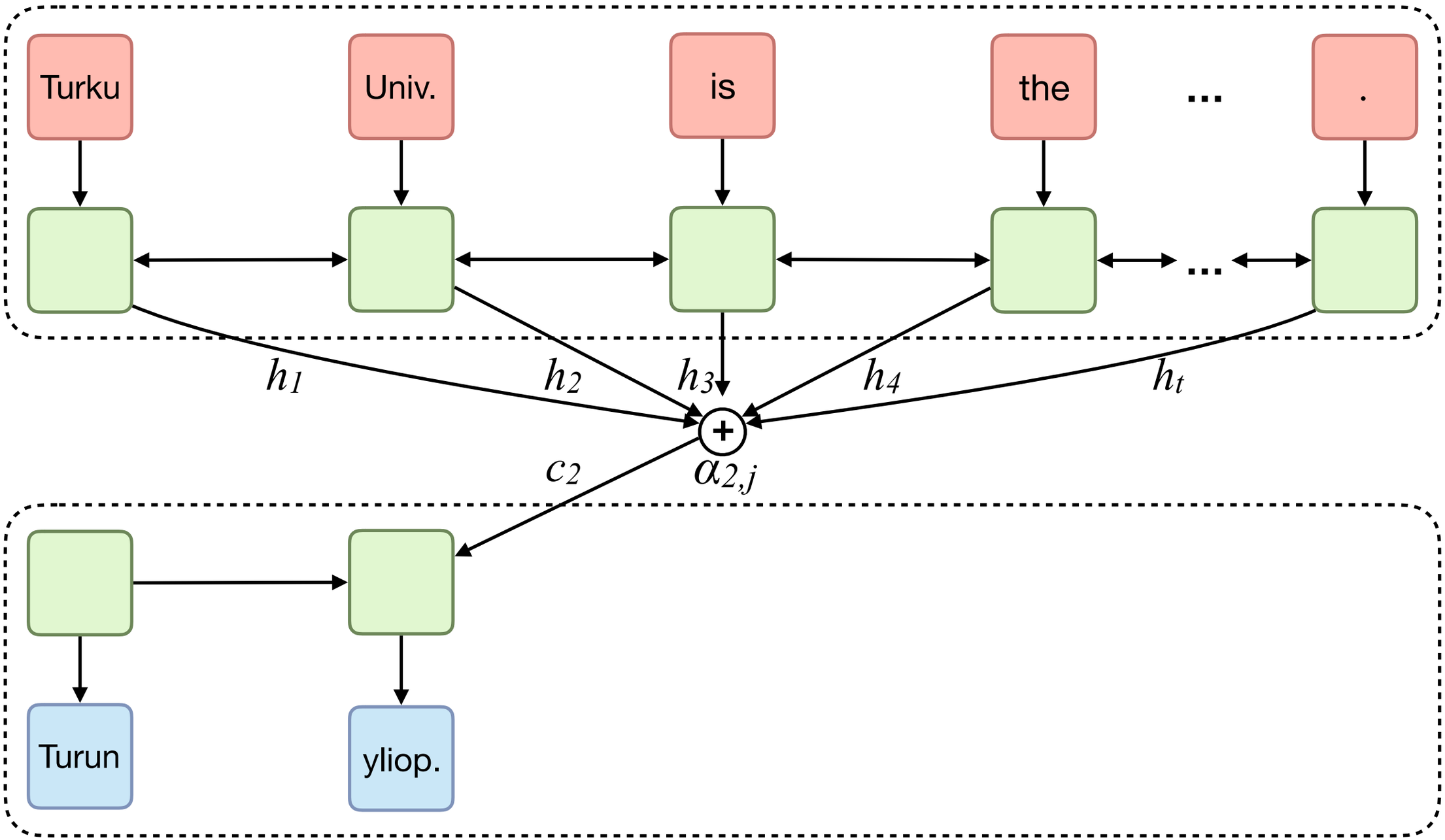


Figure from Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

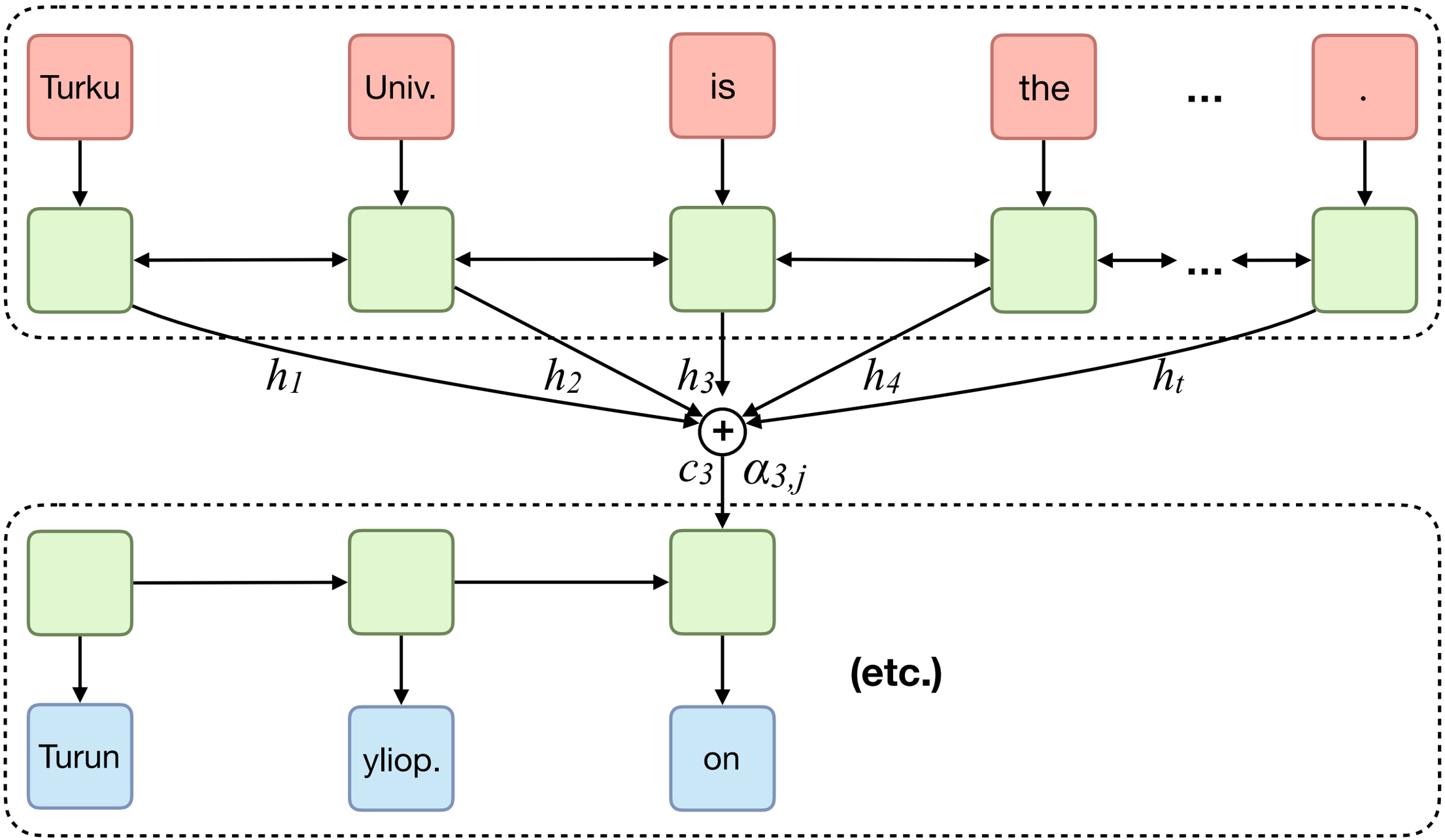
Neural attention: decoding



Neural attention: decoding



