Procesamiento de Lenguaje Natural

Clase 12 – Modelos Secuenciales

Ph.D. Rubén Manrique

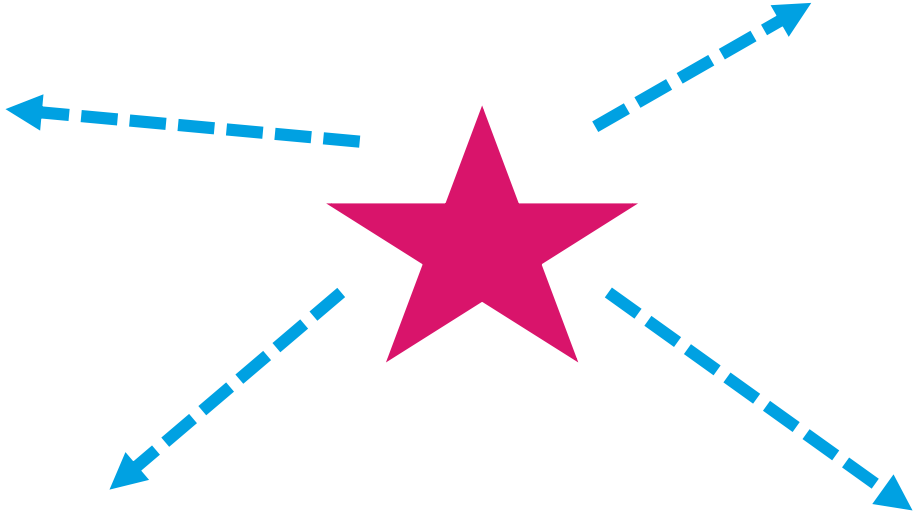
rf.manrique@uniandes.edu.co

Maestría en Ingeniería de Sistemas y Computación

- Chapter 9: Deep Learning Architectures for Sequence Processing

¿Por qué la secuencia es importante?

Suponga una estrella en movimiento. ¿A partir de la posición actual pueden predecir la siguiente posición?



¿Y ahora?

Con una alta probabilidad

El lenguaje es un fenómeno naturalmente temporal.

- Las redes feedforward utilizan entradas de tamaño fijo, junto con pesos asociados, para capturar todos los aspectos relevantes de un ejemplo a la vez.

- Esto hace que sea difícil lidiar con secuencias de diferente longitud y no capta importantes aspectos temporales del lenguaje.

El lenguaje es de naturaleza secuencial

Al que madruga Dios le _____.

Bogotá es la capital de _____.

Yo crecí en Brasil, pero ahora vivo en Bogotá. Gracias a eso hablo fluidamente _____ y _____.

Dependencias a corto y largo plazo.

Modelos de Lenguaje

$$P(fish|Thanks\ for\ all\ the)$$

- Los modelos de lenguaje nos dan la capacidad de asignar tal probabilidad condicional a cada posible palabra siguiente.

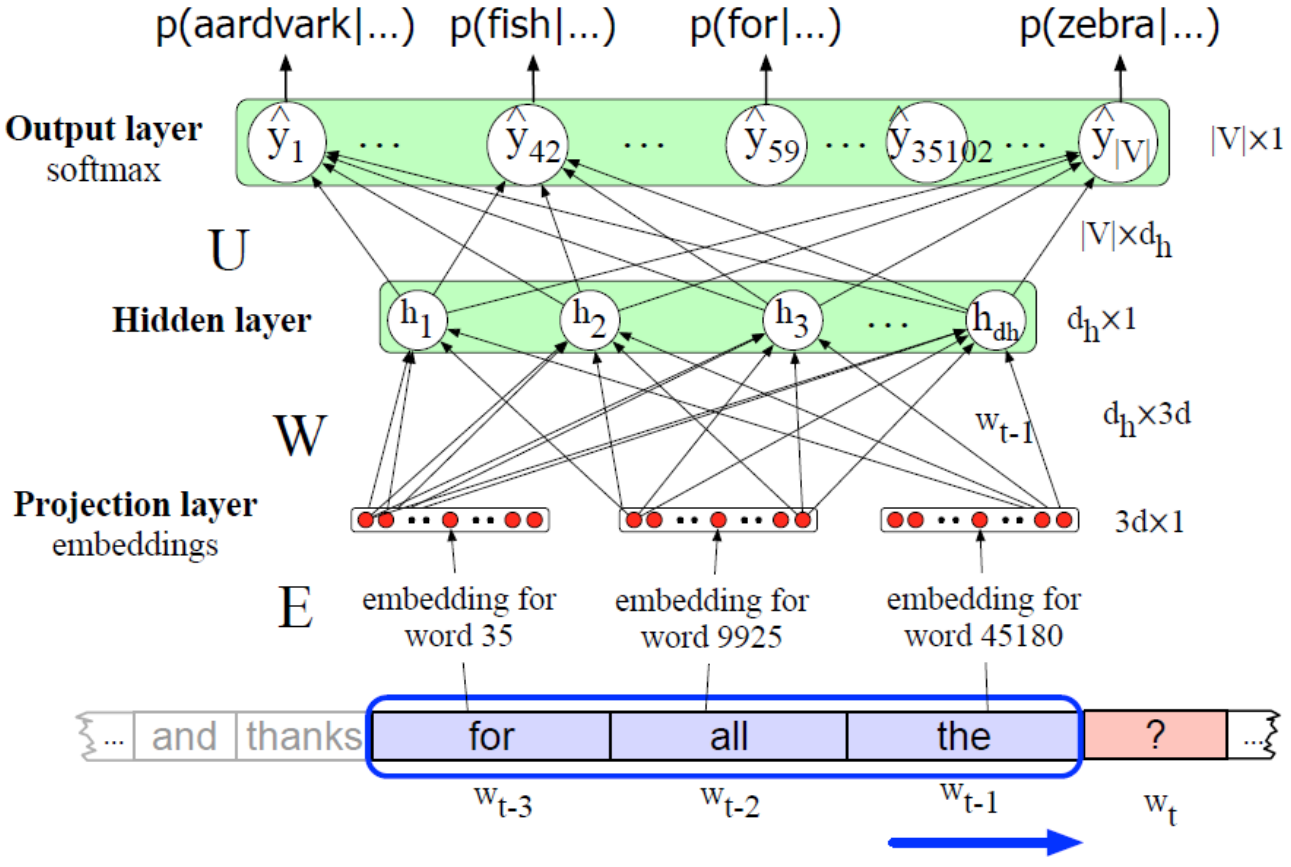
- También podemos asignar probabilidades a sentencias completas usando estas probabilidades condicionales en combinación con la regla de la cadena:

$$P(w_{1:n}) = \prod_{i=1}^{n} P(w_i|w_{n<i})$$

- Generar sentencias: **autoregressive generarion.**

Work-around

- El enfoque de ventana deslizante empleado con modelos de lenguaje neuronal.



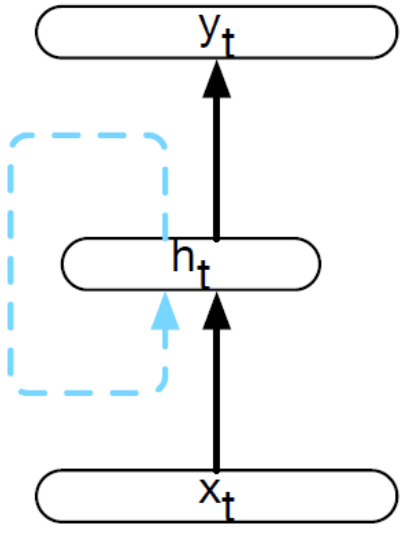
Solución problemática por:

- Las decisiones tomadas en una ventana no tienen ningún impacto en las decisiones posteriores.

- Limita el contexto del cual se puede extraer la información; cualquier cosa fuera de la ventana del contexto no tiene ningún impacto en la decisión que se está tomando.

