ATTENTIVE LANGUAGE MODELS

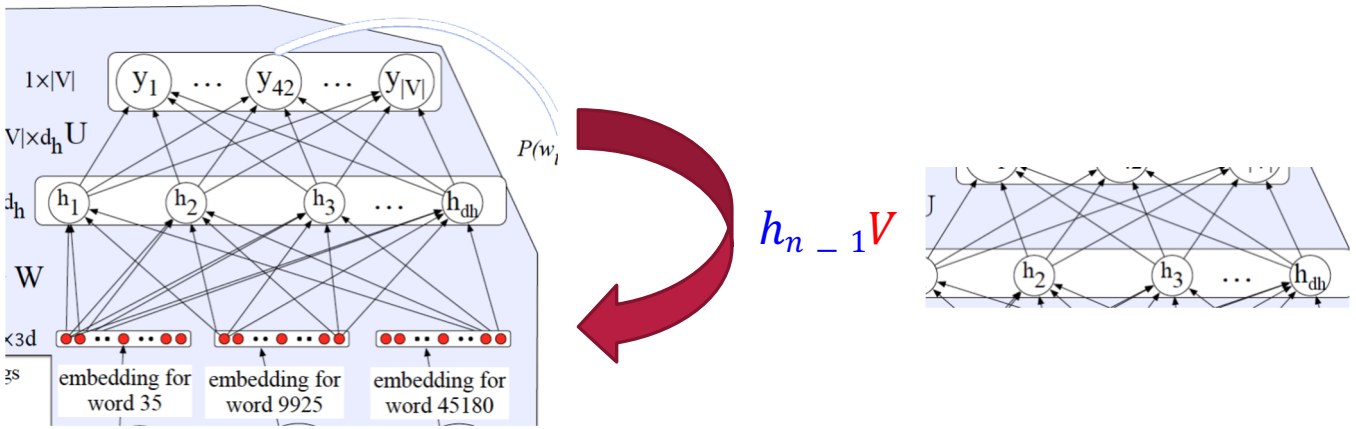
Recurrent Neural LM

$$Y_n = softmax(hnU)$$

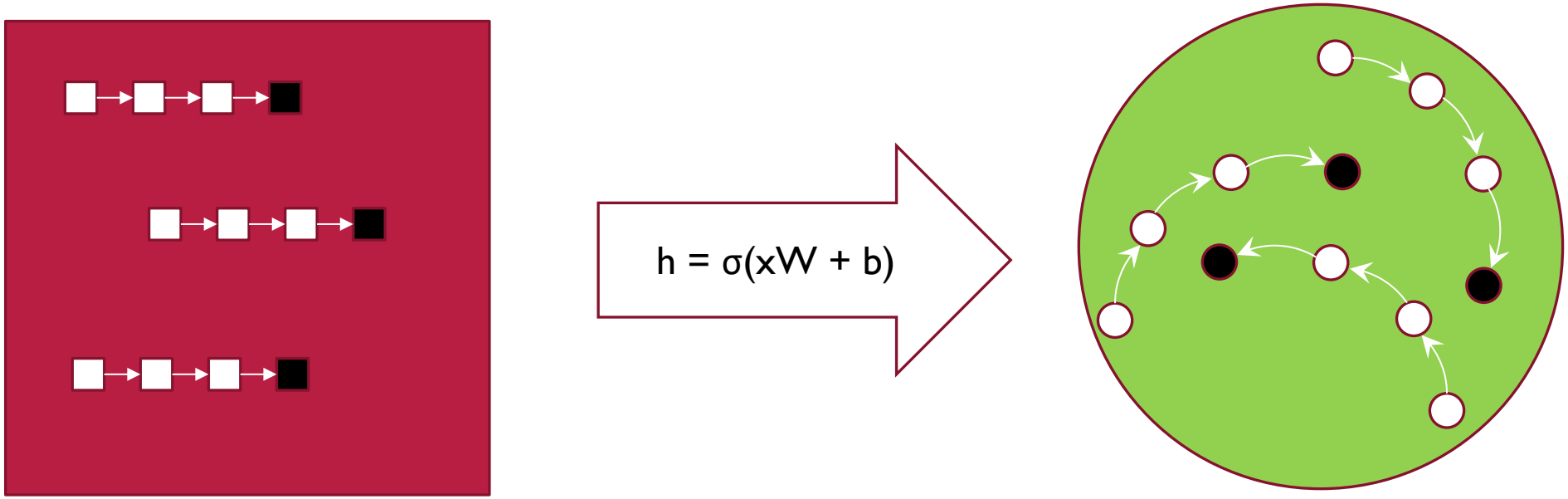
$$\mathrm{h_n} = \tanh(Xn\,W + b + h_{n\_1}V)$$