Neural attention: decoding



Neural attention

Calculating attention weights $\alpha_{i,j}$

cf. softmax

$$\alpha_{i,j} = \exp(e_{i,j}) / \sum_k \exp (e_{i,k})$$

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

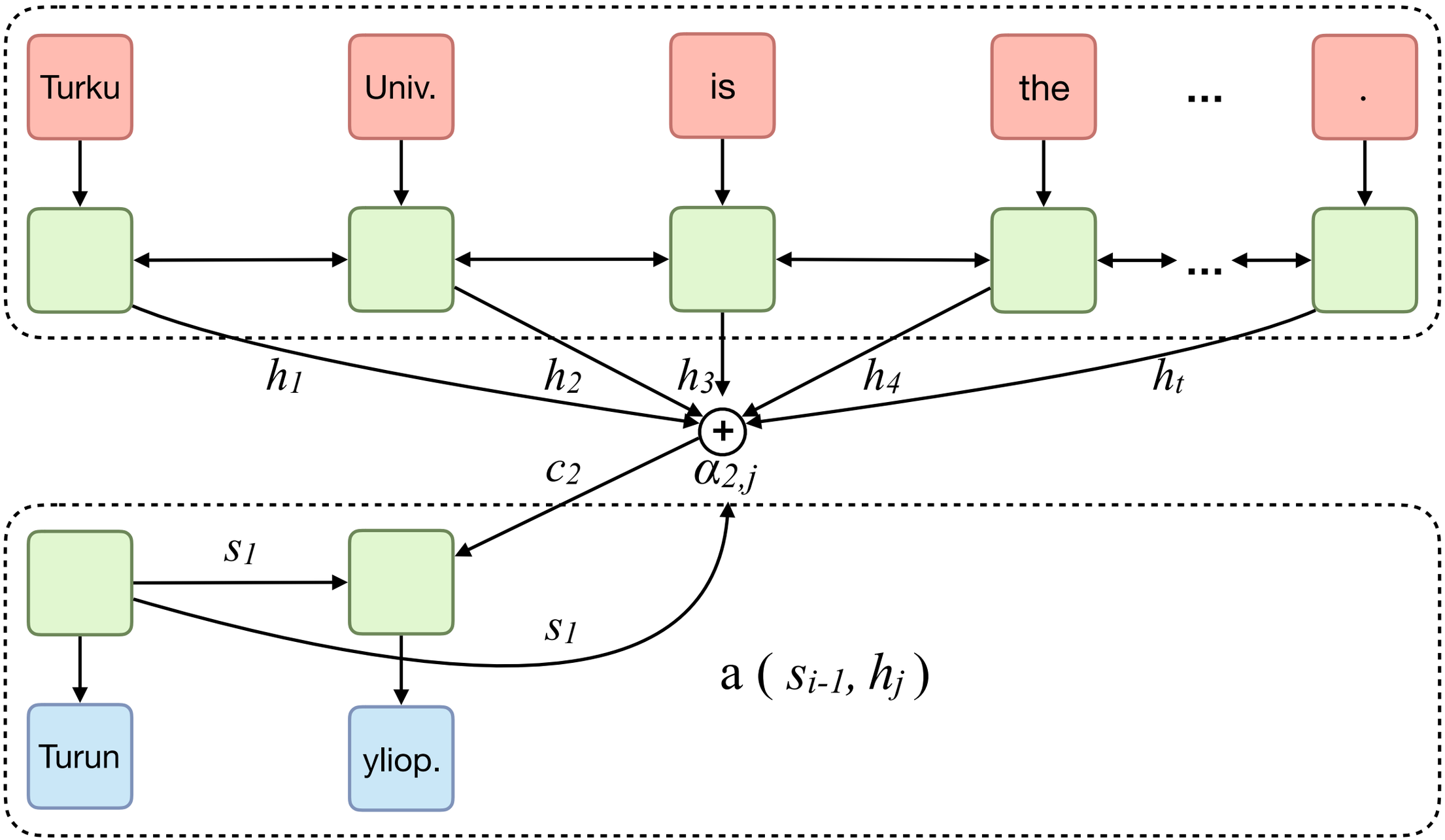
alignment function, learned (small!) NN

Where s_i is the i th state of the decoder

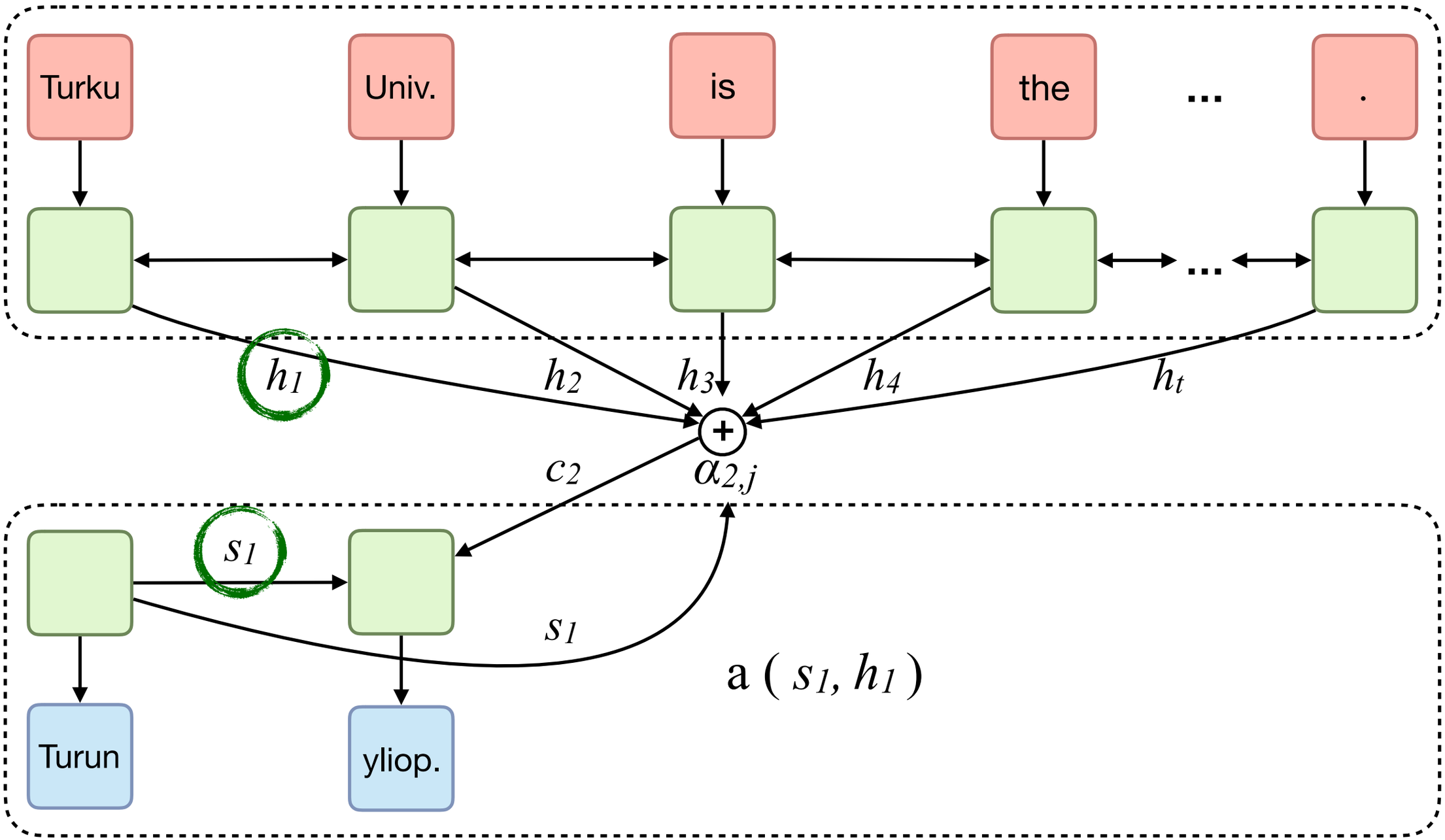
(Note: all $\alpha_{i,j} > 0$, $\sum_j \alpha_{i,j} = 1$ for all i)

Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

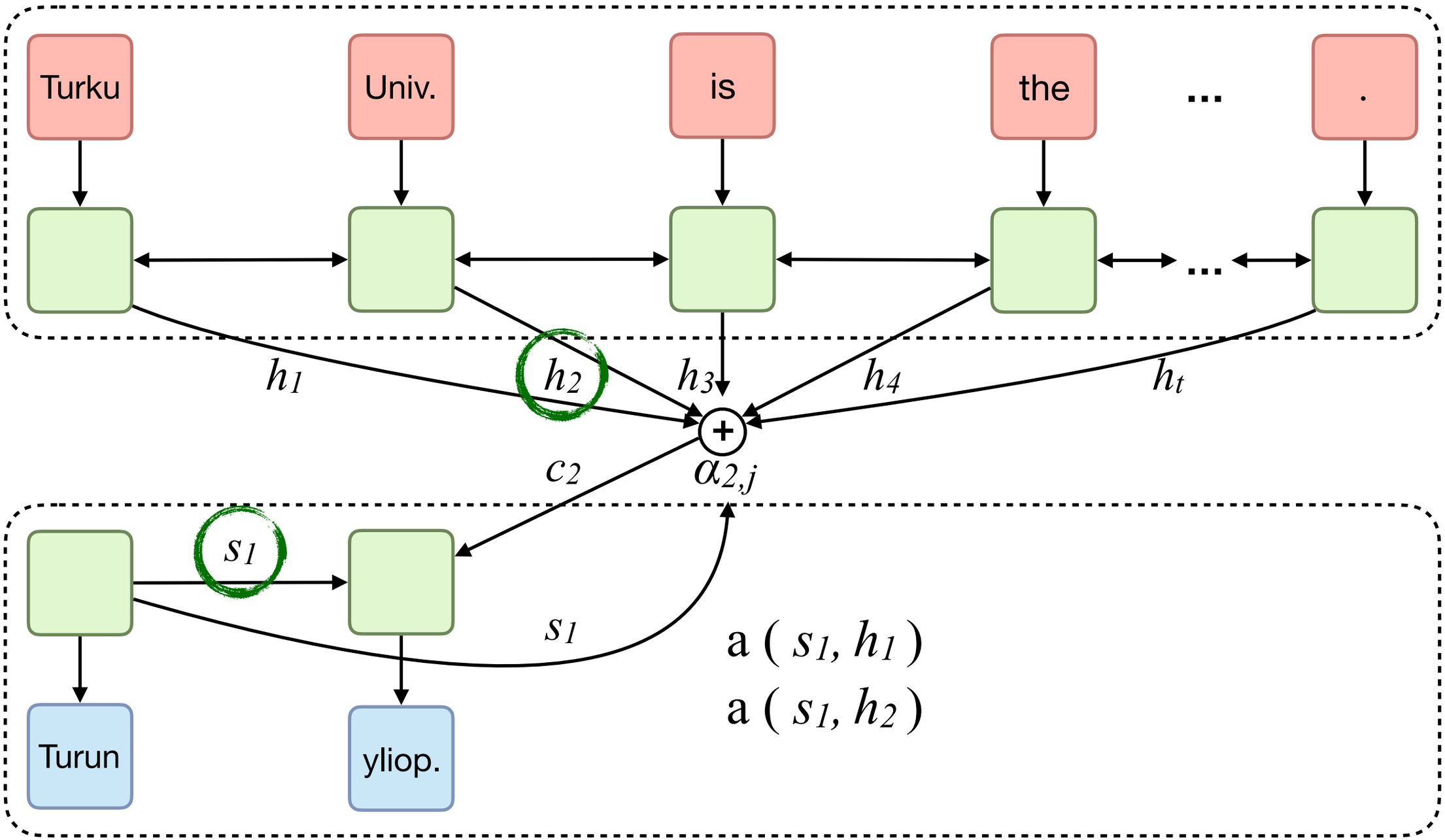
Neural attention



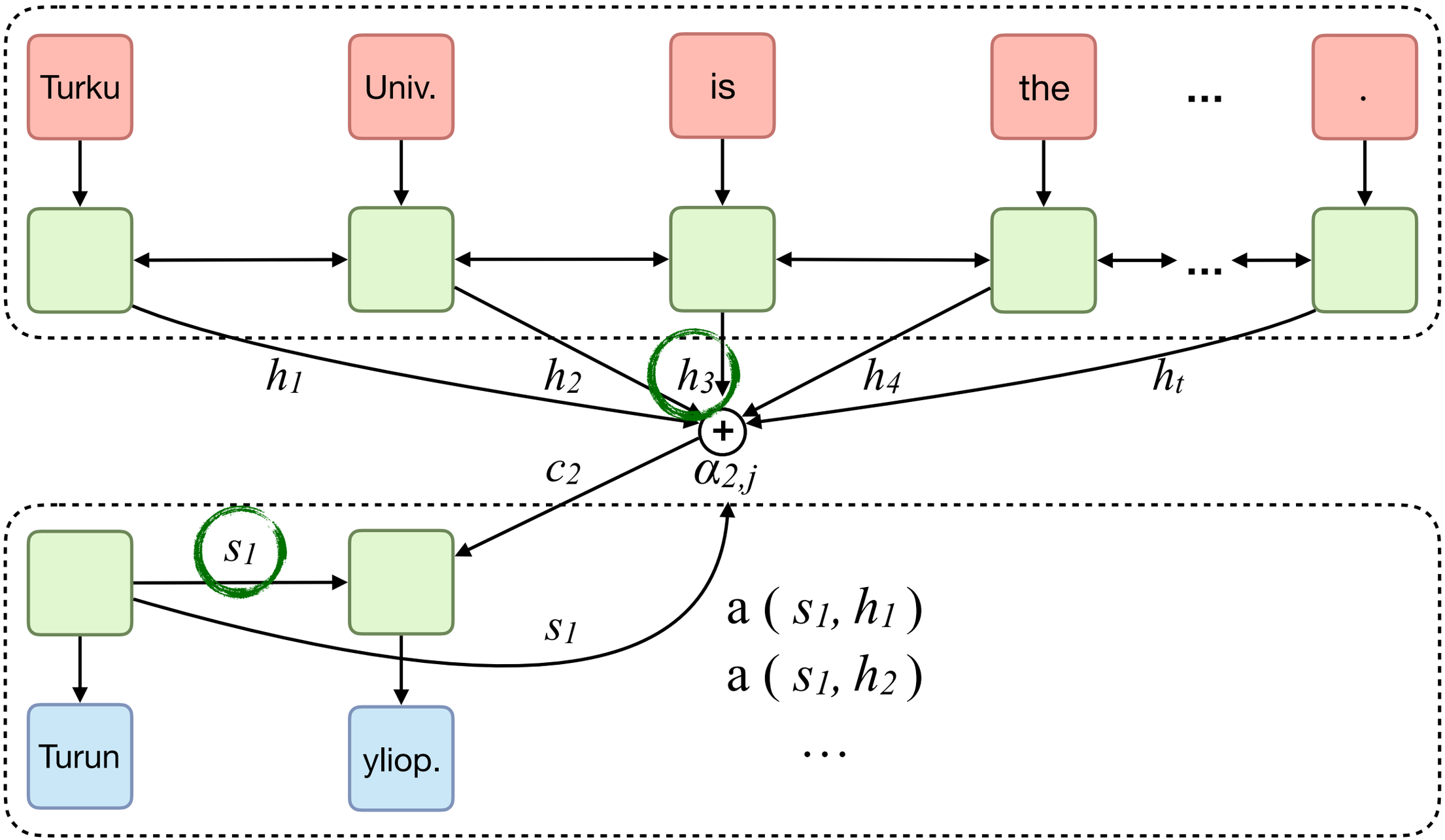
