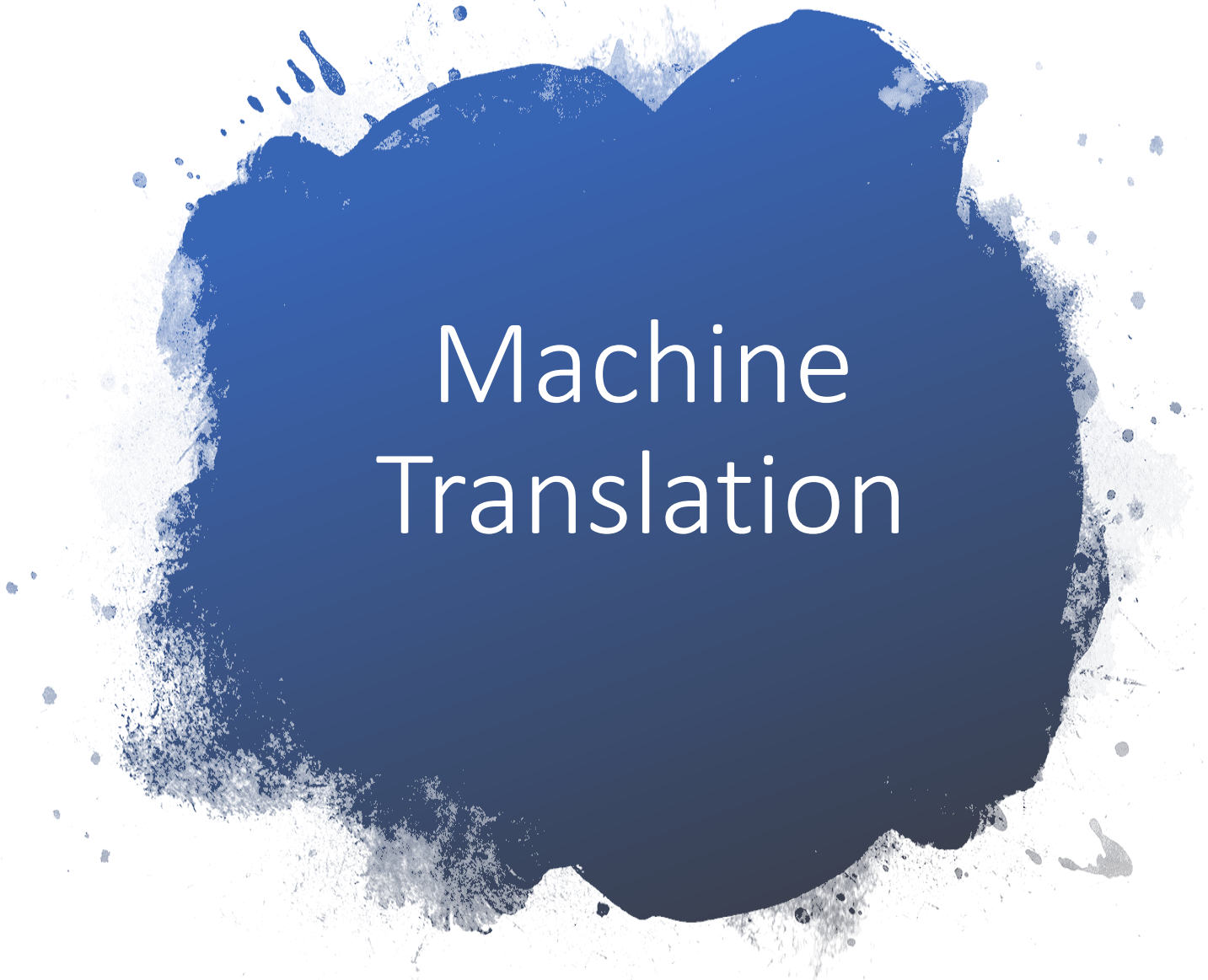
Procesamiento de Lenguaje Natural

Clase 14 – Seq2Seq

Ph.D. Rubén Manrique

rf.manrique@uniandes.edu.co

Maestría en Ingeniería de Sistemas y Computación

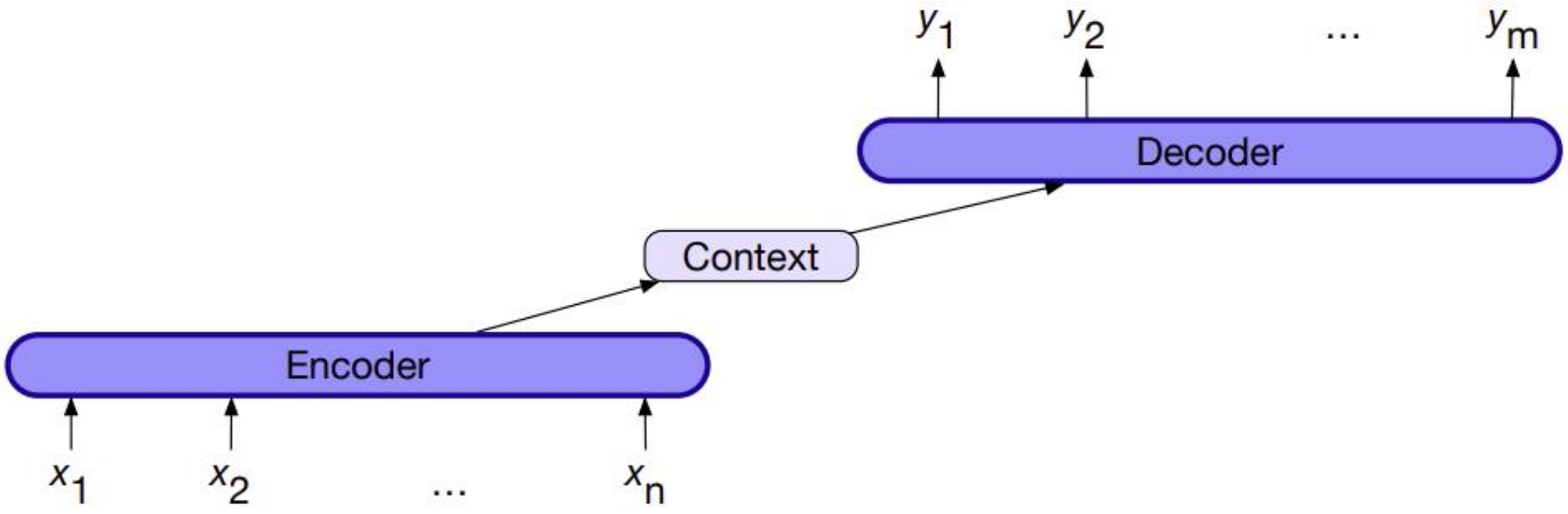

Machine
Translation

Encoder-decoder network (seq2seq)

- Sequence to sequence network, an architecture that can be implemented with RNNs or with Transformers.

- Machine translation is exactly such a task: the words of the target language don't necessarily agree with the words of the source language in number or order.
  - English: He wrote a letter to a friend
  - Japanese: tomodachi ni tegami-o kaita (friend to letter wrote)
  - Spanish: bruja verde.
  - English: green witch.