$$X_n\,W + b$$

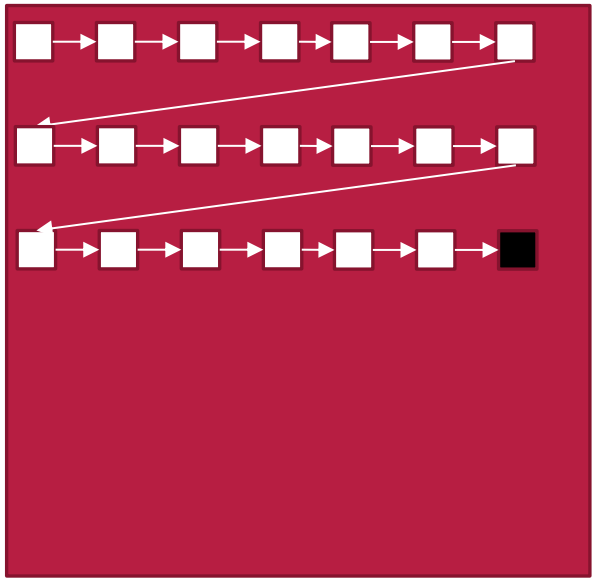
$$X_n$$



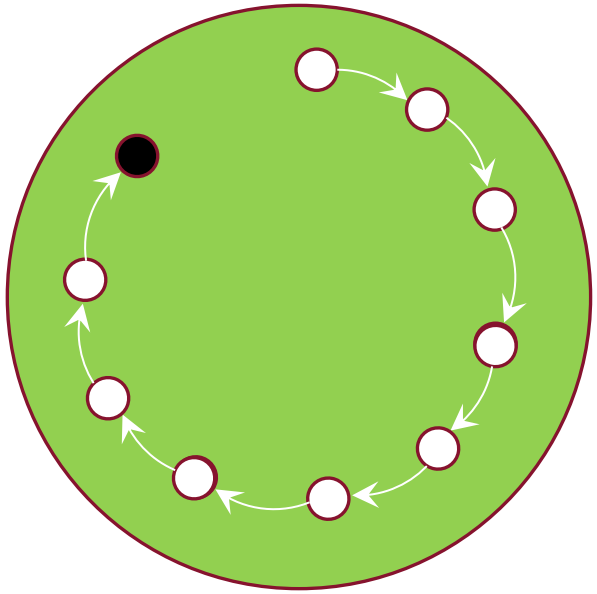
Neural LM



Recurrent Neural LM



$$h_n = \sigma(xW + b + h_{n-1}V)$$



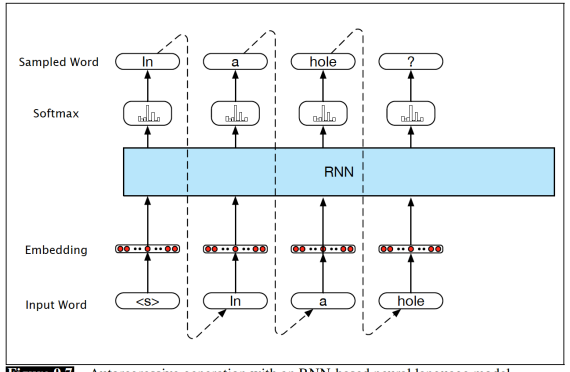
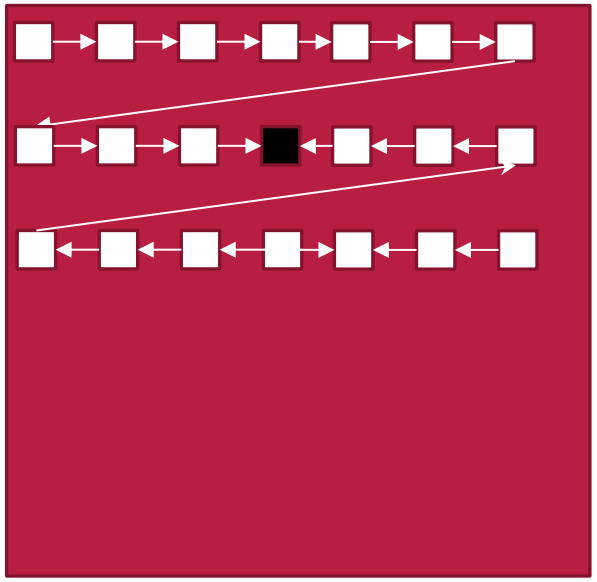
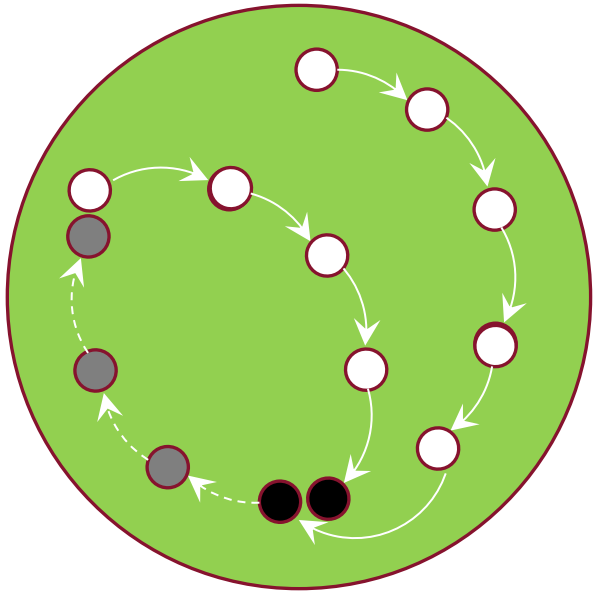


Figure 9.7 Autoregressive generation with an RNN-based neural language model.

Recurrent Neural LM: Bidirection



$$h_i = \sigma(xW + b + h_{i-1}V)$$
$$h_i = \sigma(xW' + b' + h_{i+1}V)$$



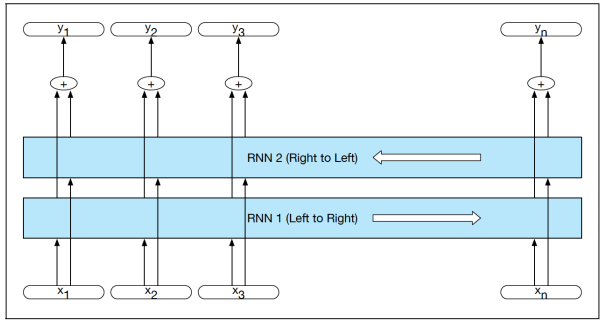
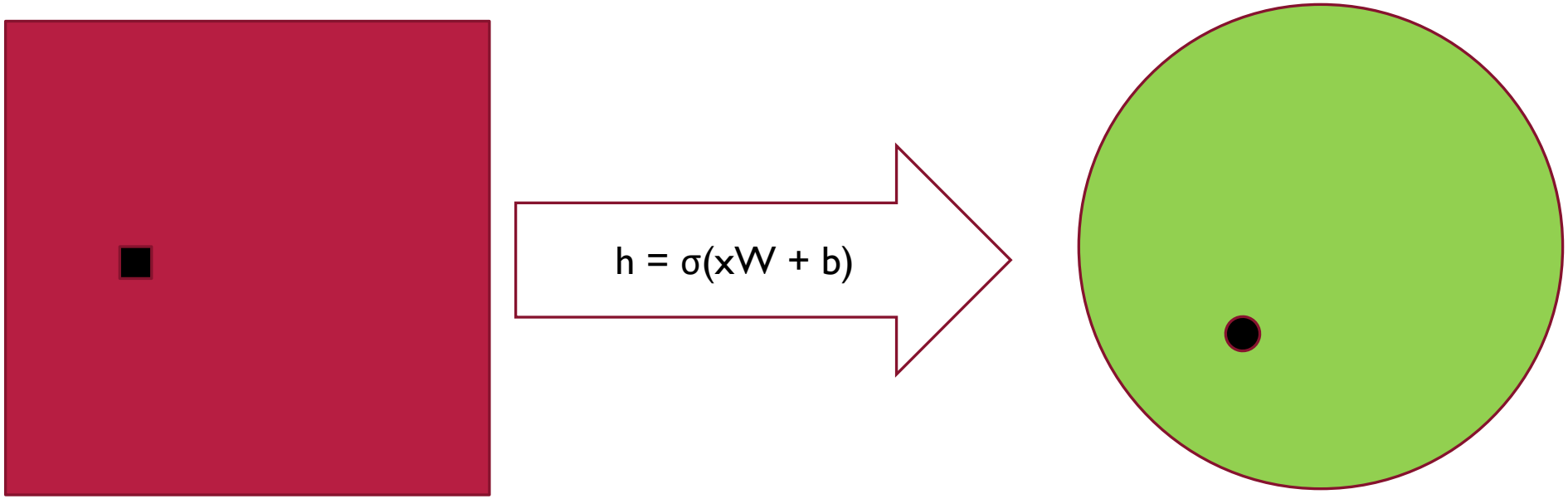


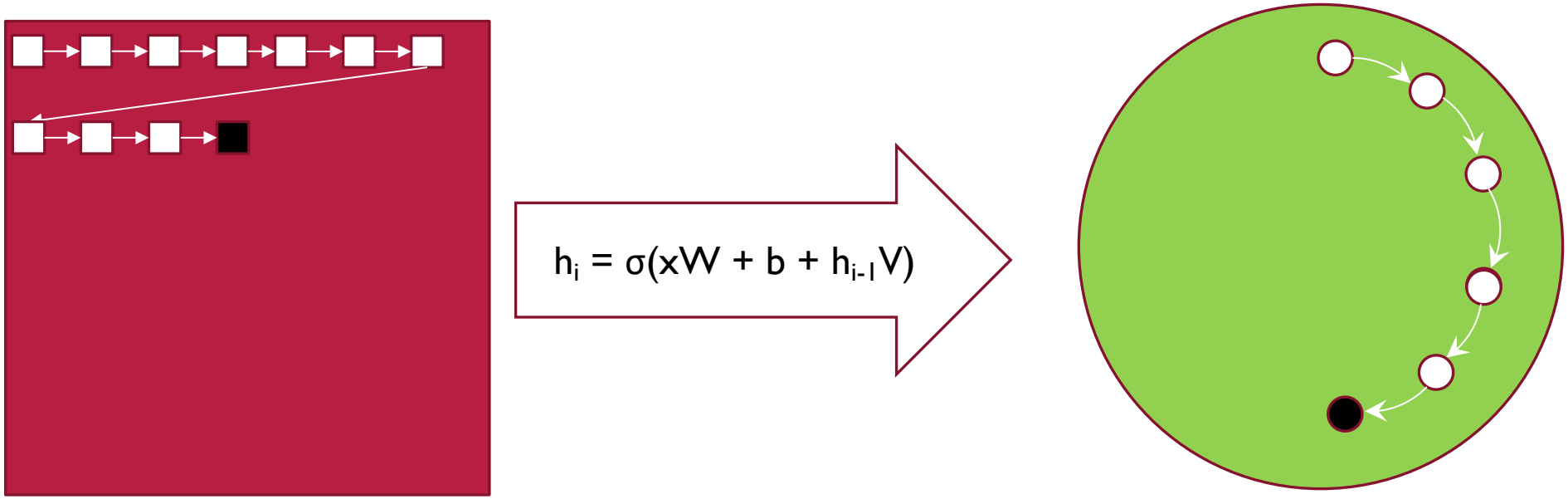
Figure 9.11 A bidirectional RNN. Separate models are trained in the forward and backward directions with the output of each model at each time point concatenated to represent the state of affairs at that point in time. The box wrapped around the forward and backward network emphasizes the modular nature of this architecture.

