La inteligencia artificial generativa en la traducción de idiomas

Cómo se explica y cómo se aplica en el Grado en Traducción e Interpretación

Emilio Rodríguez Vázquez de Aldana

Departamento de Informática y Automática

Facultad de Traducción y Documentación

La inteligencia artificial generativa en la traducción de idiomas

- Asignaturas estrictamente dedicadas a las **tecnologías de la traducción (TT)** en el Grado en Traducción e Interpretación de la USAL

Informática básica (B) 6 crs.
Recursos tecnológicos para la traducción (O) 6 crs.
Gestión de proyectos (O) 3 crs.
Localización y Localización II (P) 6 y 3 crs.
Posedición (de traducción automática)(TA) (P) 3 crs.

Objetivo:
Entender la TA neuronal, usar cerebros artificiales para traducir, adaptarlos y evaluarlos

- **Objetivo general**: conocer, comprender y saber utilizar las que emplean los profesionales y las que están en investigación y desarrollo

La inteligencia artificial generativa en la traducción de idiomas

- ¡**Tecnologías de la traducción**! ¡Qué tiempos tan convulsos!

2014	Primer motor de traducción automática neuronal (TAN)
2016	Google pone en explotación su versión neuronal, reemplazando para muchos idiomas al antiguo modelo estadístico
2017	Aparece DeepL Se presenta la arquitectura ***Transformer*** (*Attention Is All You Need*): la propuesta original fue para la ¡TA!
2018-2022	Implantación masiva de ***Transformer***: Hugging Face Traducción automática multilingüe (transferencia de conocimiento); corpus sintéticos (*Backtranslations*) Traducción automática neuronal basada en documentos
2023 (Mayo)	La traducción con los **LLM** (GPT-3.5, GPT-4, LLaMA, etc.), basados en *Transformer*, presenta resultados de calidad parejos (en determinadas situaciones) que las redes neuronales específicamente ideadas para traducir.

Curso 2016-17
Explicábamos y creábamos motores de TA estadísticos. Introducimos la TAN

Curso 2017-18
Abandonamos la TA estadística y nos dedicamos solo a la TAN

Cursos 2018 a 22
Recogiendo los cambios en TAN

Cambios a la vista para el curso 2023-24:
La TA con LLM (**modelos preentrenados**) frente la TA con **modelos supervisados**.

La inteligencia artificial generativa en la traducción de idiomas

Lecciones

Historia de la traducción automática y métodos principales

- **Material profesor**

Historia de la TA y métodos principales

- **Lectura**

Kenny, Dorothy (2022): Human and machine translation. En Kenny, D. (Ed.), Mach
artificial intelligence (pp 23–49, capítulo 2). Berlin: Language Science Press

- **Documentación de interés**

Intento (2022): The State of Machine Translation. En línea: https://inten.to/machine-trans

Información sobre Chat GPT-3 en Wikipedia

Cuestionario Lección 1 (solo sobre material profesor)

Qué es una red neuronal artificial: nociones básicas

- **Documentos de estudio**

Nociones básicas sobre redes neuronales artificiales

- **Cuestionario**

Cuestionario Lección 2.

- **Recursos**

Una RNA sencilla llevada a una hoja de cálculo

la Inteligencia Artificial y el caos: dónde están los límites de ChatGPT. F

Prácticas

Práctica 1: Introducción a las redes neuronales artificiales. Resolviendo todos los problemas de clasificación de *Tensorflow Playground*

- **Herramienta**

TensorFlow Playground: An interactive in-browser visualization of neural networks

- **Entrega Tarea**

Práctica de clase (lunes, 6/2/2023): Tensorflow Playground

Práctica 2: Diseñando una RNA para reconocimiento de números escritos a mano (*handwritting*). Un segundo problema de clasificación

- **Documentación**

Desarrollo de una RNA

- **Fichero para la realización de la práctica. Un cuaderno Python**

Cuaderno Python para experimentar con la RNA (ejecutamos en Google Colab)

- **Herramienta**

Entorno de ejecución: Google Colaboratory

- **Entrega Tarea**

Objetivo 1: Qué es una red neuronal.