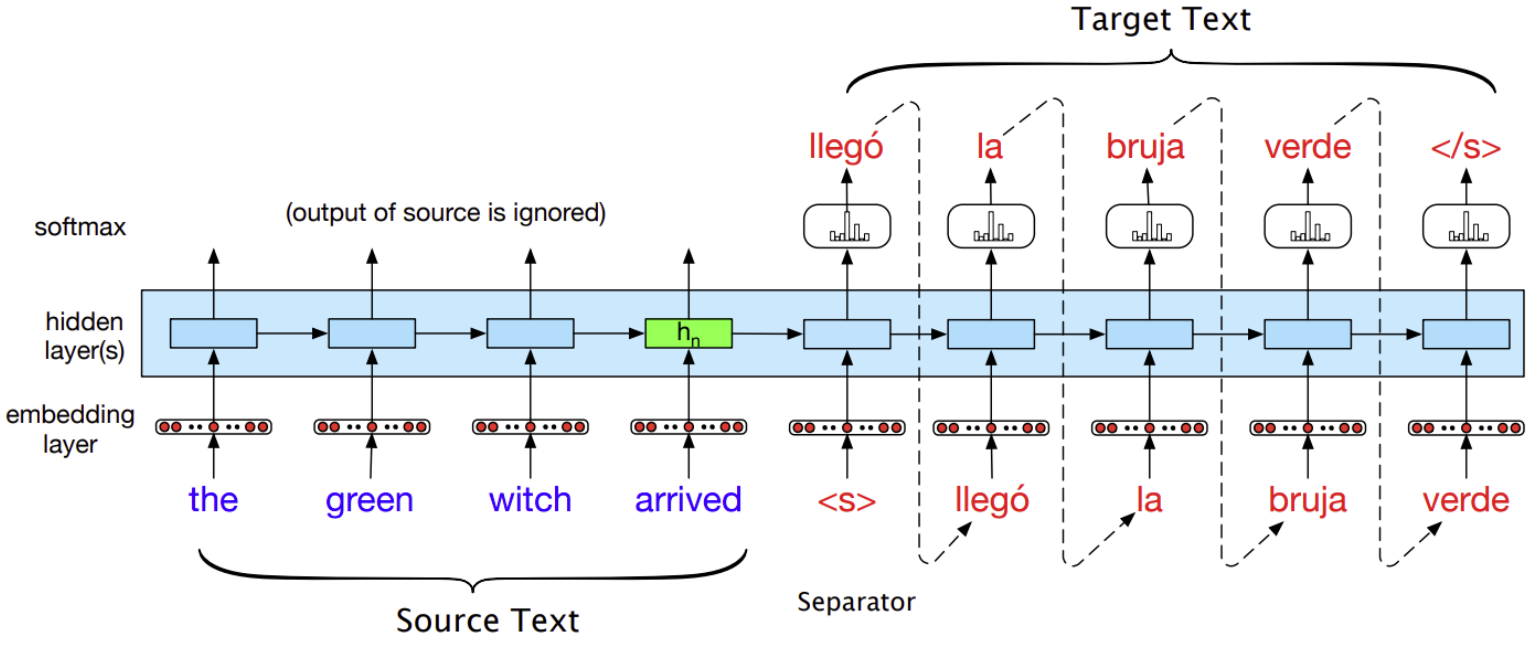
- **Lexical gap**: No word or phrase can express the exact meaning of a word in the other language



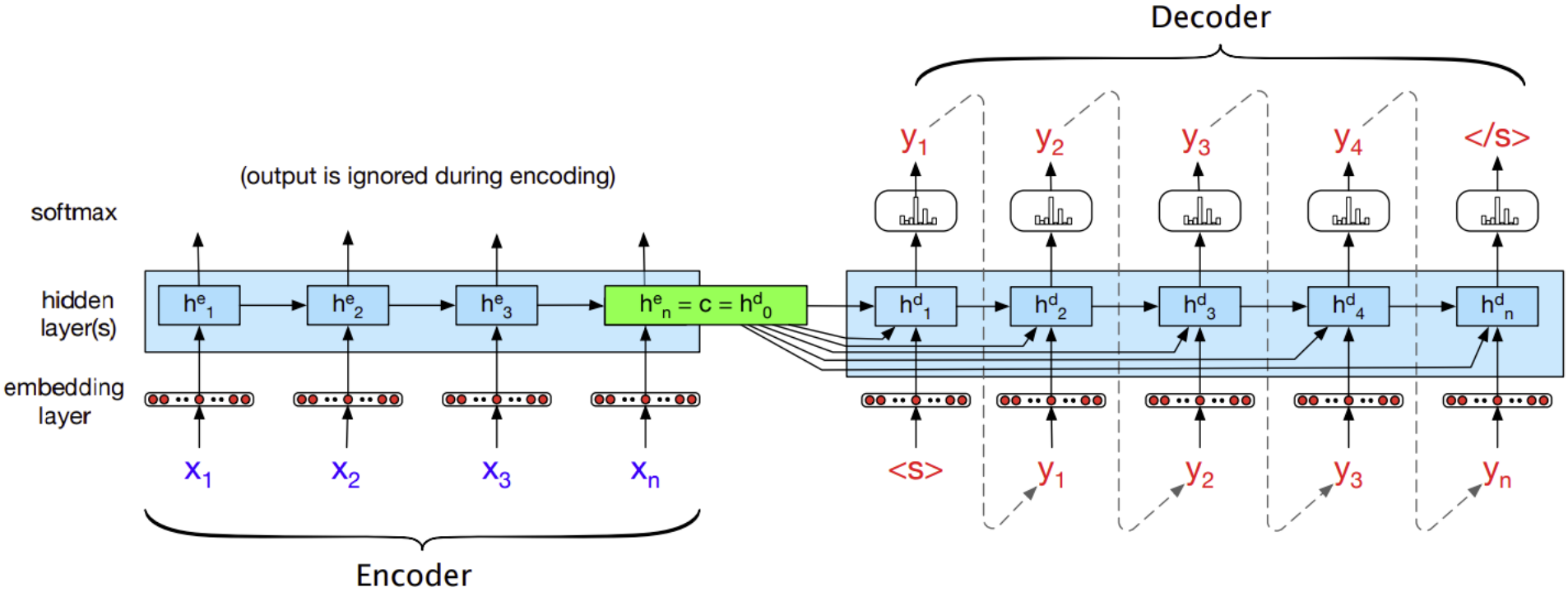
1. An encoder that accepts an input sequence, x_1^n, and generates a corresponding sequence of contextualized representations, h_1^n. LSTMs, GRUs, convolutional networks, and **transformers** can all be employed as encoders.
2. A context vector, c, which is a function of h_1^n, and conveys the essence of the input to the decoder.
3. A decoder, which accepts c as input and generates an arbitrary length sequence of hidden states h_1^m, from which a corresponding sequence of output states y_1^m, can be obtained. Just as with encoders, decoders can be realized by any kind of sequence architecture.