Neural attention



Neural attention

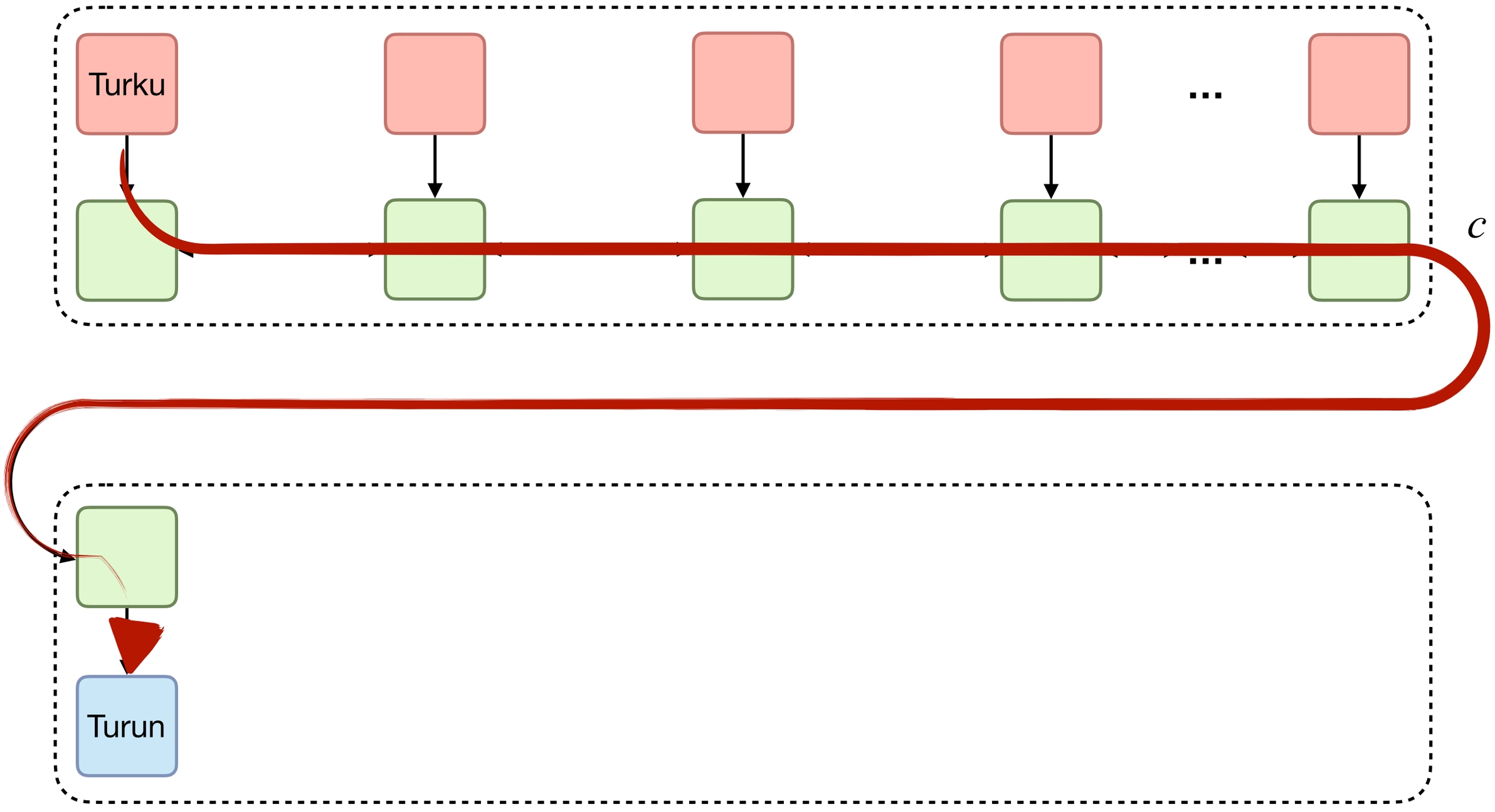


Neural attention



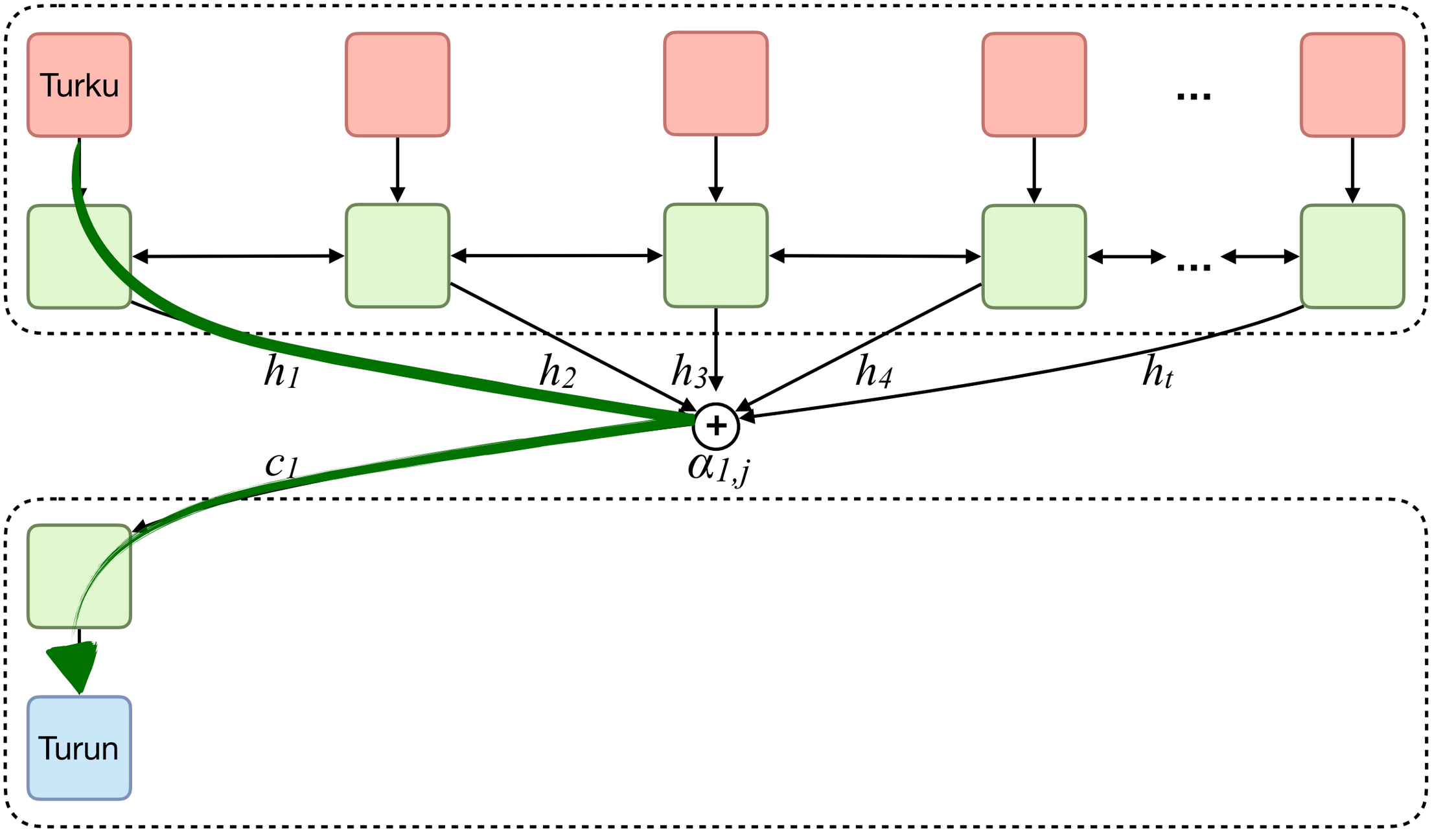