La inteligencia artificial generativa en la traducción de idiomas

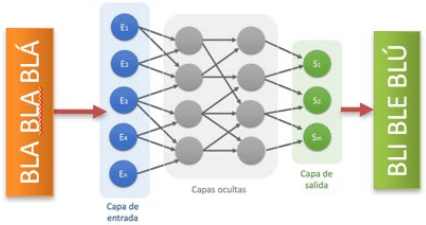
Qué es una red neuronal artificial. Nociones básicas

1. ¿Qué es una RNA?
2. Aprendizaje supervisado en RNA
3. Las RNA trabajan con números vectoriales
4. Arquitectura de las RNA
5. Una neurona en acción: su procesamiento
6. Entendiendo las RNA con TensorFlow Playground
7. Recapitulación y conclusiones
8. Bibliografía y fuentes

Acercamiento intuitivo a las RNA:
Estructura de la lección

La inteligencia artificial generativa en la traducción de idiomas

1. ¿Qué es una RNA?



- **Cerebro:**
  - Se estima que el cerebro humano tiene alrededor de 100M de neuronas y 100B (10^{14}) de conexiones entre ellas. Fuente: Investigación y Ciencia
- **RNA**

Numerous design choices can be made to scale neural network capacity, including adding more layers or making the hidden representations wider. Continuing our study on training deeper networks for translation, we utilized GPipe [4] to train 128-layer Transformers with over 6 billion parameter. Increasing the model capacity resulted in significantly improved performance across all languages by an average of 5 BLEU points. We also studied other properties of very deep networks, including the depth-width trade-off, trainability challenges and design choices for scaling Transformers to over 1500 layers with 84 billion parameters.

  - Miembros de Google Research experimentando con diferentes tamaños de RNA ideadas para traducir (Bapna y Firat: 2019)
    - ¡Ojo! Convertir sus "billions". Es decir, entre $6*10^9$ y $84*10^9$
  - La RNA *Chat GPT-3* contiene 175.000M de parámetros
    - Un parámetro en RNA es una conexión entre dos neuronas. Toda conexión (parámetro) es ajustable en fase de aprendizaje

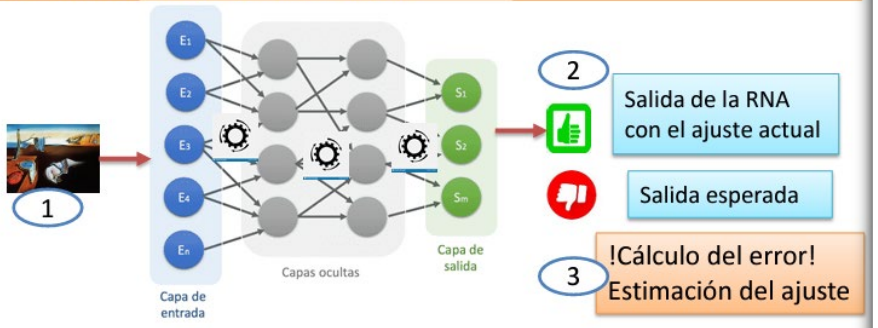
6

2. Aprendizaje supervisado en RNA

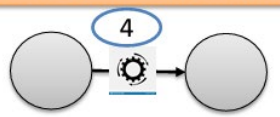
El entrenamiento lo realiza un algoritmo. Este algoritmo busca ajustar las conexiones de las neuronas en función de la salida de la RNA con la salida esperada. Este proceso, dato a dato, hasta completar la colección de entrenamiento, se suele repetir muchas veces. A cada pasada completa de la colección se le llama *epoch*. Es decir, la fase de aprendizaje conlleva N *epochs*

Entrenamiento

Recorrer colección N *epochs*.



Subir o bajar el valor de la conexión según error. Se distribuye el ajuste entre todas las conexiones de sus neuronas. **Cada conexión entre cada neurona es un parámetro ajustable**