- En segundo lugar, el uso de ventanas dificulta que las redes aprendan patrones sistemáticos que surgen por ejemplo de palabras compuestas.

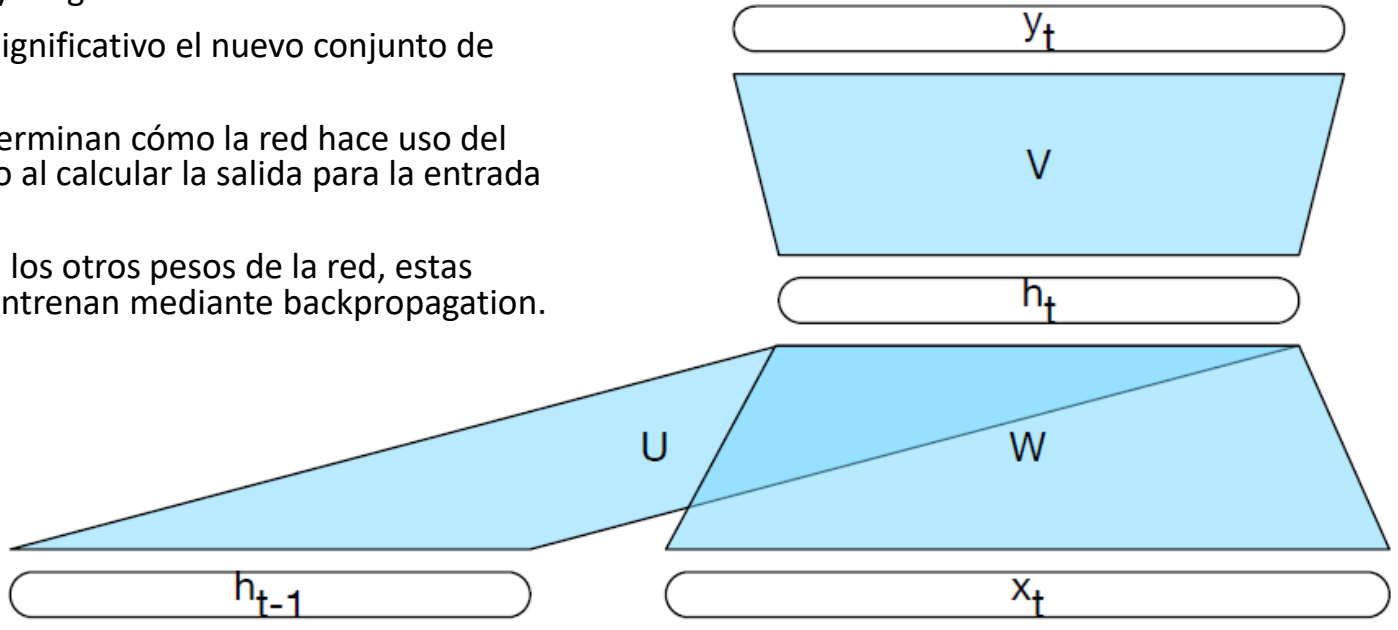
Redes Neuronales Recurrentes



- **Recurrent link**: This link augments the input to the computation at the hidden layer with the value of the *hidden layer from the preceding point in time*.
- The hidden layer from the previous time step provides a form of memory, or context, that encodes earlier processing and informs the decisions to be made at later points in time.

Desdoblando la recurrencia (I)

- De fondo no hay ningún cambio con la feedforward.
- El cambio mas significativo el nuevo conjunto de pesos U.
- Estos pesos determinan cómo la red hace uso del contexto pasado al calcular la salida para la entrada actual.
- Al igual que con los otros pesos de la red, estas conexiones se entrenan mediante backpropagation.

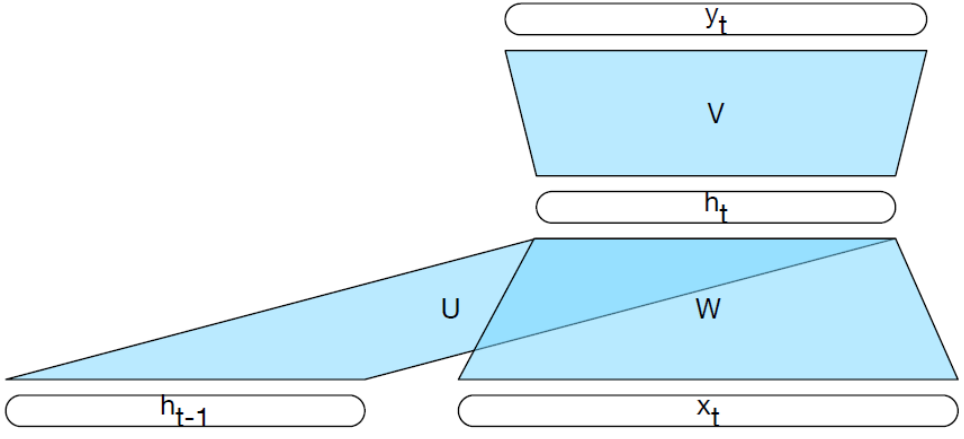


Desdoblando la recurrencia (II)

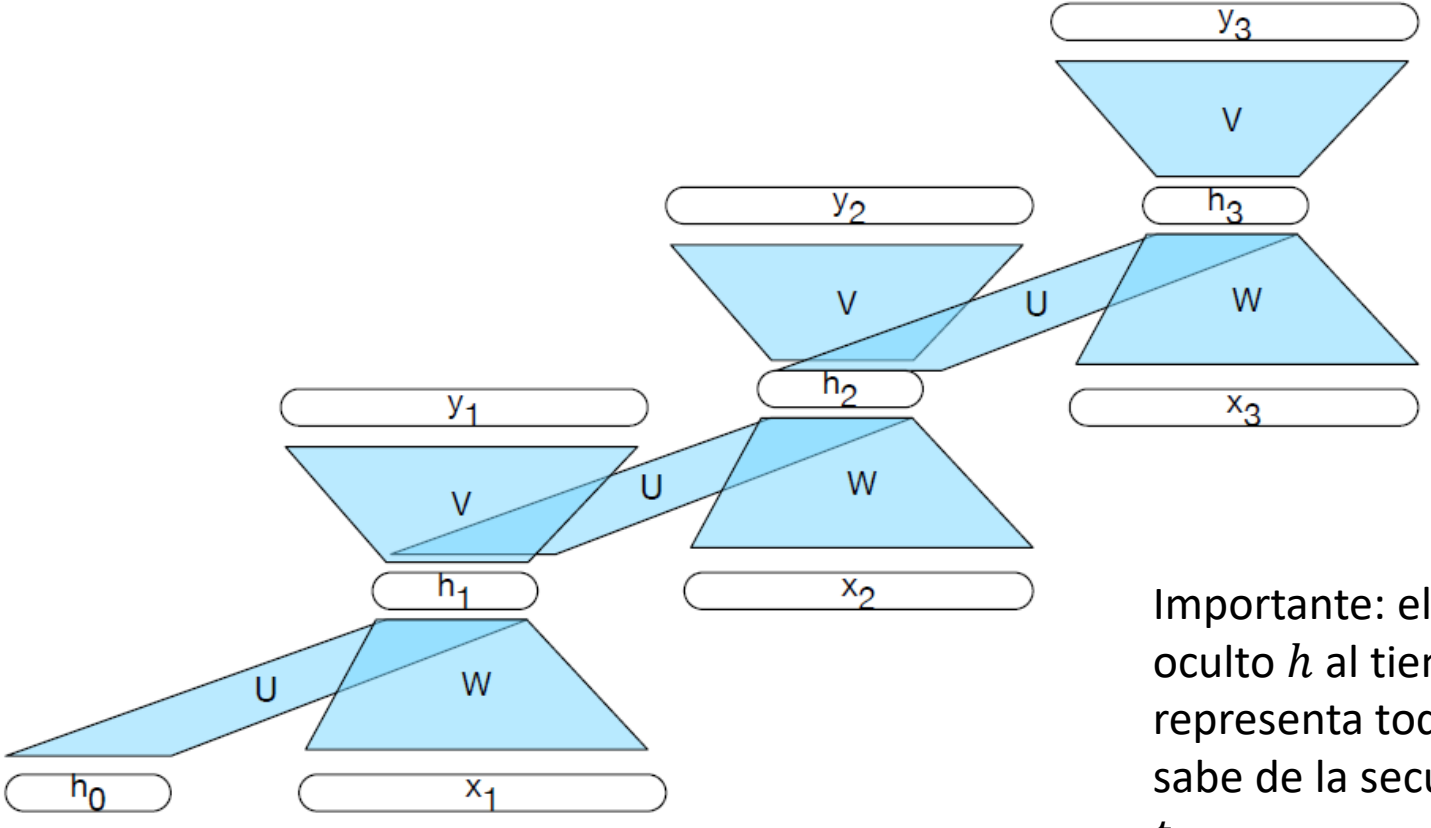
$$\boldsymbol{h_t = g(W x_t + U h_{t-1})}$$