Neural attention

Seq-to-seq RNN: input-output distance grows with sentence length



Neural attention

Attention: constant distance from input to output



Applications: MT

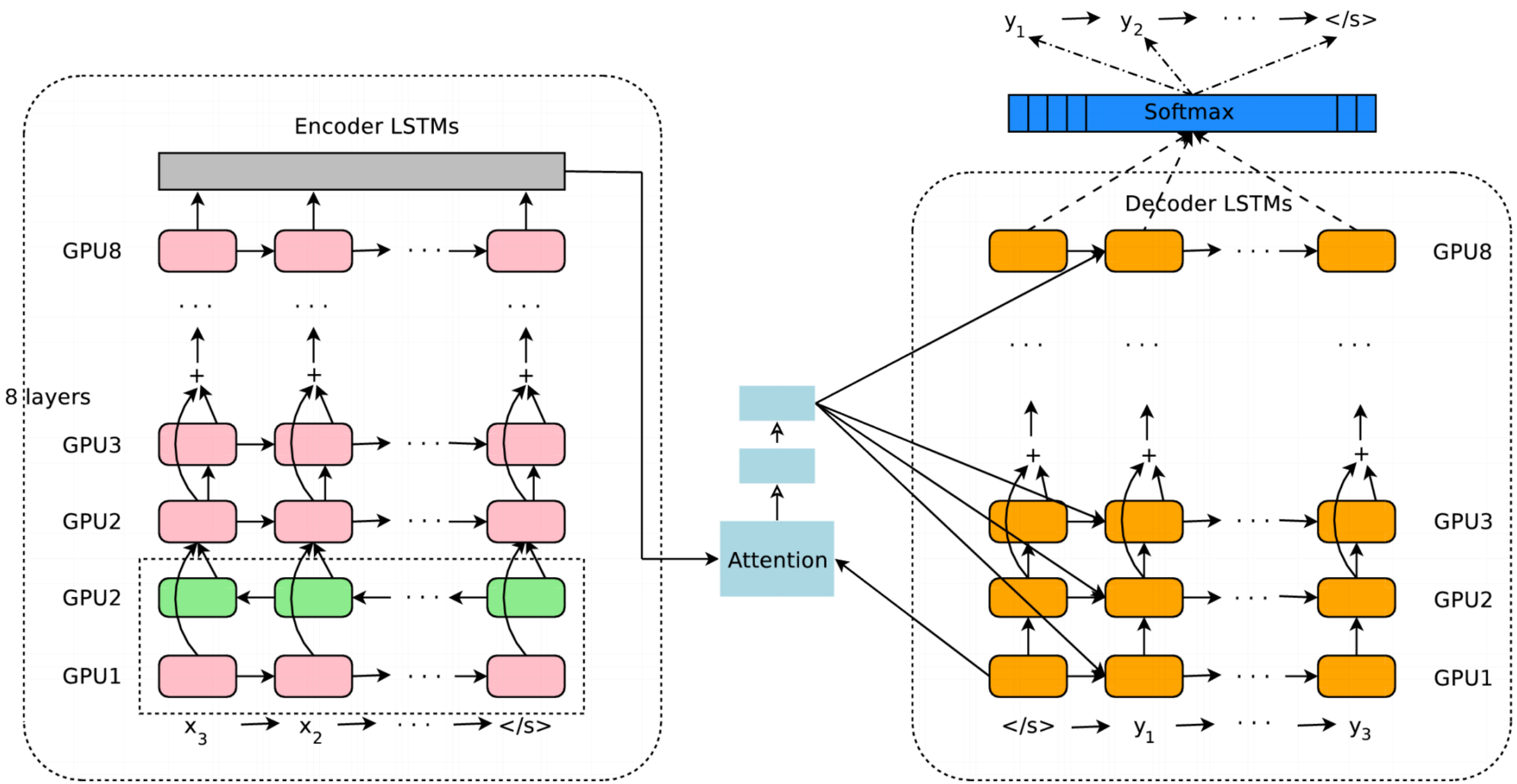


Figure from Wu *et al*. (2016) *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*