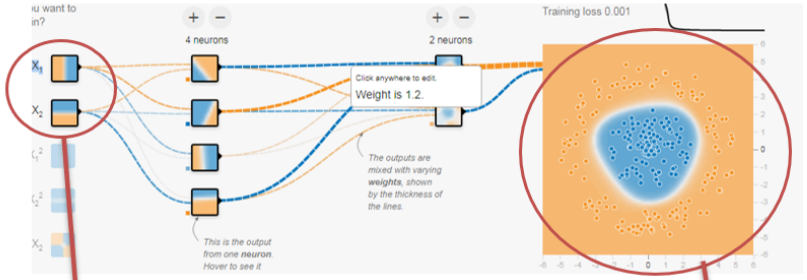
10

Acercamiento intuitivo a las RNA: Ejemplos del material docente

La inteligencia artificial generativa en la traducción de idiomas

6. Entendiendo las RNA con TensorFlow Playground

- ¿Cuántas capas? ¿Cuántas neuronas por capa?
  - Las neuronas de la capa de entrada y las de la capa de salida las "condiciona" el problema a tratar

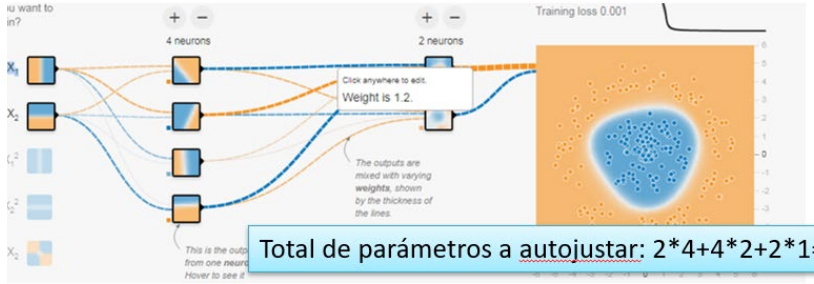


Queremos clasificar puntos pertenecientes a un espacio bidimensional: nos basta con el valor del eje de las X y de las Y. Dos neuronas de entrada para enviar por cada una la coordenada de las X y de las Y

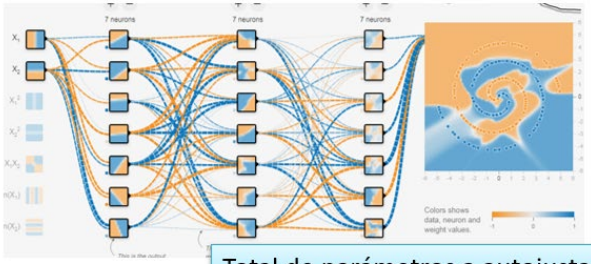
Una neurona de salida un número, que expresa, con una valor positivo (azul) o negativo (naranja), cómo clasifica nuestra RNA el punto dado en la entrada a través de las coordenadas X e Y . Ese valor es el **color del fondo**. Cuanto más intenso (naranja o azul) más claro está a quién pertenece ese punto del espacio. Los puntos en blanco (zona blanca) es zona difícil de establecer a quién pertenece (sin certeza). ¡La vida misma!: las fronteras siempre son territorios problemáticos

6. Entendiendo las RNA con TensorFlow Playground

- El conjunto de parámetros que se "autoajustarán" durante el entrenamiento será, por tanto, el total de conexiones entre neuronas