$$\boldsymbol{y_t = g(V h_t)}$$

$$\boldsymbol{W} \in \mathbb{R}^{d_h \times d_{in}}$$
$$\boldsymbol{U} \in \mathbb{R}^{d_h \times d_h}$$
$$\boldsymbol{V} \in \mathbb{R}^{d_{out} \times d_h}$$



Desdoblando la recurrencia (III)



Importante: el estado oculto h al tiempo t representa todo lo que se sabe de la secuencia hasta t.

Desdoblando la recurrencia (IV)

function FORWARDRNN($x, network$) **returns** output sequence y

$h_0 \leftarrow 0$
for $i \leftarrow 1$ **to** LENGTH(x) **do**
    $h_i \leftarrow g(U\ h_{i-1} + W\ x_i)$
    $y_i \leftarrow f(V\ h_i)$
return y

RNN como modelo de lenguaje

- Se evita la restricción de contexto limitada inherente a los modelos de N-gramas, ya que el estado oculto incorpora información sobre todas las palabras precedentes desde el principio de la secuencia.

$e_t = \boldsymbol{E}^T x_t$ A cada paso se trae el embedding de la palabra actual x_t

x_t es un one-hot vector $|V| \times 1$.

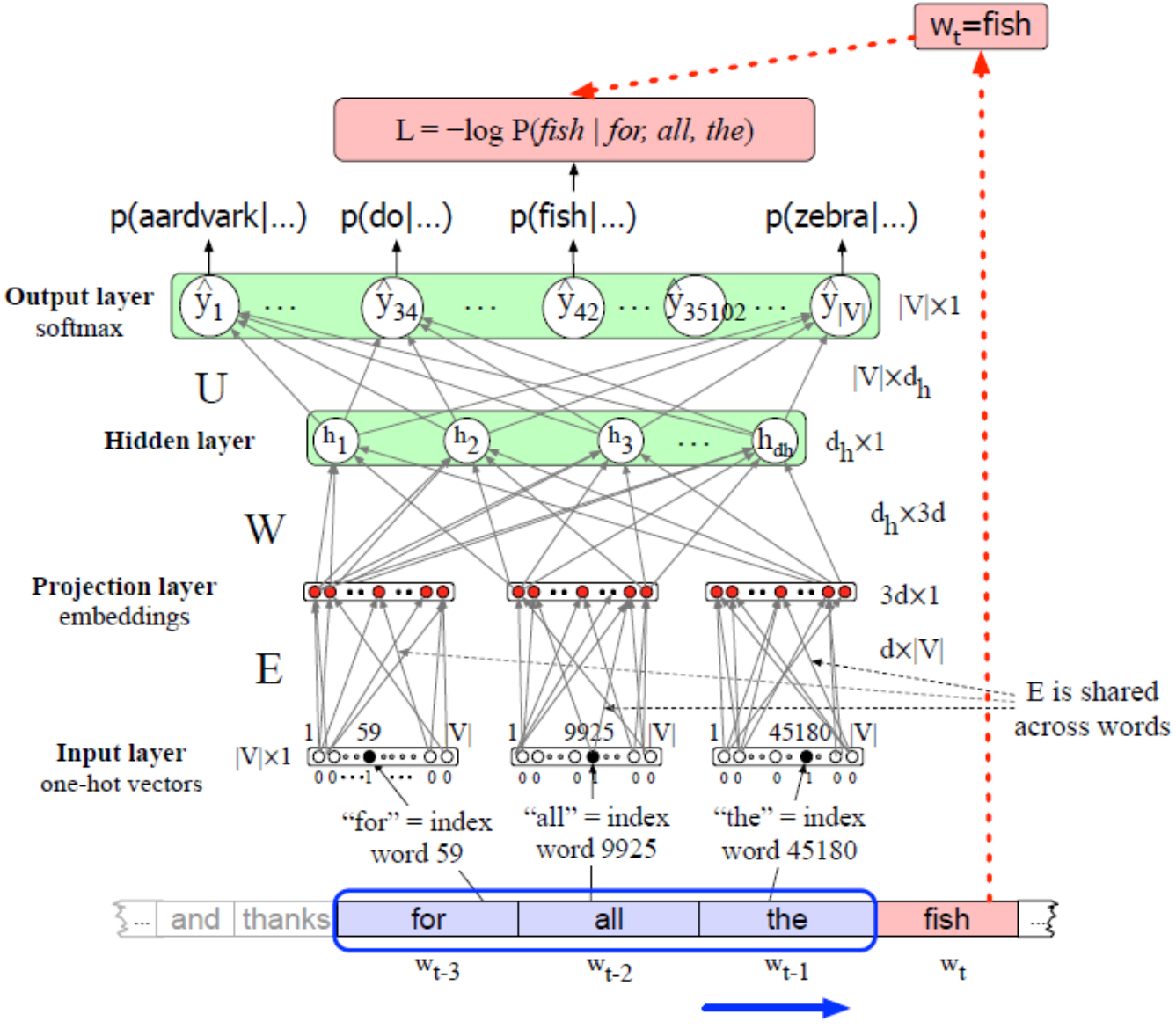
$\boldsymbol{h_t} = g(\boldsymbol{W e_t} + \boldsymbol{U h_{t-1}})$ Calculo de la capa oculta con retroalimentación.