Encoder-Decoder with RNNs (I)

- We only have to make one slight change to turn this language model with **autoregressive generation** into a translation model:
  - Add an sentence separation marker at the end of the source text, and then simply concatenate the target text

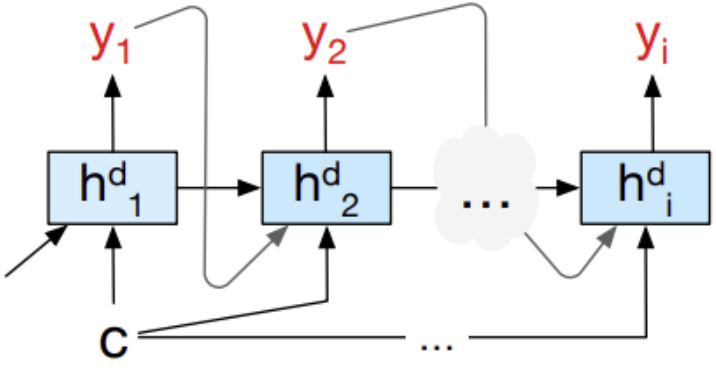


Encoder-Decoder with RNNs (II)



- Superscripts e and d distinguish the hidden states of the encoder and the decoder.

Encoder-Decoder with RNNs (III)



$$
\begin{aligned}
c &= h_n^e \\
h_0^d &= c \\
h_t^d &= g(\hat{y}_{t-1}, h_{t-1}^d, c) \\
z_t &= f(h_t^d) \\
y_t &= \text{softmax}(z_t)
\end{aligned}
$$