Total de parámetros a autojustar: 2*4+4*2+2*1=18



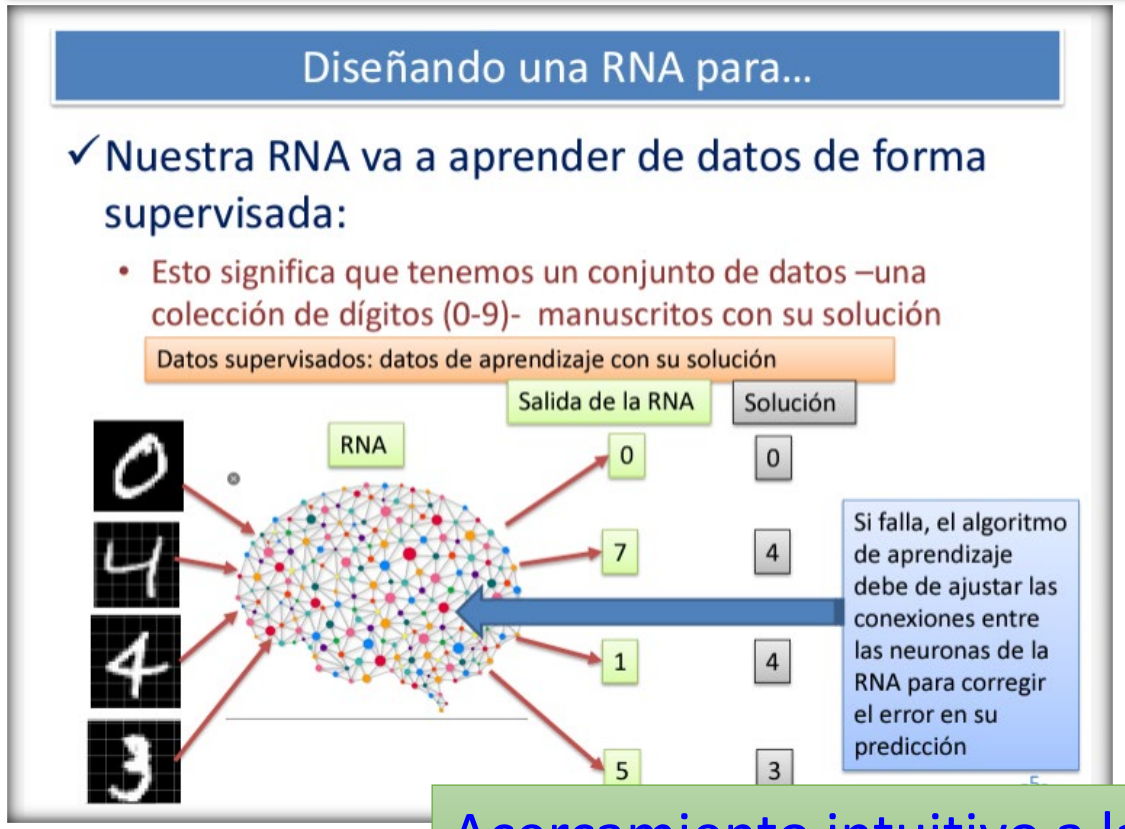
Total de parámetros a autojustar: 2*7+7*7+7*7+7*7+7*1=119