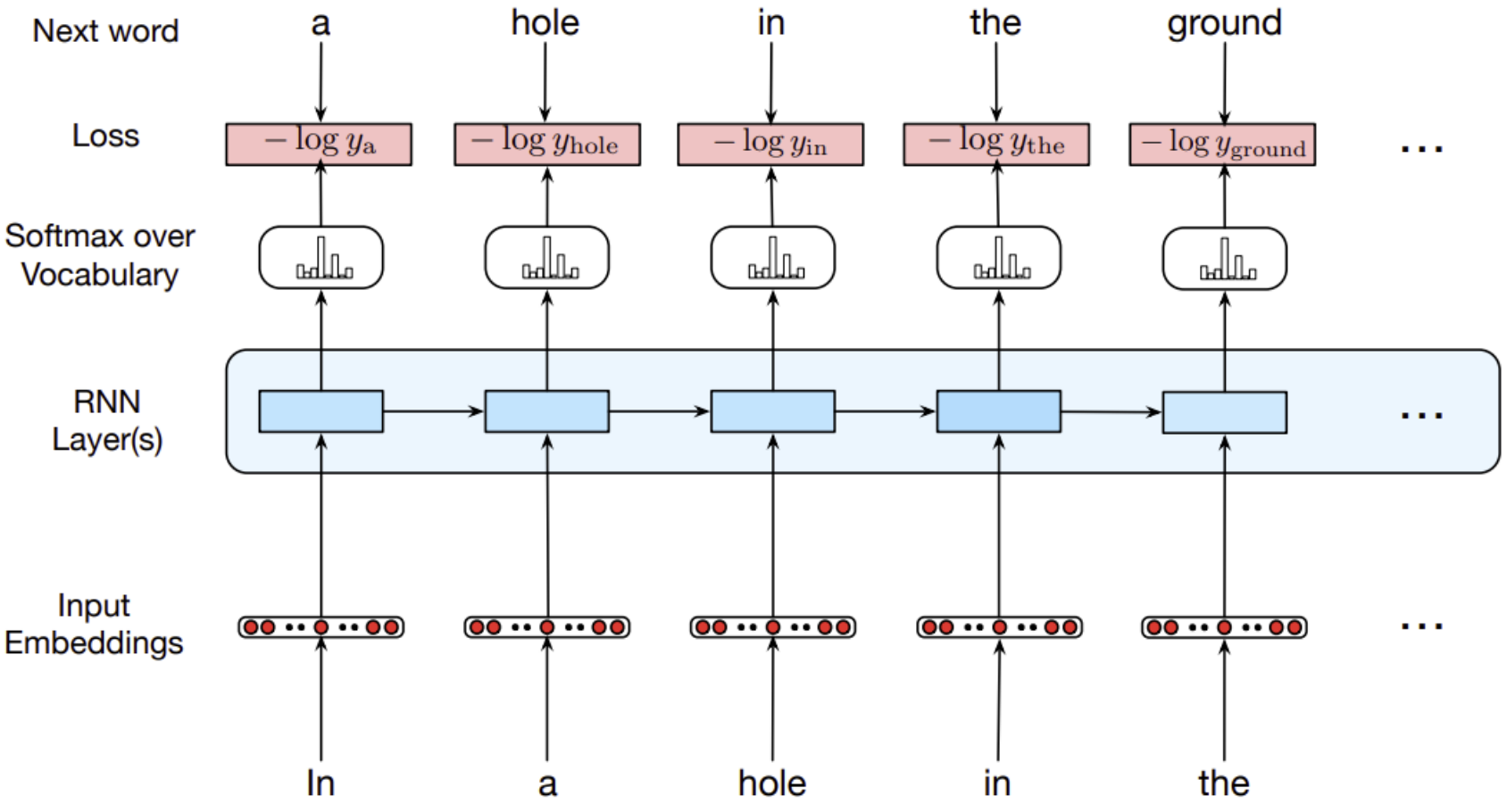
$\boldsymbol{y_t} = softmax(\boldsymbol{V h_t})$ Normalización de los scores en una distribución de probabilidad.

$P(w_{t+1} = i | w_{1:t}) = y_t^i$ La probabilidad de una palabra en un tiempo t corresponde a su componente en y.

$$P(w_{1:n}) = \prod_{i=1}^{n} y_{w_i}^i$$

Recordando...


w_t=fish
L = −log P(*fish* | *for, all, the*)
p(aardvark|...)
p(do|...)
p(fish|...)
p(zebra|...)
Output layer
softmax
$\hat{y}_1$ ... $\hat{y}_{34}$ ... $\hat{y}_{42}$ ... $\hat{y}_{35102}$ ... $\hat{y}_{|V|}$
|V|×1
U
|V|×d_h
Hidden layer
h_1 h_2 h_3 ... h_{dh}
d_h×1
W
d_h×3d
Projection layer
embeddings
3d×1
E
d×|V|
E is shared
across words
Input layer
one-hot vectors
|V|×1
1 59 |V|
1 9925 |V|
1 45180 |V|
"for" = index
word 59
"all" = index
word 9925
"the" = index
word 45180
... and thanks for all the fish ...
w_{t-3} w_{t-2} w_{t-1} w_t



$$\frac{1}{T}\sum_{t=1}^{T} L_{CE}$$

Y el entrenamiento…..?

$$L_{CE} = -\sum_{w \in V} y_w^t \log \hat{y}_w^t$$

Autoregressive generation

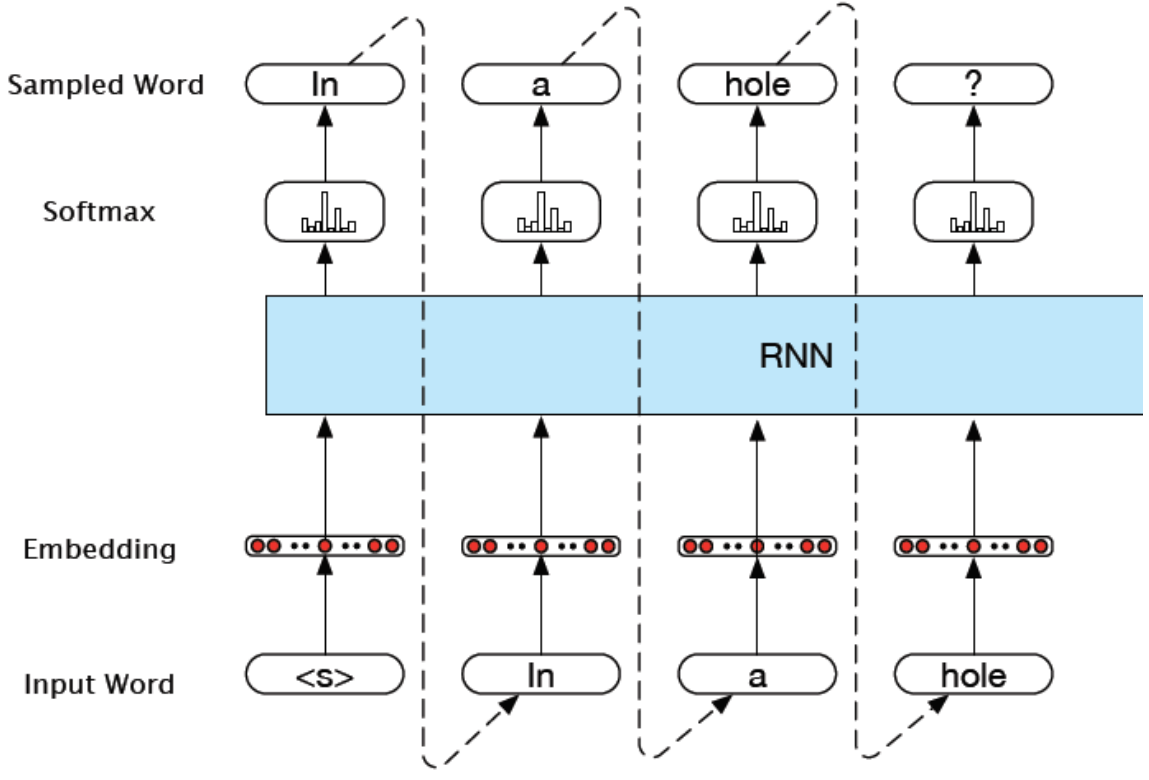
- To begin, sample a word in the output from the softmax distribution that results from using the beginning of sentence marker, <s>, as the first input.

- Use the word embedding for that first word as the input to the network at the next time step, and then sample the next word in the same fashion.