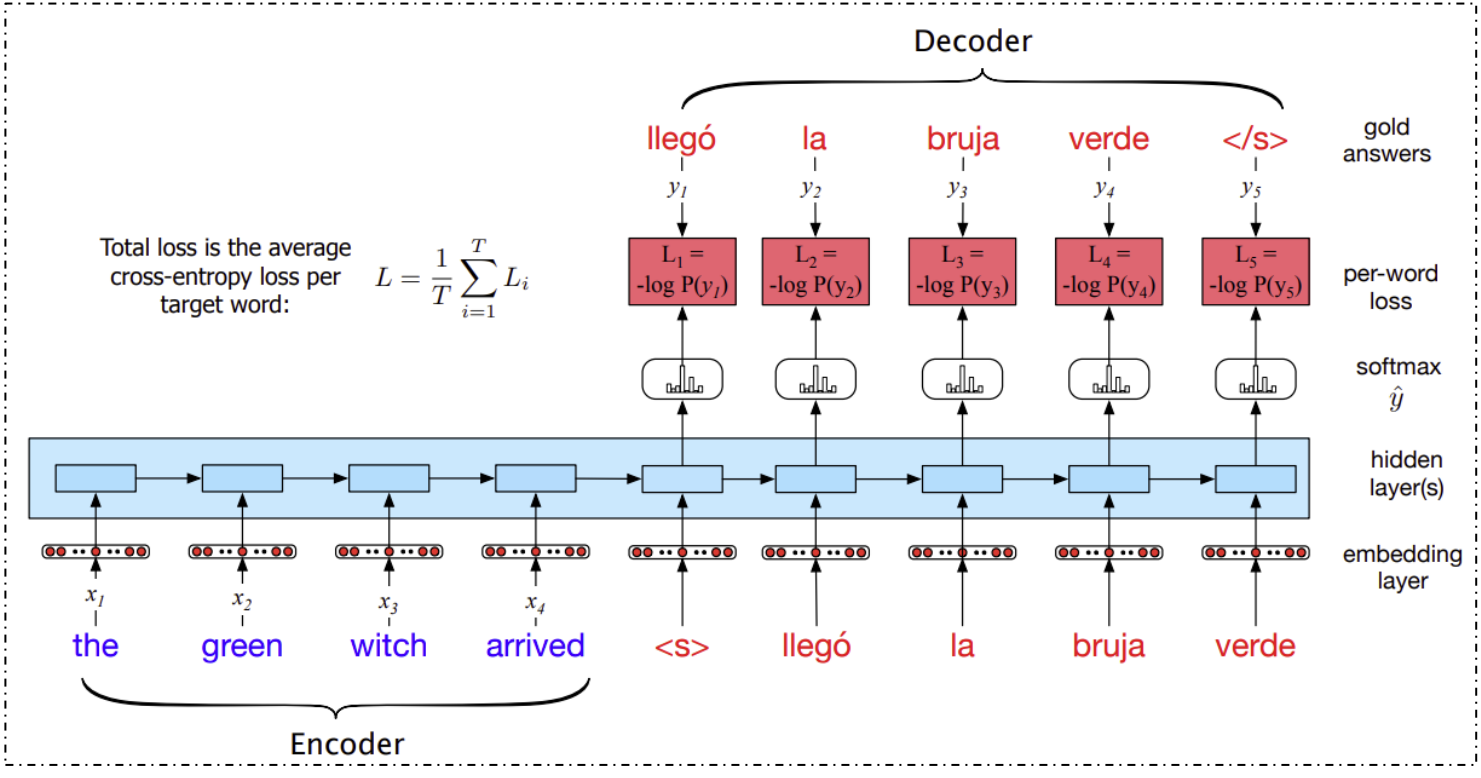
Puede ser RNN, LSTM, GRU

- One weakness of this approach as described so far is that the influence of the context vector, c, will wane as the output sequence is generated.
- A solution is to make the context vector c available at each step in the decoding process by adding it as a parameter to the computation of the current hidden state.

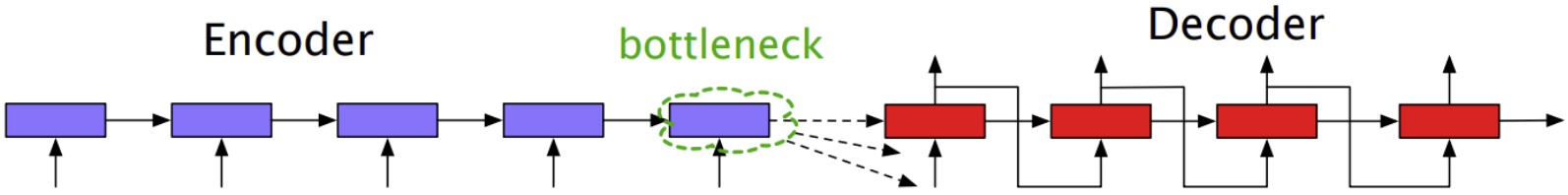
Encoder-Decoder with RNNs (IV) - Training

Teacher forcing: Note that in the decoder we usually don't propagate the model's softmax outputs,, but use teacher forcing to force each input to the correct gold value for training.

The training data typically consists of sets of sentences and their translation.



Attention (I)



Context must represent absolutely everything about the meaning of the source text, since the only thing the decoder knows about the source text is what's in this context vector.

Long sentences, may not be equally well represented in the context vector.

Gracias por la atención

¿Tiene alguna pregunta?