Encoder-Decoder

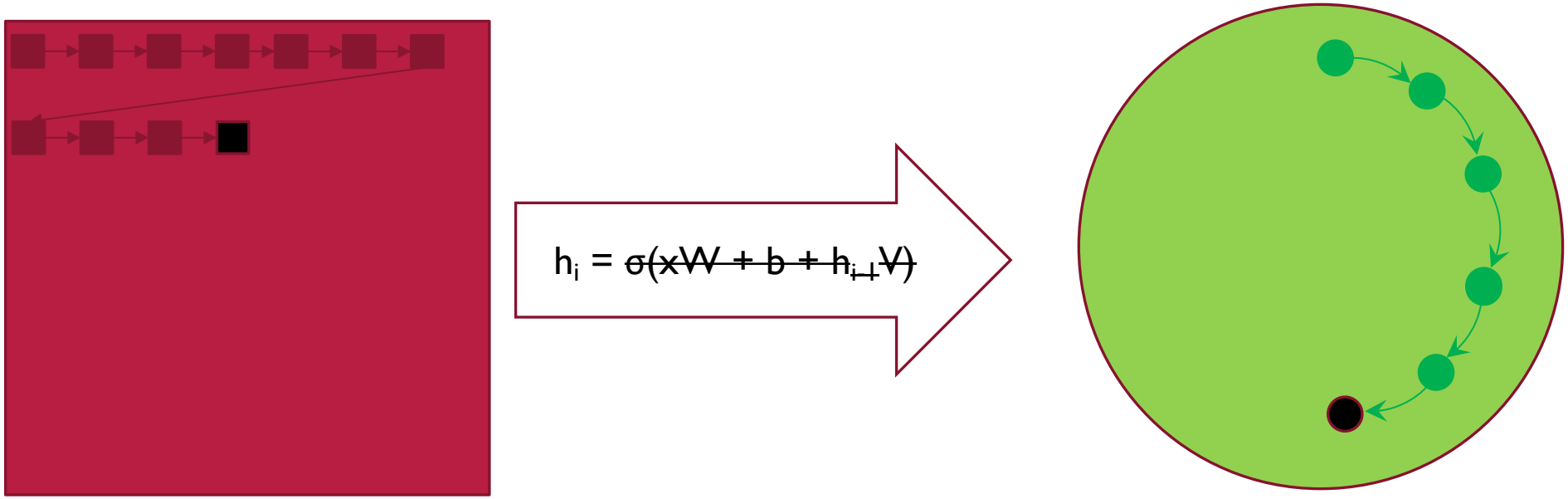
Recurrent Neural LM: Encoder-Decoder



Recurrent Neural LM: Encoder-Decoder

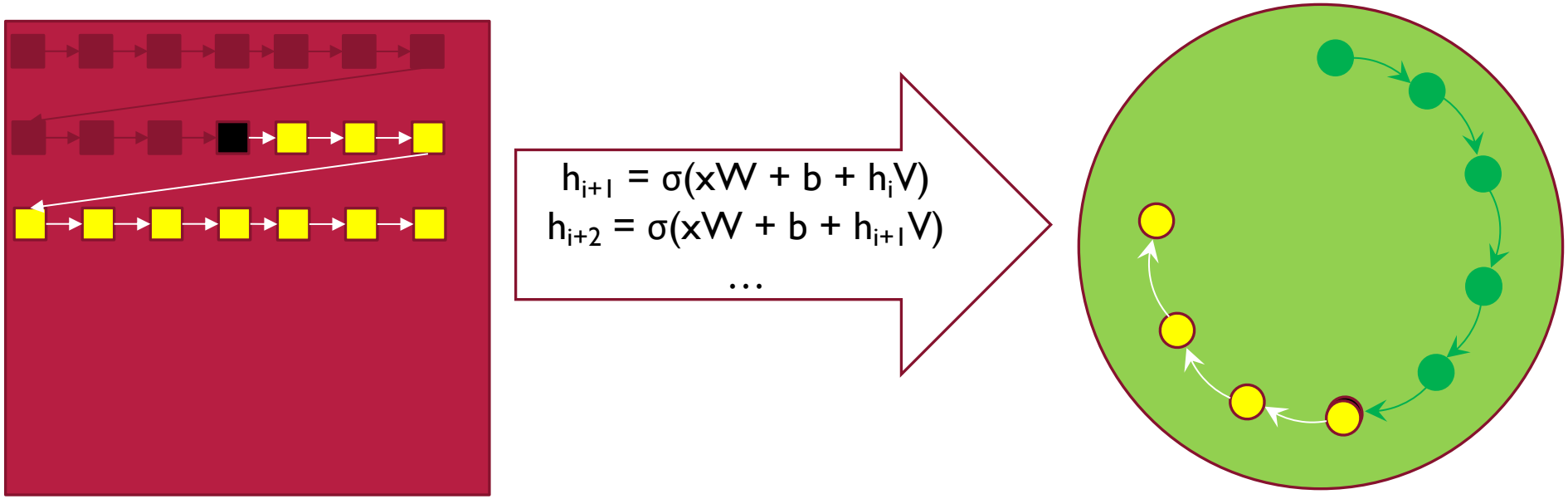


Recurrent Neural LM: Encoder-Decoder



The context helps x to reach its position. Then, we don't need them anymore!

Recurrent Neural LM: Encoder-Decoder



From the final position of x, we can continue and generate other tokens (autogeneration)

Recurrent Neural LM: Encoder-Decoder

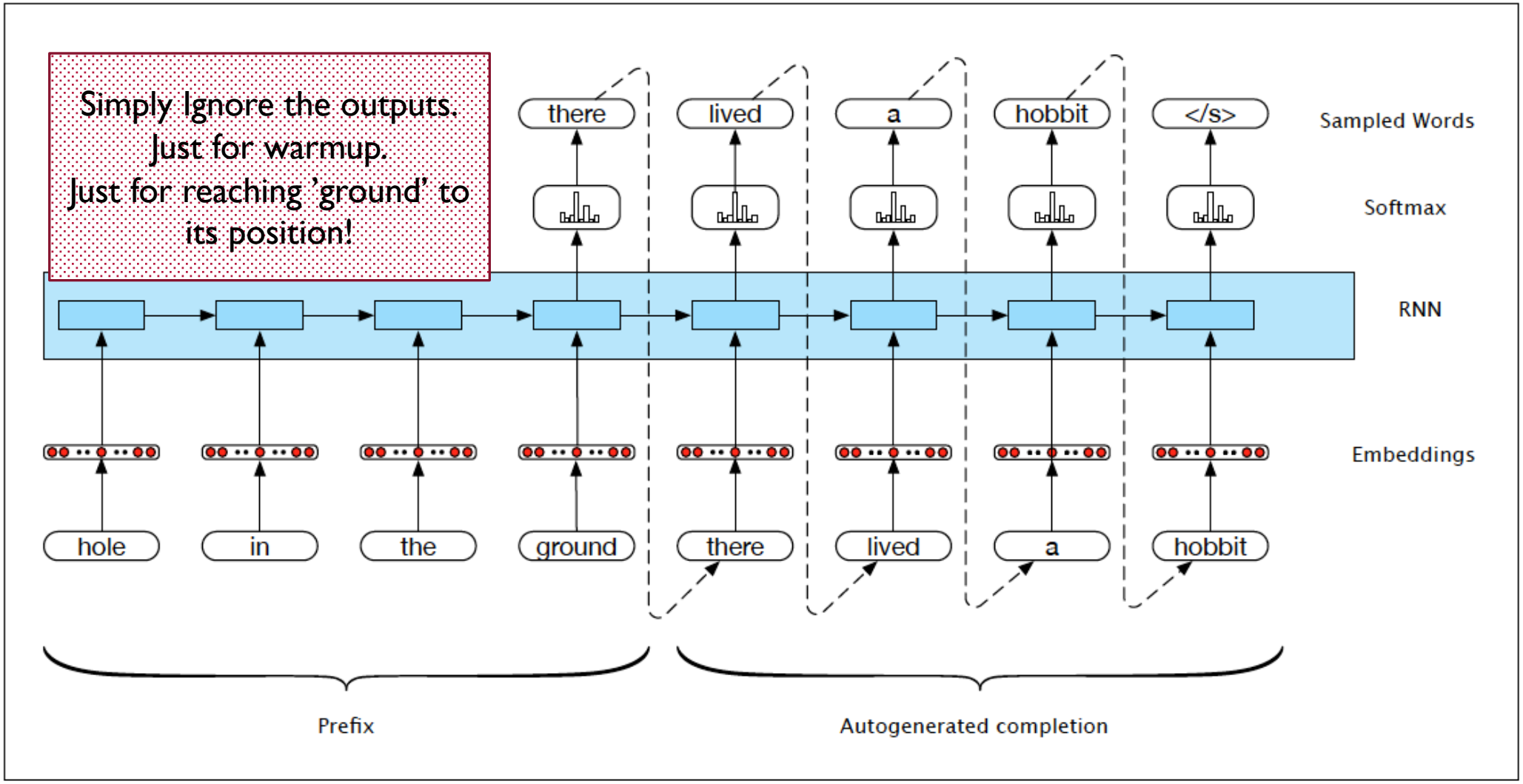
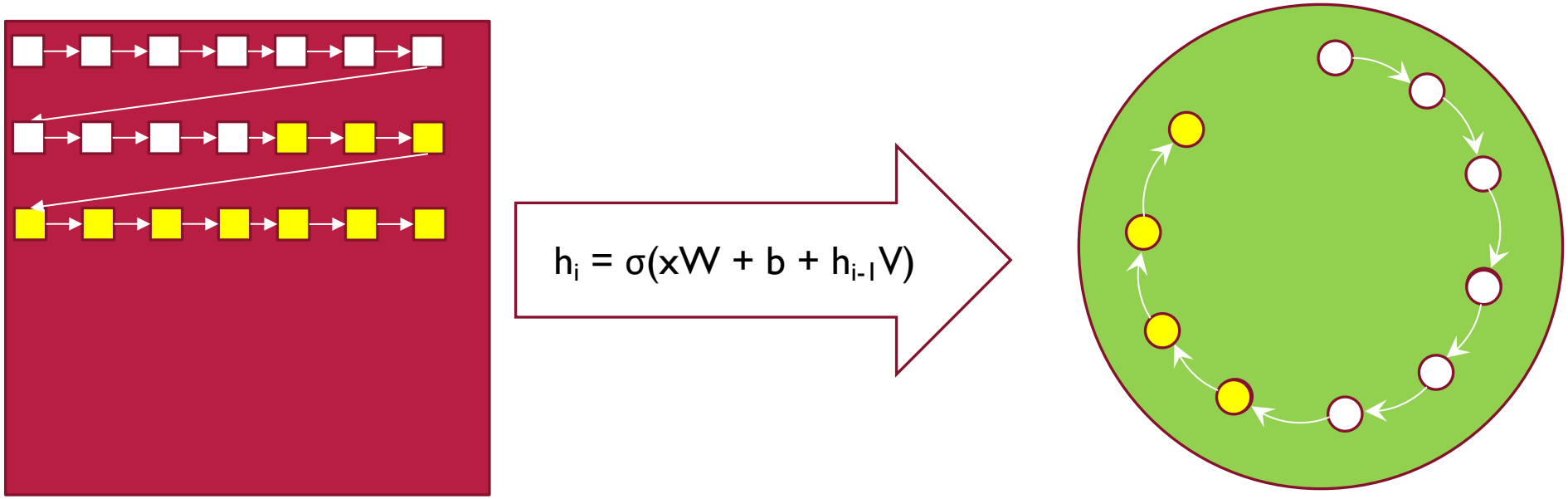