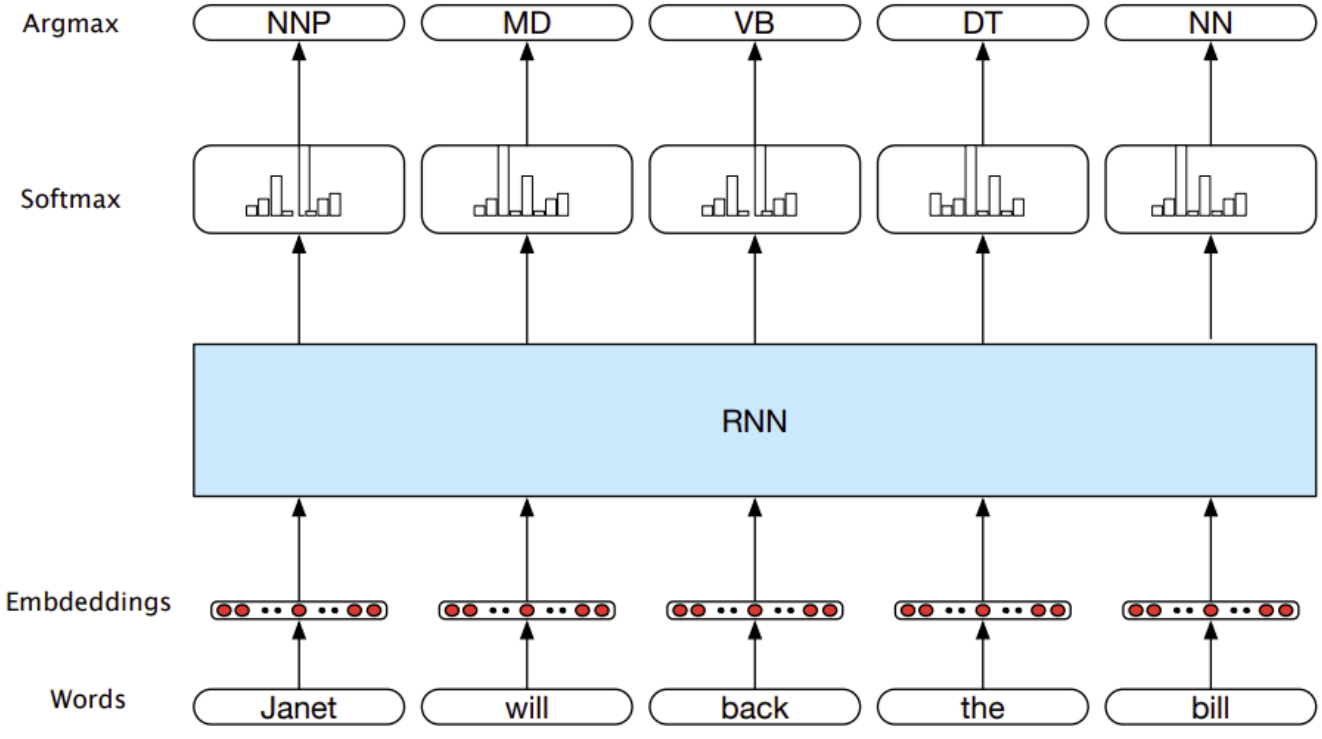
- Continue generating until the end of sentence marker, </s>, is sampled or a fixed length limit is reached.

Autoregressive generation

- Concepto importante para: traducción, resumen y sistema Q&A (MtoM, sequence to sequence).
- En lugar de simplemente usar <s> para comenzar, podemos proporcionar un contexto más rico y apropiado para la tarea.