Applications: MT

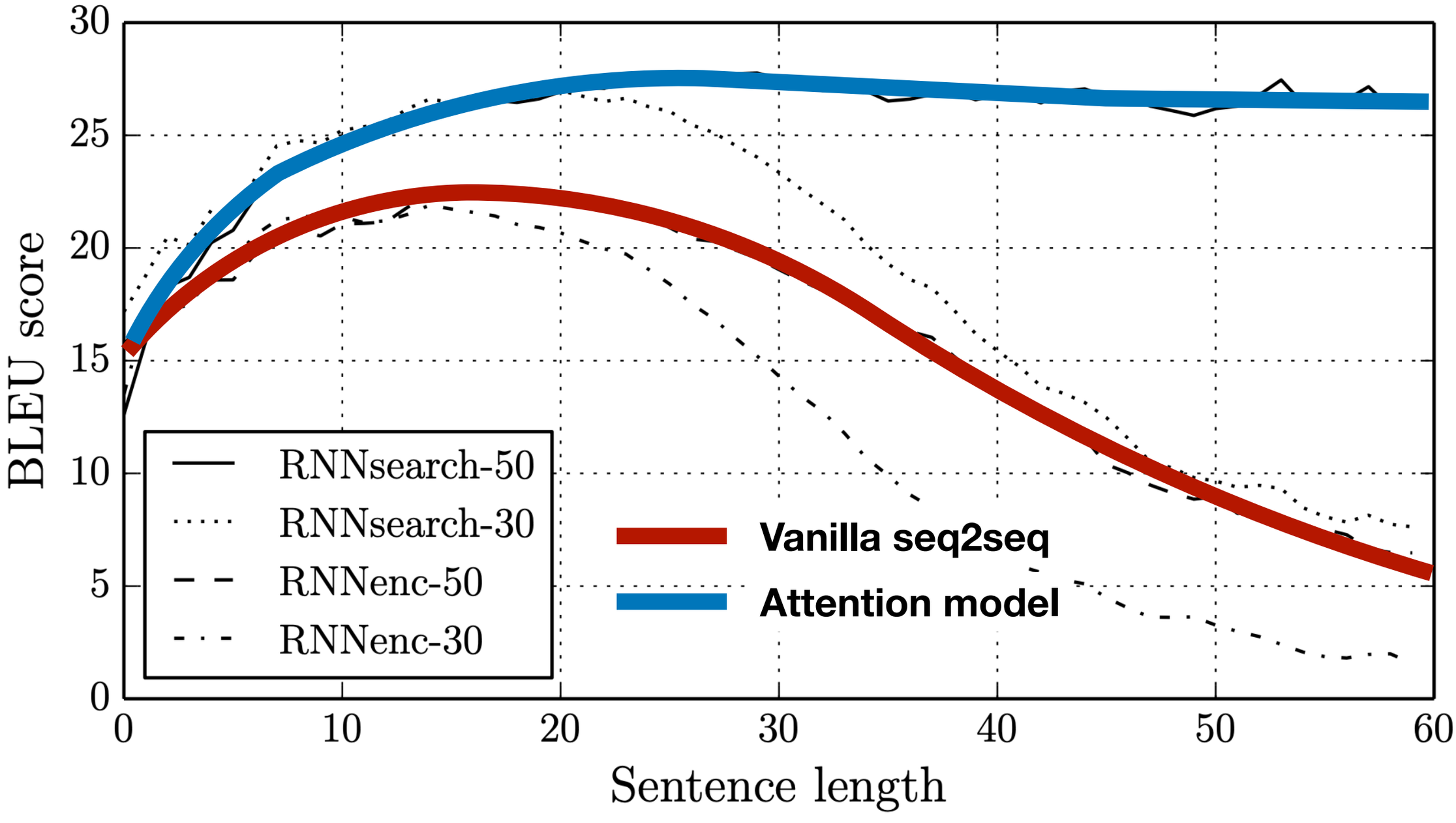


Figure from Bahdanau *et al*. (2014) *Neural machine translation by jointly learning to align and translate*

Applications: captioning

Figure 4. Examples of attending to the correct object (*white* indicates the attended regions, *underlines* indicated the corresponding word)

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.

Figure from Xu *et al*. (2015) *Show, Attend and Tell: Neural Image Caption Generation with Visual Attention*

Applications: QA

by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

ent119 identifies deceased sailor as **X** , who leaves behind a wife

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

Figure from Hermann *et al*. (2015) *Teaching machines to read and comprehend*

Applications: QA

On interpretability:
Jain and Wallace (2019)
Attention is not Explanation
Wiegreffe and Pinter (2019)
Attention is not not Explanation

by *ent423* , *ent261* correspondent updated 9:49 pm et , thu march 19 , 2015 (*ent261*) a *ent114* was killed in a parachute accident in *ent45* , *ent85* , near *ent312* , a *ent119* official told *ent261* on wednesday . he was identified thursday as special warfare operator 3rd class *ent23* , 29 , of *ent187* , *ent265* . `` *ent23* distinguished himself consistently throughout his career . he was the epitome of the quiet professional in all facets of his life , and he leaves an inspiring legacy of natural tenacity and focused

. . .

ent119 identifies deceased sailor as **X** , who leaves behind a wife

by *ent270* , *ent223* updated 9:35 am et , mon march 2 , 2015 (*ent223*) *ent63* went familial for fall at its fashion show in *ent231* on sunday , dedicating its collection to `` mamma '' with nary a pair of `` mom jeans '' in sight . *ent164* and *ent21* , who are behind the *ent196* brand , sent models down the runway in decidedly feminine dresses and skirts adorned with roses , lace and even embroidered doodles by the designers ' own nieces and nephews . many of the looks featured saccharine needlework phrases like `` i love you ,

. . .

X dedicated their fall fashion show to moms

Figure from Hermann *et al*. (2015) *Teaching machines to read and comprehend*

Applications: summarization

Input: Article 1st sentence	Model-written headline
metro-goldwyn-mayer reported a third-quarter net loss of dlrs 16 million due mainly to the effect of accounting rules adopted this year	mgm reports 16 million net loss on higher revenue
starting from july 1, the island province of hainan in southern china will implement strict market access control on all incoming livestock and animal products to prevent the possible spread of epidemic diseases	hainan to curb spread of diseases
australian wine exports hit a record 52.1 million liters worth 260 million dollars (143 million us) in september, the government statistics office reported on monday	australian wine exports hit record high in september

Figure from Liu and Pan (2016) Text summarization with TensorFlow

Summary: attention

Vanilla encoder-decoder architectures have a **limited memory bottleneck**

Attention allows all **encoder states** to serve as **memory** that the decoder can selectively access

Attention weights are calculated as a function of encoder and decoder state alignment (similarity) and select what part of the memory to access

Attention mechanisms have been applied to a variety of models, providing advances in the state of the art in e.g. machine translation

Resources

Chris Manning: Neural Machine Translation and Models with Attention | Stanford CS 224N [esp. 45:00-]
https://www.youtube.com/watch?v=IxQtK2SjWWM

Andrew Ng: C5W3L07 Attention Model Intuition [and follow-up]
https://www.youtube.com/watch?v=SysgYptB198

Weng (2018) Attention? Attention!
https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Sutskever et al. (2014) Sequence to Sequence Learning with Neural Networks
https://arxiv.org/abs/1409.3215

Bahdanau *et al*. (2014) Neural machine translation by jointly learning to align and translate
https://arxiv.org/abs/1409.0473