Figure 10.1 Using an RNN to generate the completion of an input phrase.

Recurrent Neural LM: Bitext Translation

For machine, they are just sequences of tokens!



<s>Prague Stock Market falls to minus by the end of the trading day</s><s>Die Prager Börse stürzt gegen Geschäftsschluss ins Minus</s>

Recurrent Neural LM: Bitext Translation

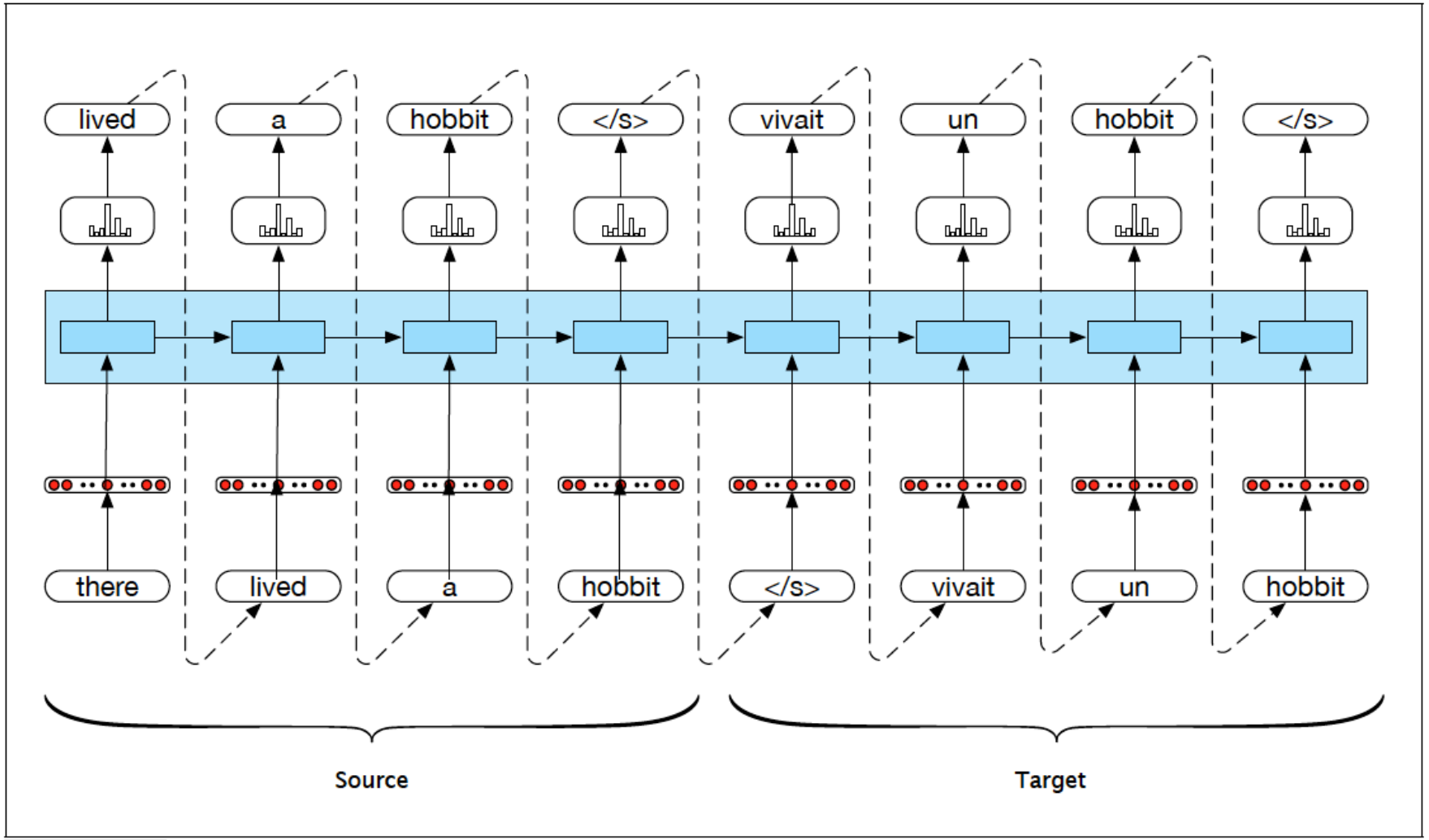
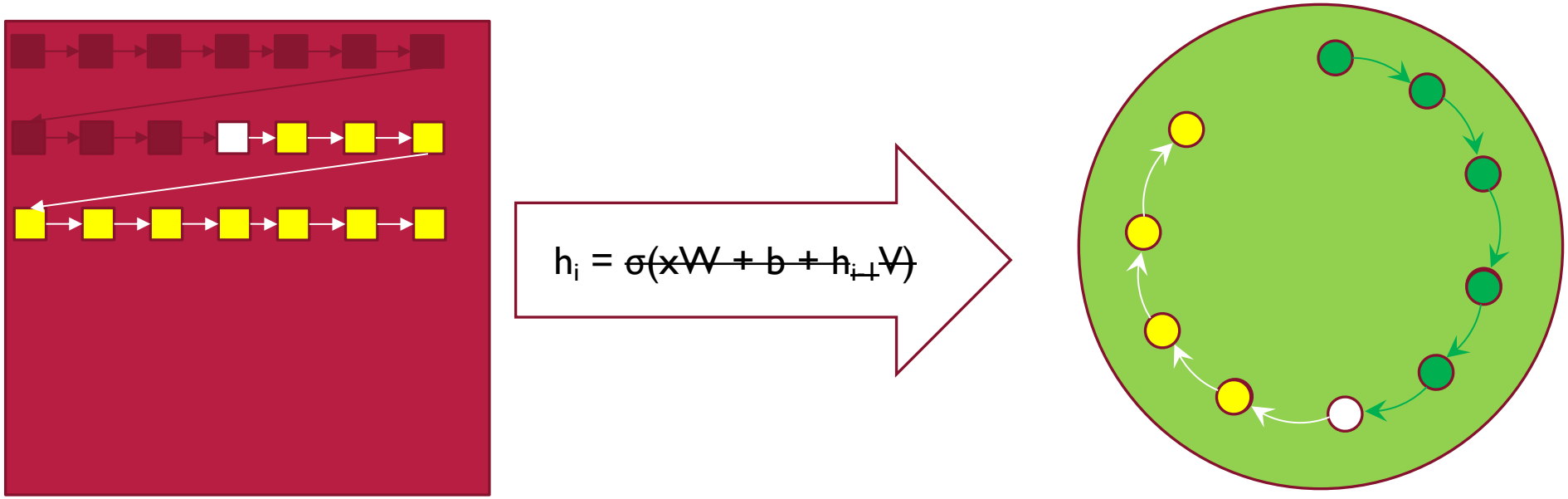