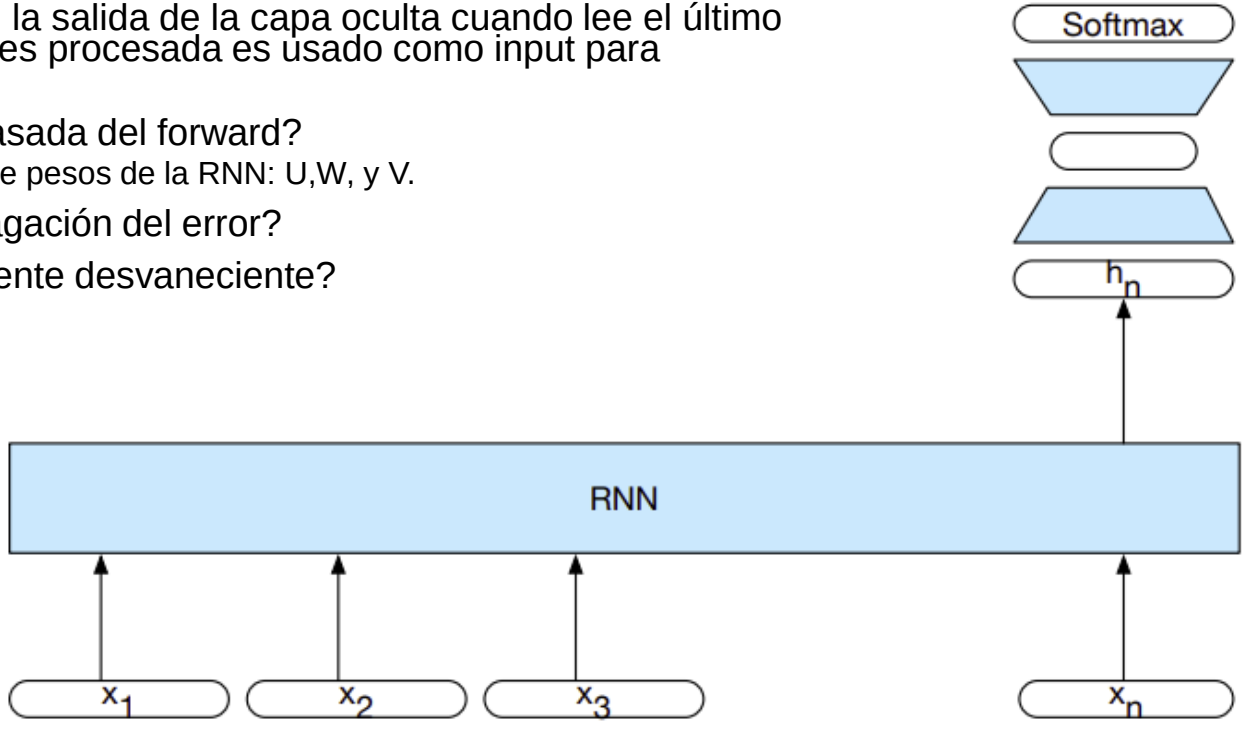


Aplicaciones (1): Etiquetado POS

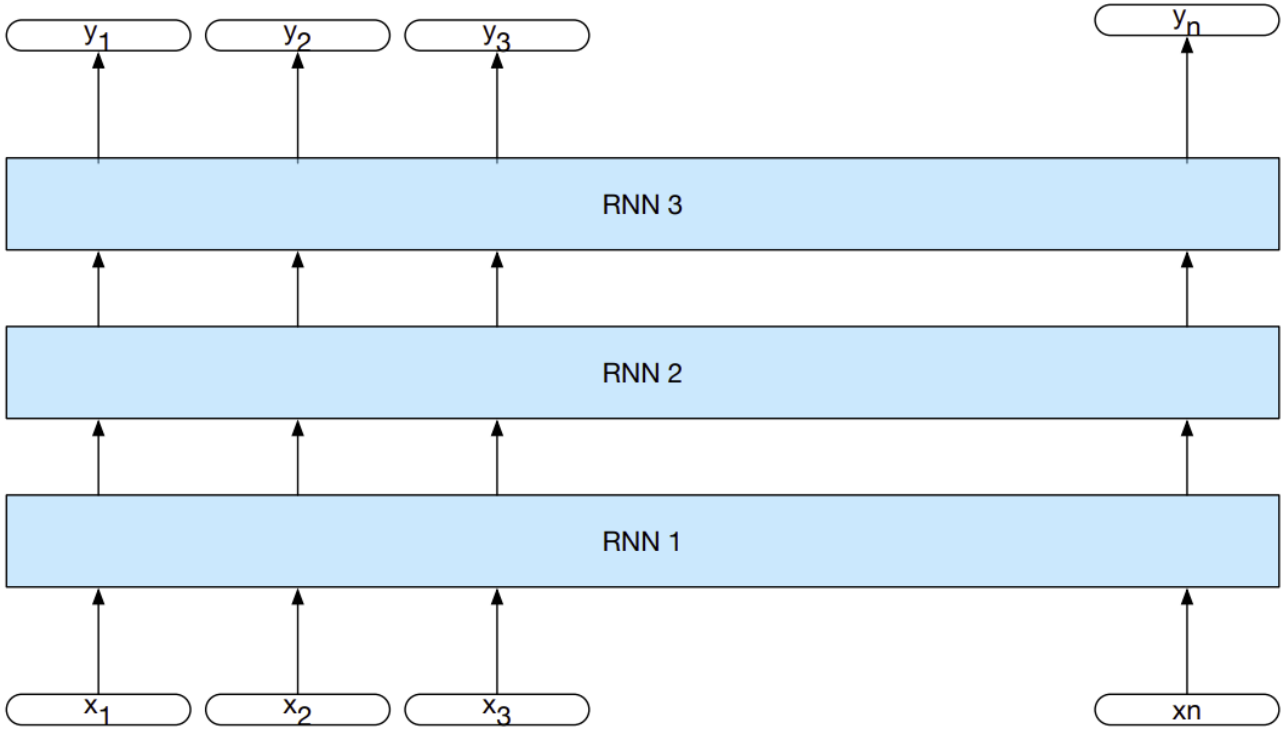


Aplicaciones (1): Clasificación de secuencias

- Sentimientos/spam por ejemplo.
- Nótese la arquitectura, la salida de la capa oculta cuando lee el último token de la secuencia es procesada es usado como input para feedforward.
- Como debe ser una pasada del forward?
  - Recuerden los sets de pesos de la RNN: U,W, y V.
- Como cambia la propagación del error?
- Que pasa con el gradiente desvaneciente?
- End-to-end training.



Stacked RNNs



Como debe ser una pasada del forward?

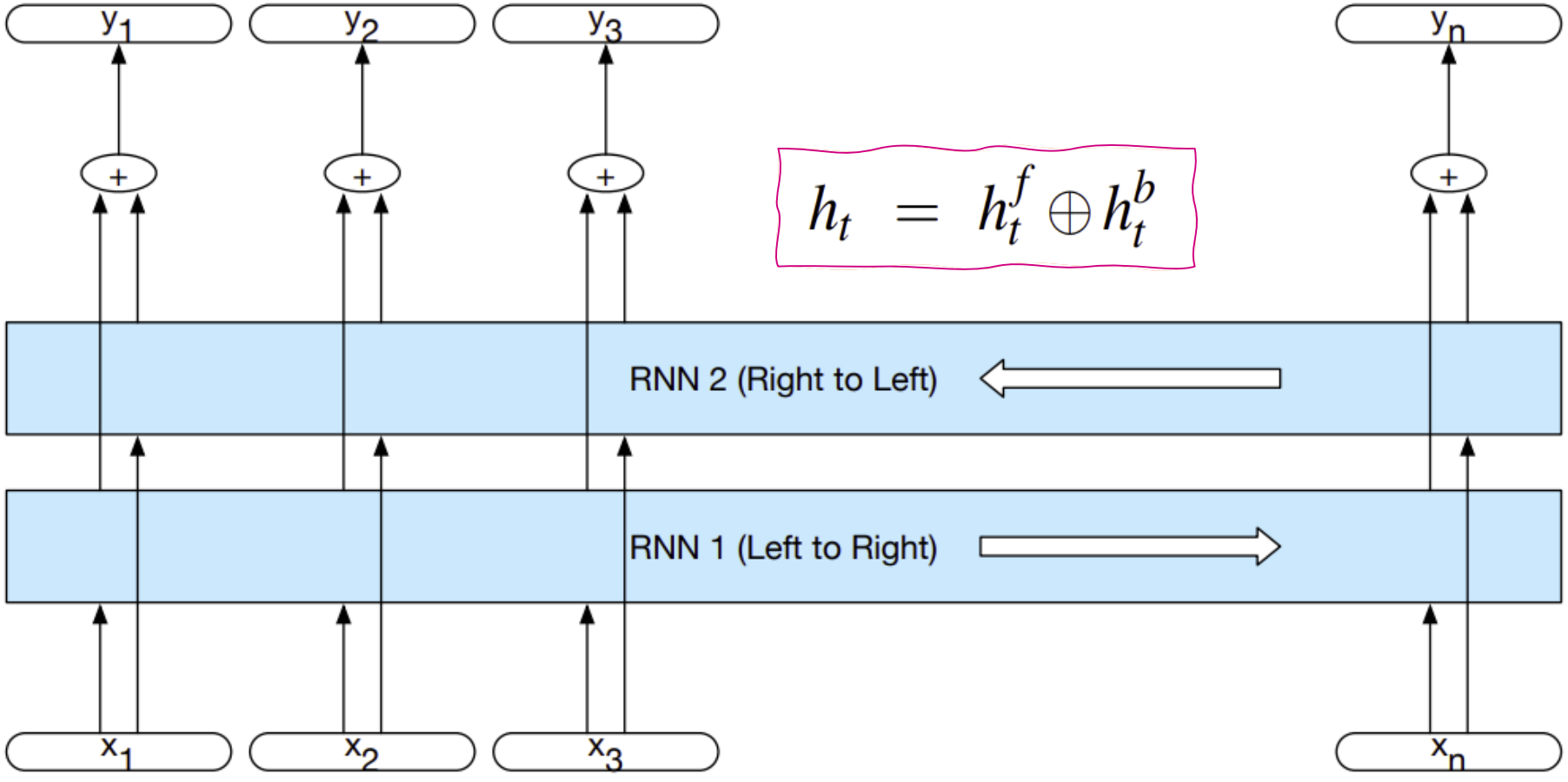
Recuerden los sets de pesos de la RNN: U,W, y V.

RNNs Bidireccionales

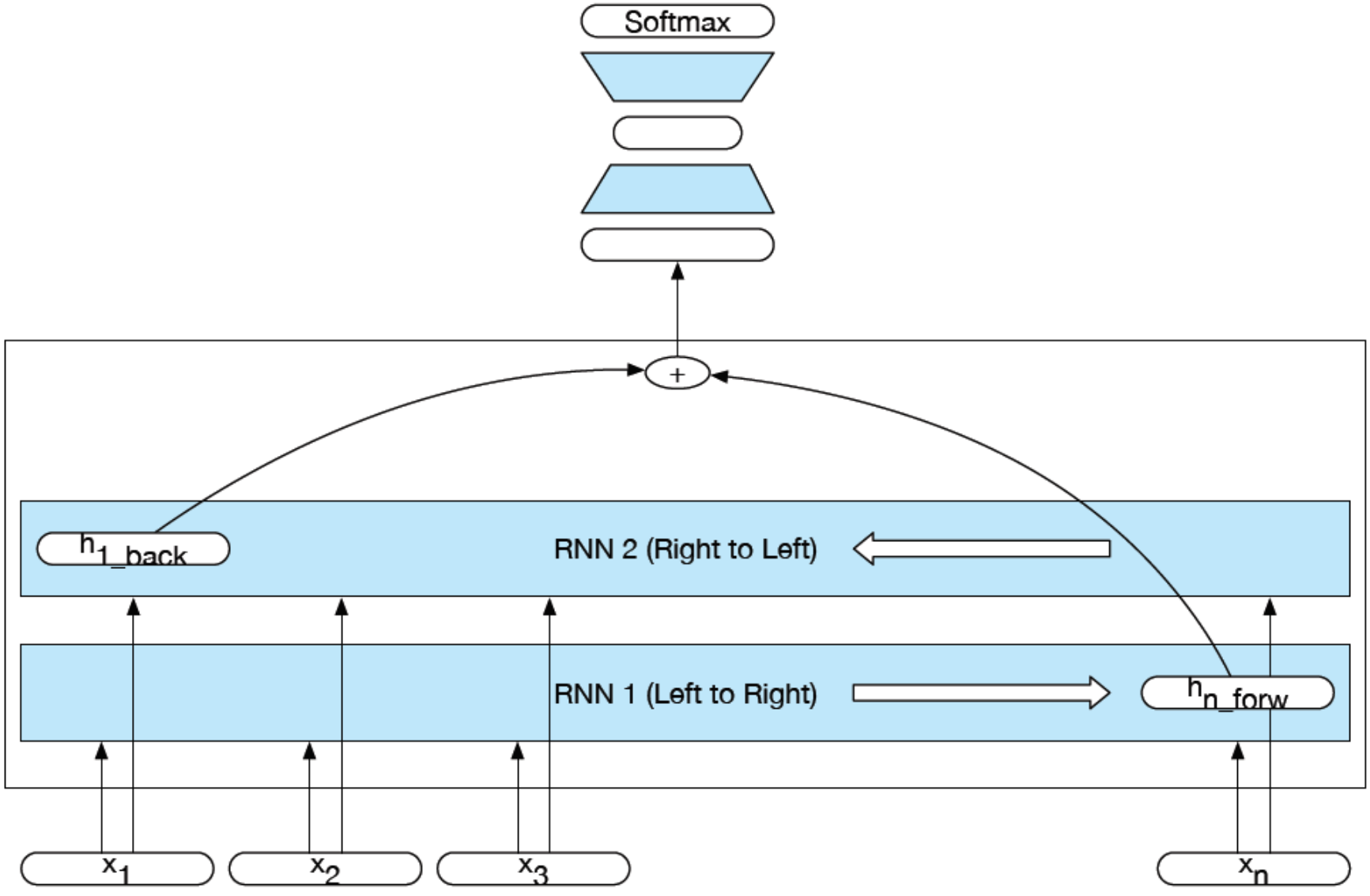
- Solo se puede construir cuando tenemos acceso a la secuencia completa.