Figure 10.2 Training setup for a neural language model approach to machine translation. Source-target bi-texts are concatenated and used to train a language model.

Recurrent Neural LM: Bitext Translation

For machine, they are just sequences of tokens!



<s>Prague Stock Market falls to minus by the end of the trading day</s><s>Die Prager Börse stürzt gegen Geschäftsschluss ins Minus</s>

Recurrent Neural LM: Encoder-Decoder

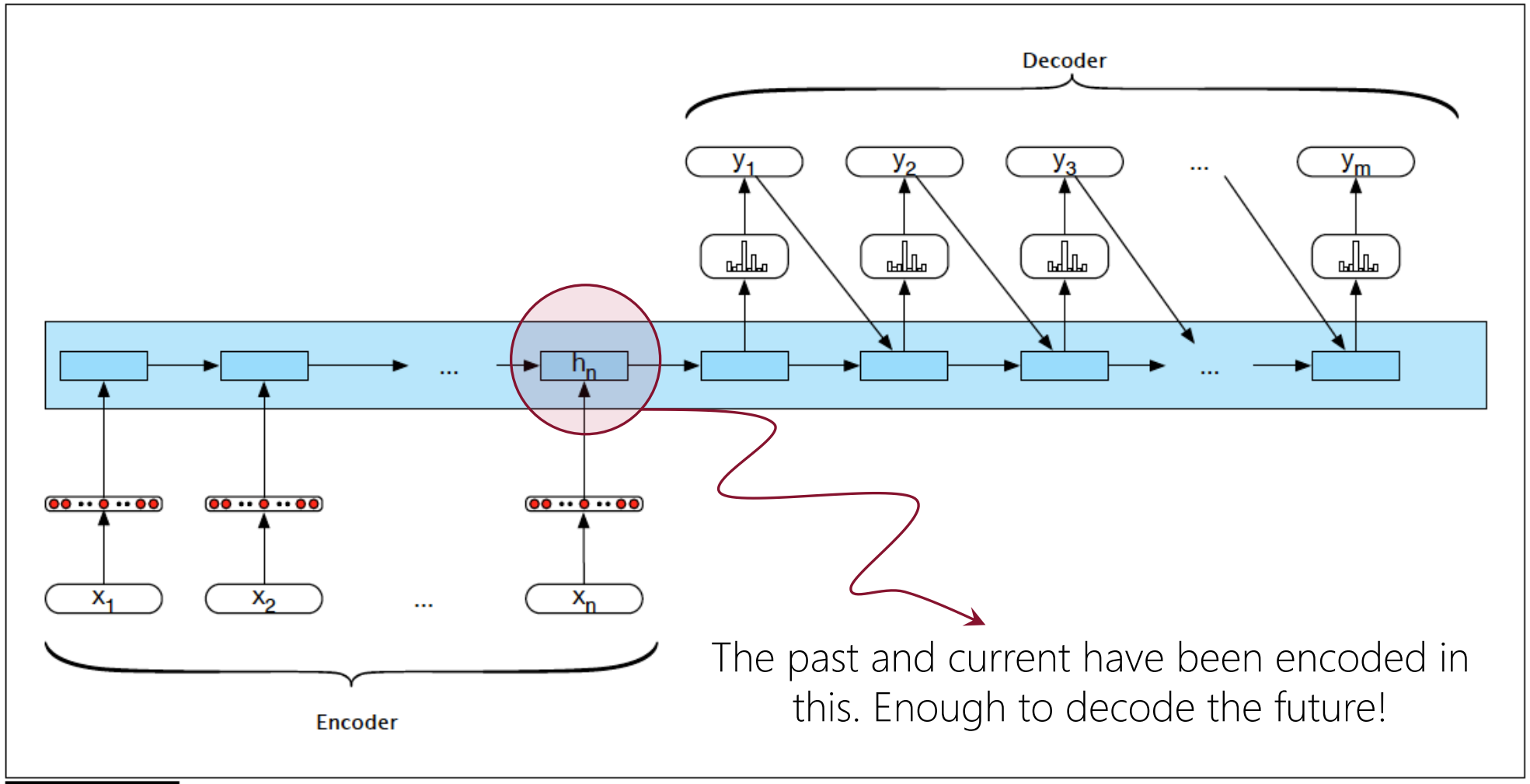


Figure 10.3 Basic RNN-based encoder-decoder architecture. The final hidden state of the encoder RNN serves as the context for the decoder in its role as h_0 in the decoder RNN.

Recurrent Neural LM: Encoder-Decoder

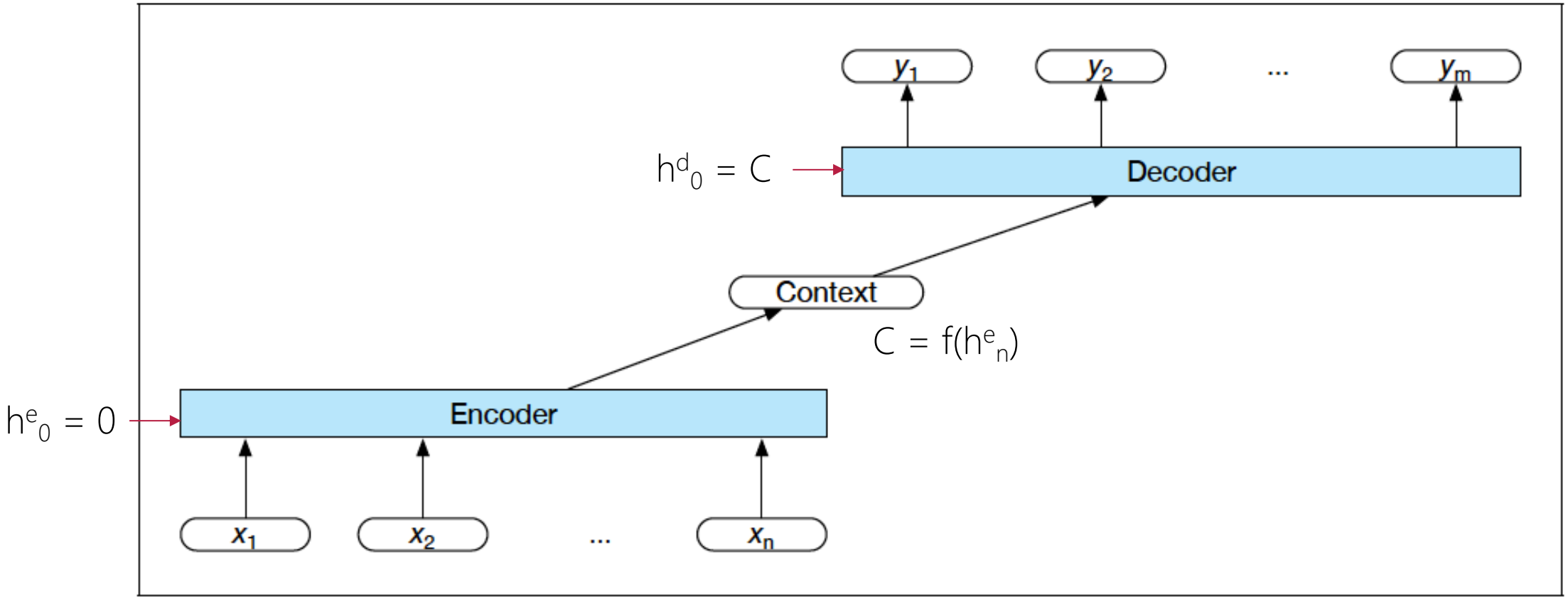


Figure 10.4 Basic architecture for an abstract encoder-decoder network. The context is a function of the vector of contextualized input representations and may be used by the decoder in a variety of ways.