- Un Bi-RNN consta de dos RNN independientes, uno donde la entrada se procesa desde el principio hasta el final y el otro desde el final hasta el principio.

- Luego combinamos las salidas de las dos redes en una sola representación que captura los contextos izquierdo y derecho de una entrada en cada momento.



- Para clasificación de secuencias hay un problema, cual?


Softmax
+
h_{1_back}
RNN 2 (Right to Left)
RNN 1 (Left to Right)
h_{n_forw}
x_1
x_2
x_3
x_n

Mejoras a las RNN: LSTMs y GRUs

Al que madruga Dios le _____.

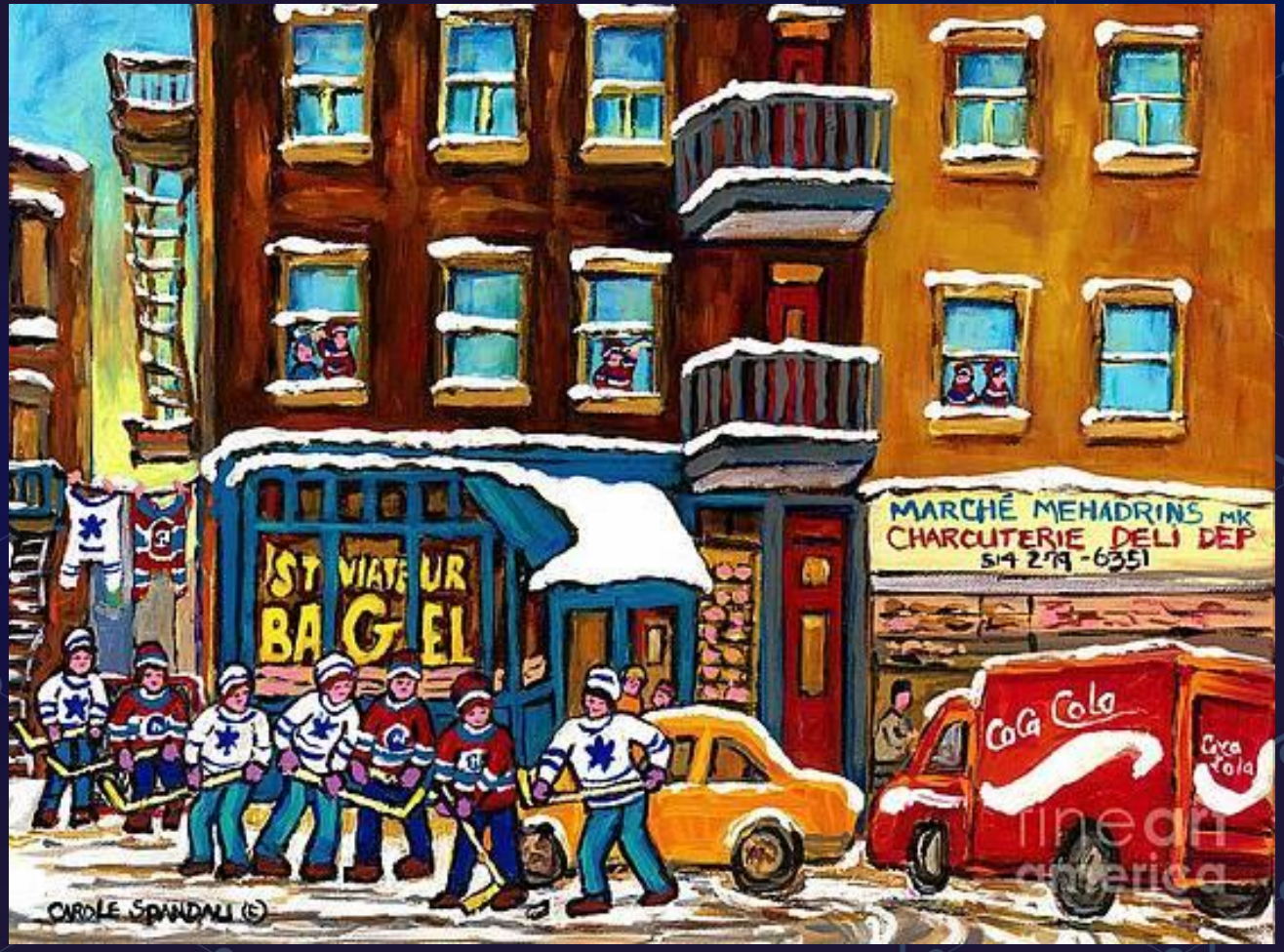
Bogotá es la capital de _____.