Sequence2Sequence

Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to sequence learning with neural networks. In NIPS.

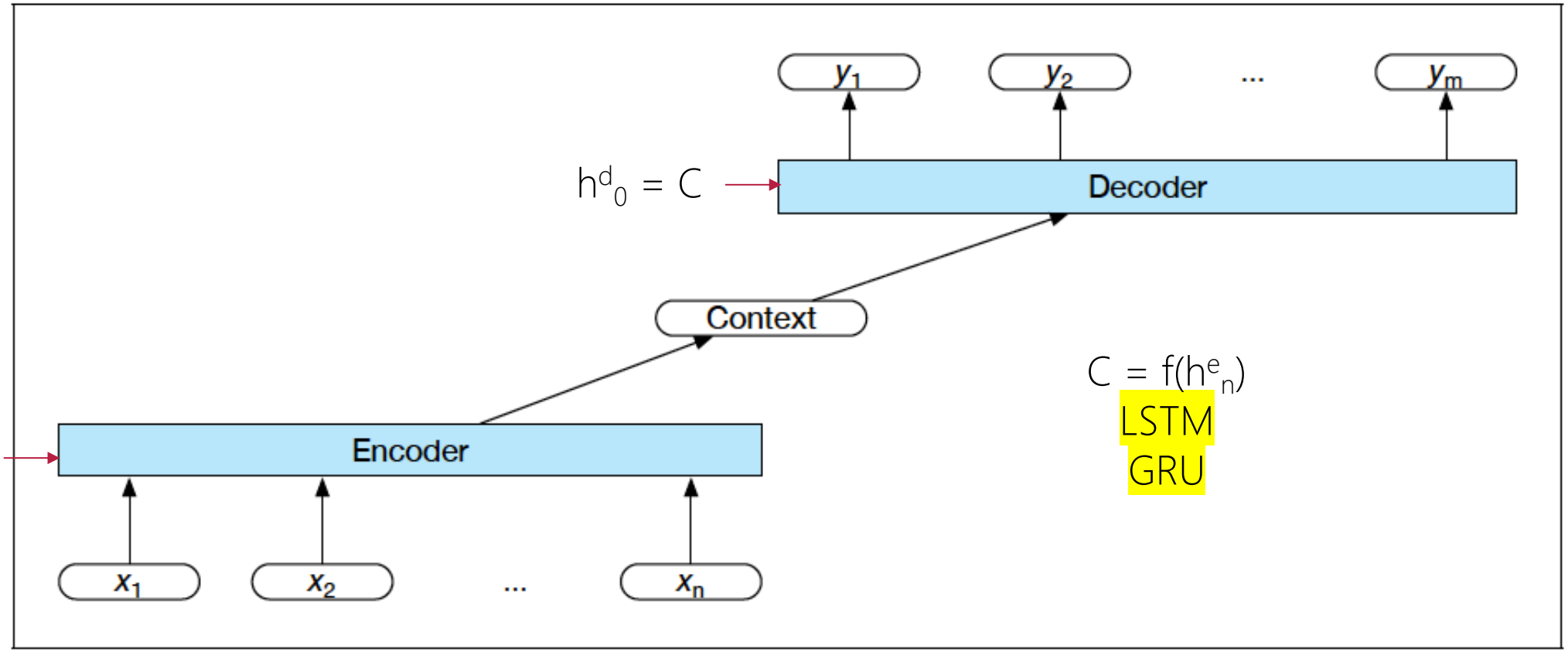


Figure 10.4 Basic architecture for an abstract encoder-decoder network. The context is a function of the vector of contextualized input representations and may be used by the decoder in a variety of ways.

Question to Answers

Anything to Sequence

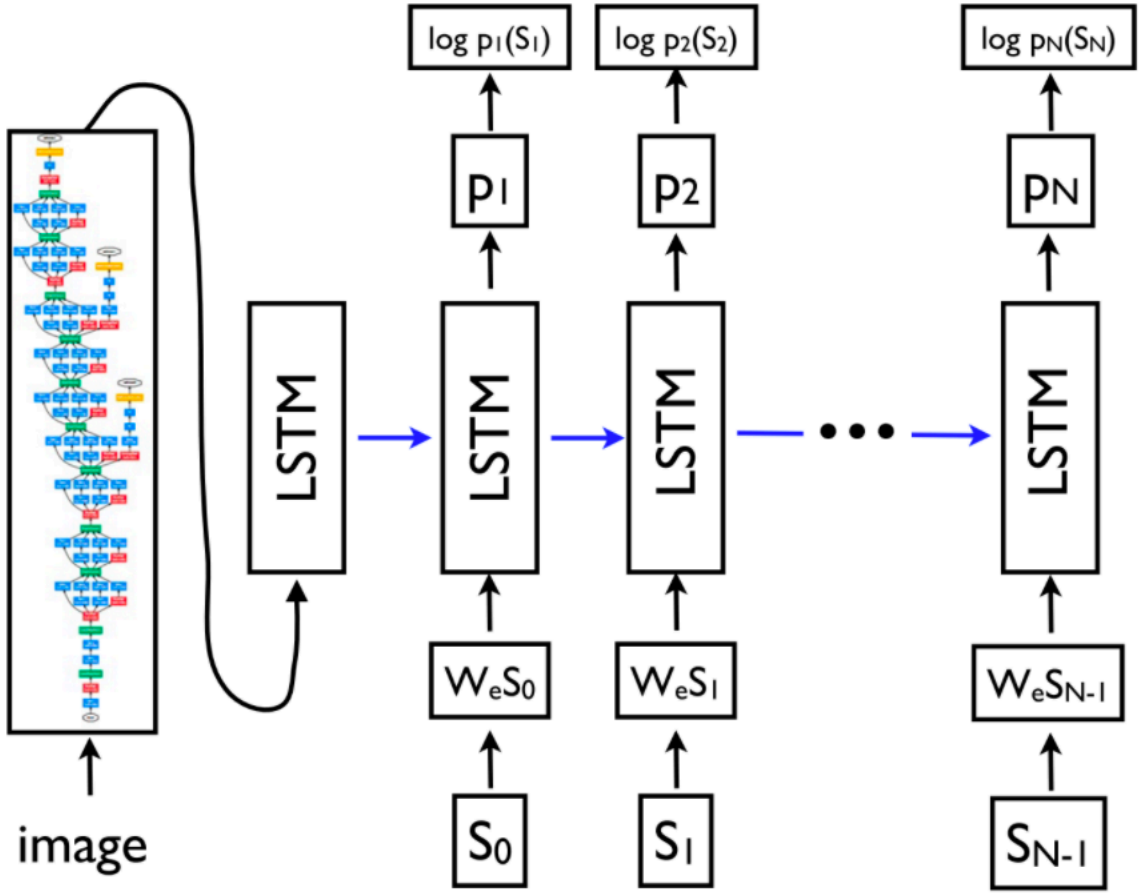


Figure 9: Generating a caption based on an image (Vinyals et al., 2015)

Look into the Past based on future!

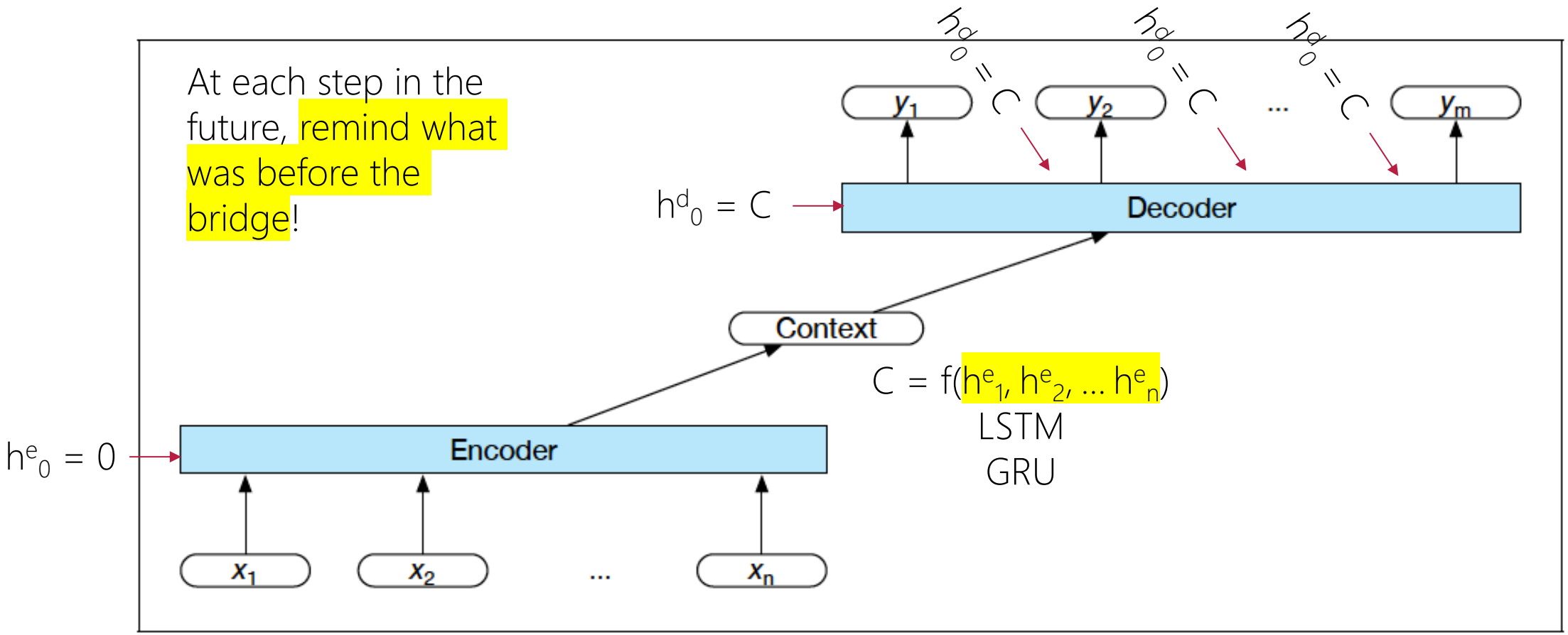
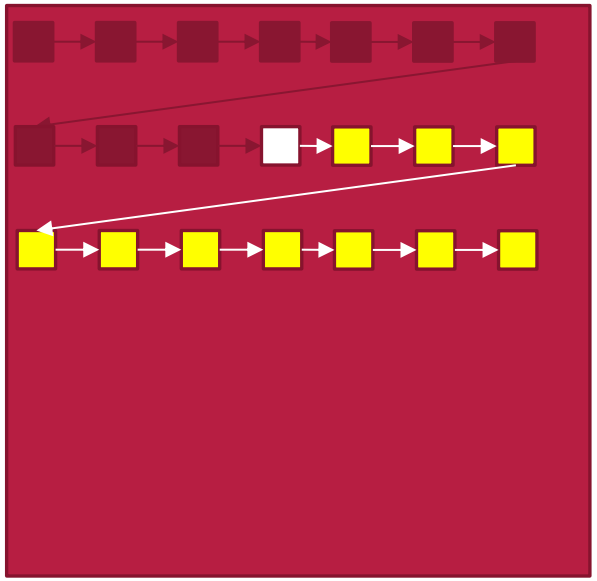


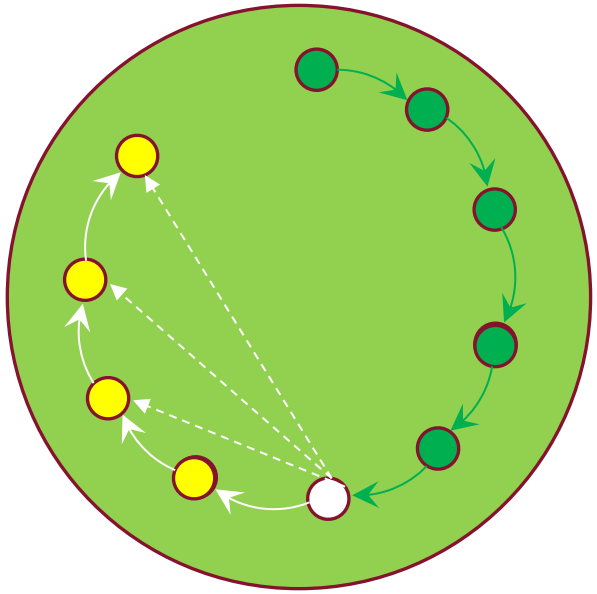
Figure 10.4 Basic architecture for an abstract encoder-decoder network. The context is a function of the vector of contextualized input representations and may be used by the decoder in a variety of ways.

Recurrent Neural LM: Bitext Translation

For machine, they are just sequences of tokens!



$h_i = \sigma(xW + b + h_{i-1}V)$



<s>Prague Stock Market falls to minus by the end of the trading day</s><s>Die Prager Börse stürzt gegen Geschäftsschluss ins Minus</s>

Look into the Past based on future!

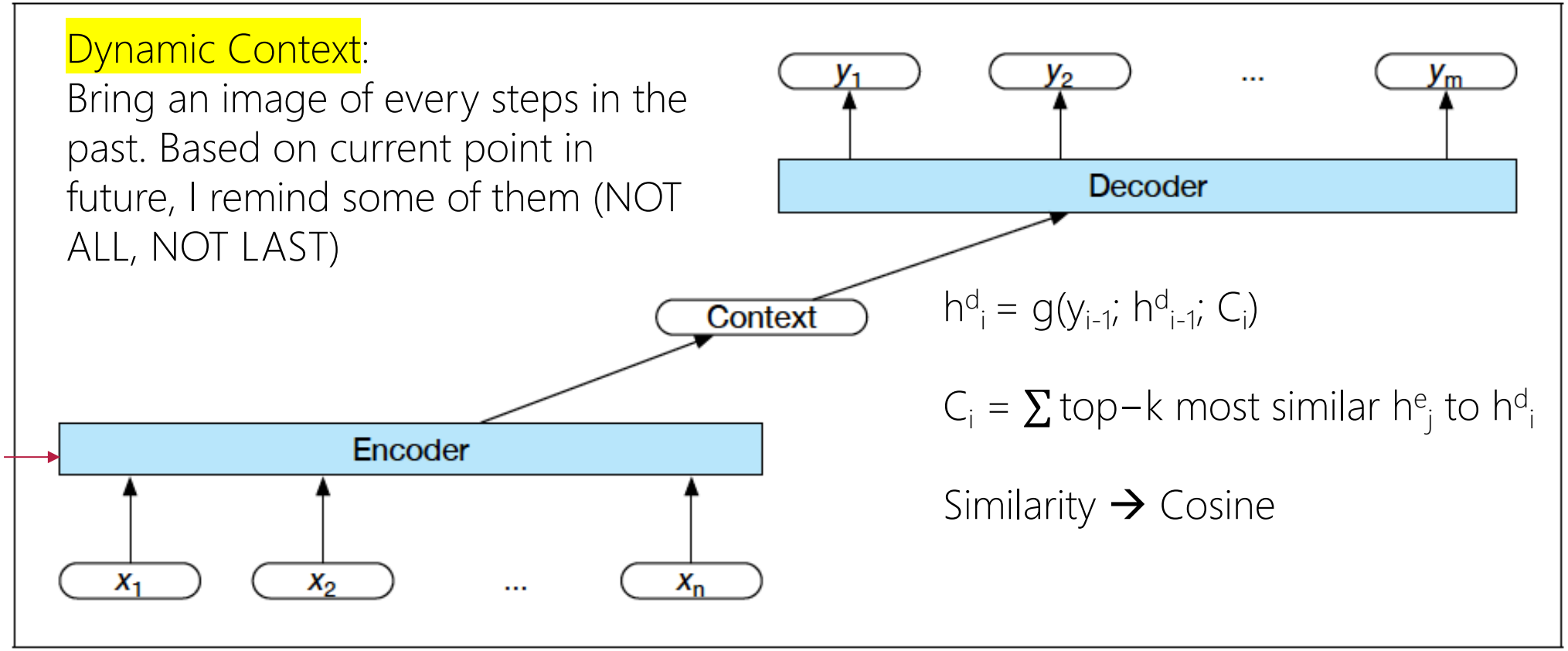


Figure 10.4 Basic architecture for an abstract encoder-decoder network. The context is a function of the vector of contextualized input representations and may be used by the decoder in a variety of ways.