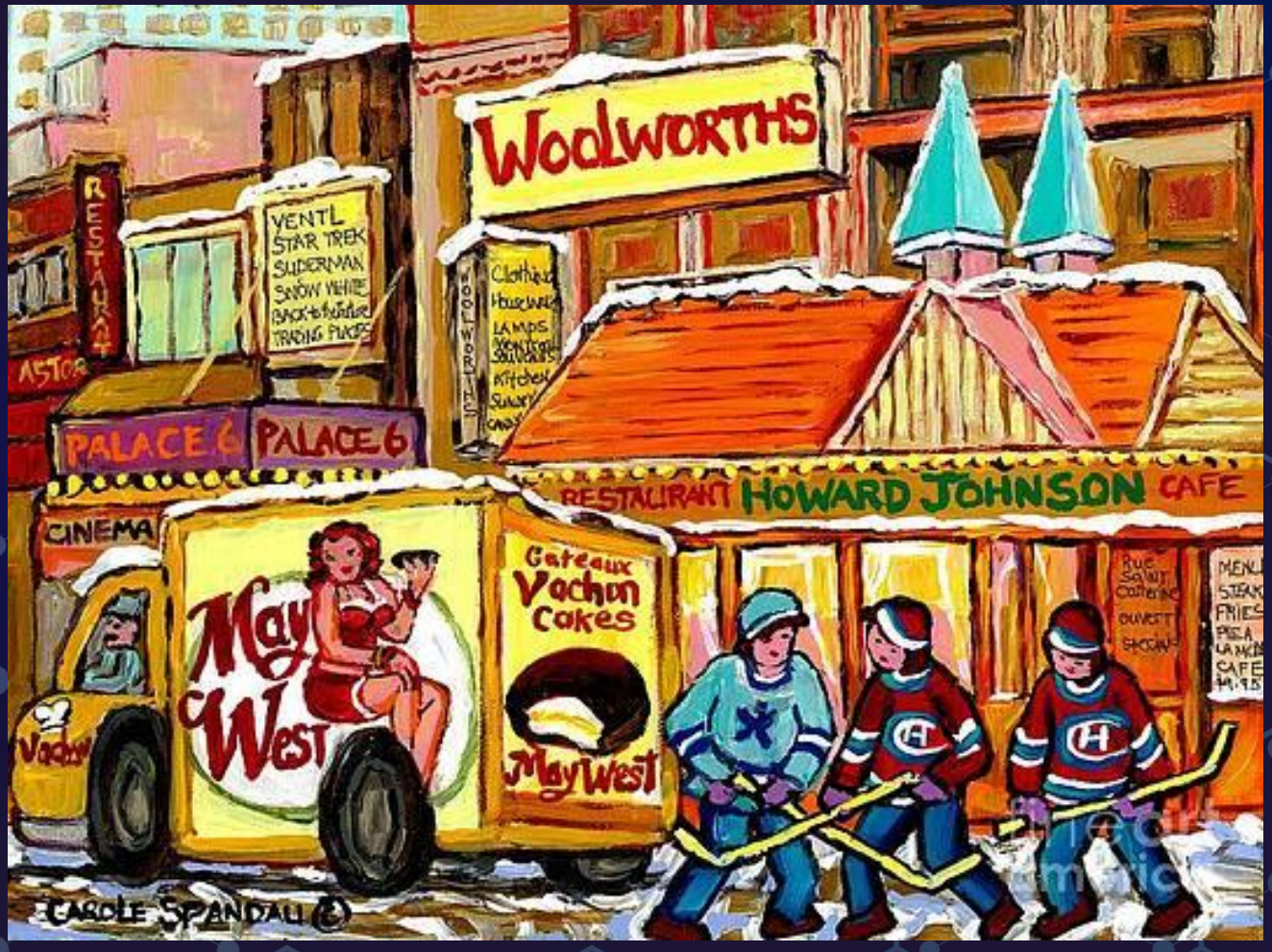
Yo crecí en Brasil, pero ahora vivo en Bogotá. Gracias a eso hablo fluidamente _____ y _____.

Dependencias a corto y largo plazo.

Una razón de la incapacidad de los RNN para transmitir información crítica es que a las capas ocultas y, por extensión, a los pesos que determinan los valores de salida de la capa oculta, se les pide que realicen dos tareas simultáneamente: **proporcionar información útil para la decisión actual** y **actualizar y llevar adelante la información requerida para decisiones futuras.**


ST MARIE
CUTE ST PAUL
Epicerie Deli
DEPANNEUR
FAMILIAL
Biere Vin
KIK


MARCHÉ MEHADRINS MK
CHARCUTERIE DELI DEP
514 279-6351
BAGEL
Coca Cola
CAROLE SPANDAU (C)


Woolworths
RESTAURANT
ASTOR
VENTE
STAR TREK
SUPERMAN
SNOW WHITE
TRADING PLACES
PALACE 6
PALACE 6
CINEMA
RESTAURANT HOWARD JOHNSON CAFE
Gateaux
Vachon
Cakes
May West
May West
CAROLE SPANDALI

$c_{t=0}$

4 jh
4c
2l

$h_{t=0}$

4jh

$c_{t=0}$

4 jh
4c
2l

$h_{t=0}$

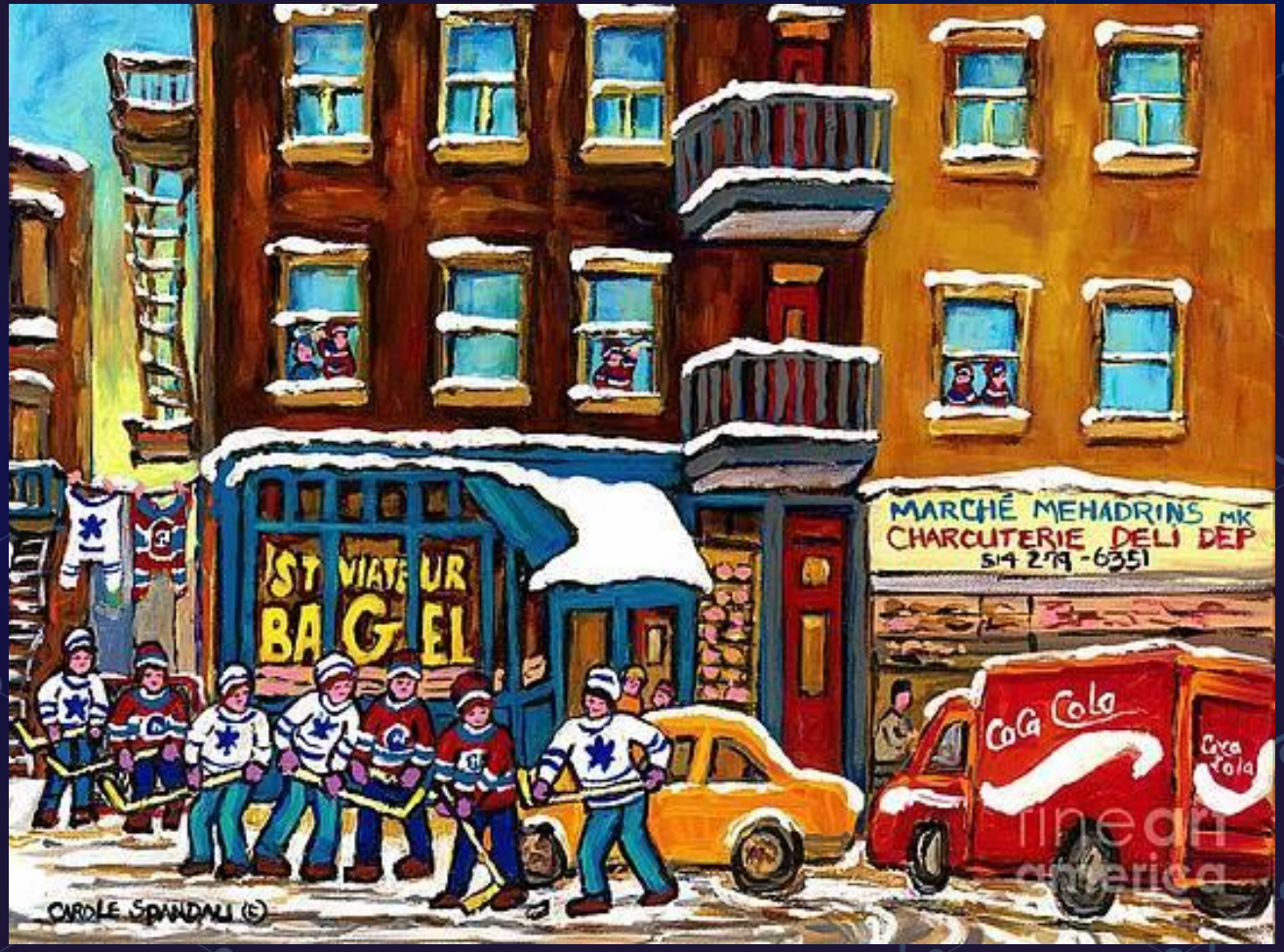
4jh

$c_{t=1}$

4jh,
7jh
2l
1CC

$h_{t=1}$

11jh



$c_{t=0}$

4 jh
4c
2l

$h_{t=0}$

4jh

$c_{t=1}$

4jh,
7jh
2l
1CC

$h_{t=1}$

11jh

$c_{t=2}$

4jh, 7jh,
3jh
2l

$h_{t=2}$

14jh

Gracias por la atención

¿Tiene alguna pregunta?