Look into the Past based on future!

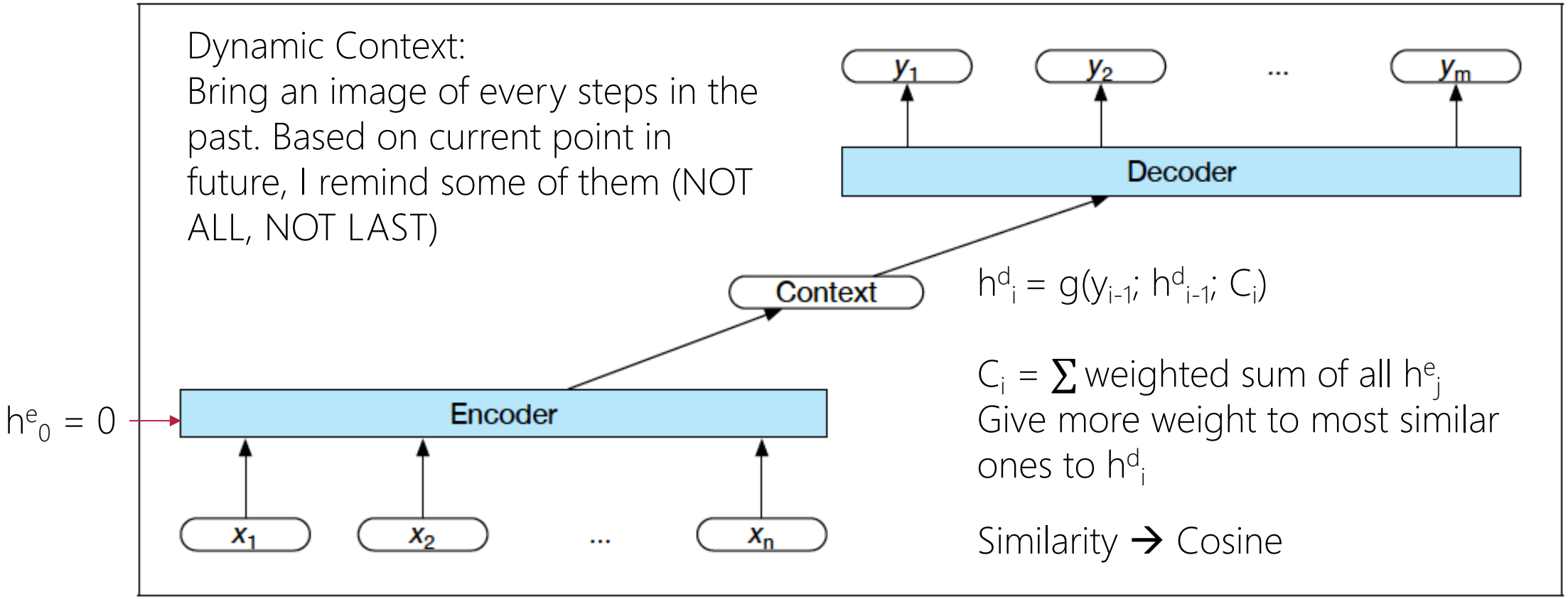
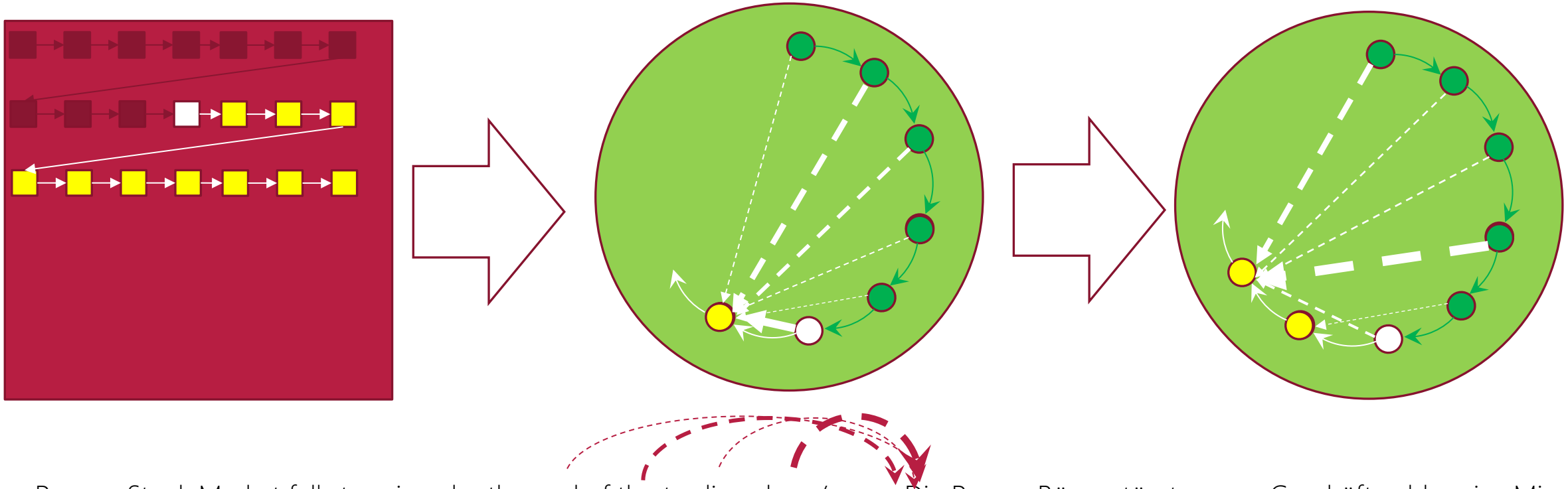


Figure 10.4 Basic architecture for an abstract encoder-decoder network. The context is a function of the vector of contextualized input representations and may be used by the decoder in a variety of ways.

Attention!

For machine, they are just sequences of tokens!



<s>Prague Stock Market falls to minus by the end of the trading day</s><s>Die Prager Börse stürzt gegen Geschäftsschluss ins Minus</s>

Attention!

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In ICLR 2015

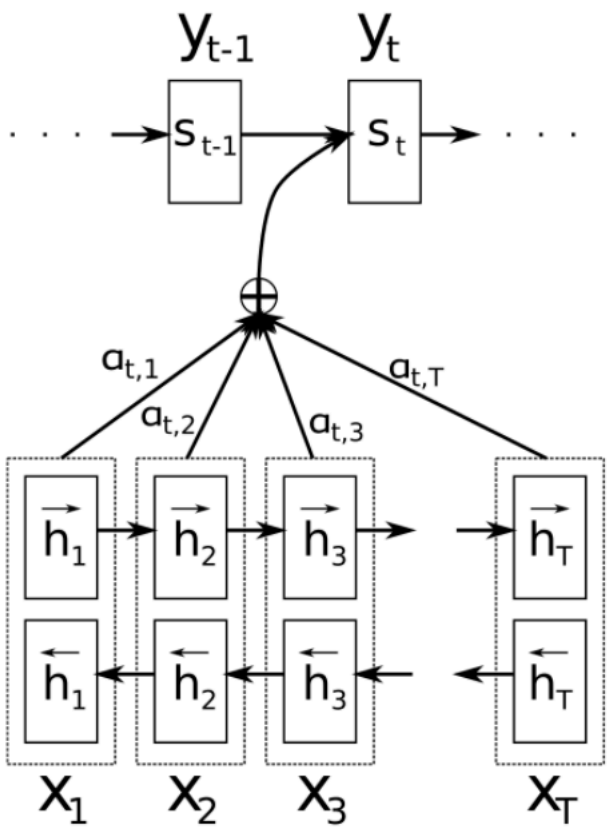


Figure 11: Attention (Bahdanau et al., 2015)

Similarity → Cosine → Learn Similarity as a functions of inputs!!

$score(h^d_{i-1}; h^e_j) = h^d_{i-1} W_s h^e_j$

$a_{ij} = softmax(score(h^d_{i-1}; h^e_j))$ for all j in encoder

$C_i = \sum_{j=1}^{n} a_{ij} h^e_j$

Attention!

A woman is throwing a frisbee in a park.

Figure 12: Visual attention in an image captioning model indicating what the model is attending to when generating the word "frisbee". (Xu et al., 2015)

Attention Is All You Need! The Transformer

BERT: Bidirectional Encoder Representations from Transformers