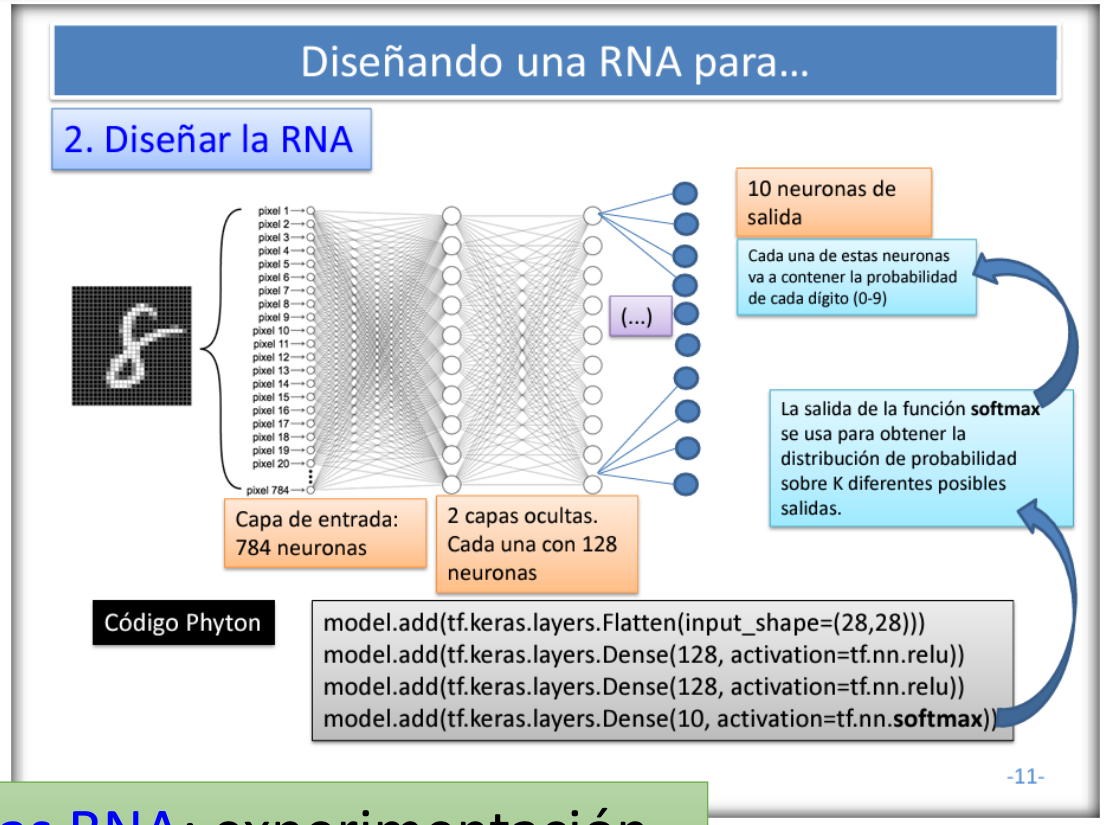
32

Acercamiento intuitivo a las RNA: experimentación

Práctica 1: TensorFlow Playground es una buena herramienta interactiva para entender qué es un RNA

La inteligencia artificial generativa en la traducción de idiomas





Acercamiento intuitivo a las RNA: experimentación

Práctica 2: Diseñando una RNA para el reconocimiento de dígitos escritos a mano

Modifican/ejecutan un cuaderno Python en *Google Colaboratory* sin tener ni idea de programación

La inteligencia artificial generativa en la traducción de idiomas

colab.research.google.com/drive/[redacted]

TA | DeepLearning | Google | Traducción Automá... | Apliaciones IA | Revistas | InterpretacionAuto... » | Otros marcadores

ReconoceNumeros0a9_clase08_02_2023.ipynb

Archivo Editar Ver Insertar Entorno de ejecución Herramientas Ayuda Últ

Comentario Compartir

+ Código + Texto Conectar

Creación de la red neuronal e indicación de algunos parámetros

Preparados los datos, creamos la red neuronal. Una capa de entrada que será la matriz 28*28 "aplanada" (Flatten), dos capas ocultas de 128 neuronas cada una y, por fin, una de salida de 10 neuronas, una para cada para cada número.

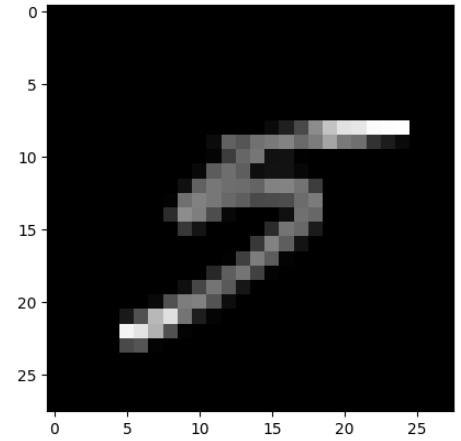
```
model= tf.keras.models.Sequential()
model.add(tf.keras.layers.Flatten(input_shape=(28, 28)))
#model.add(tf.keras.layers.Dense(128, activation=tf.nn.relu))
#model.add(tf.keras.layers.Dense(128, activation=tf.nn.relu))
model.add(tf.keras.layers.Dense(10, activation=tf.nn.softmax))
```

Tras crear las capas, indicamos cómo queremos que se comporte la red al corregir (supervisar) las predicciones erróneas. De forma elemental, cómo queremos que se distribuya el ajuste de las neuronas para corregir/aprender

```
[ ] model.compile(op
                  los
```

Mostra

```
import random
prueba=random.randint(0,len(x_testf))
plt.imshow(x_testf[prueba],cmap="gray")
plt.show()
j=0
for i in predictions[prueba]:
  print(j,"-->",round(i,5),end=" || ")
  j+=1
print()
print("El número creo que es el ",np.argmax(predictions[prueba]))
if np.argmax(predictions[prueba])!= y_test[prueba]:
            print("¡¡¡FALLÉ!!! Era el ", y_test[prueba])
else:
            print("¡¡¡ACERTÉ!!!")
```




```
0 --> 0.0 || 1 --> 0.0 || 2 --> 0.0 || 3 --> 0.0 || 4 --> 0.00014 || 5 --> 0.94035 || 6 --> 0.0 |
```

Acercamiento intuitivo a las RNA: experimentación

Práctica 2: Diseñando una RNA para el reconocimiento de dígitos escritos a mano

Modifican/ejecutan un cuaderno Python en *Google Colaboratory* sin tener ni idea de programación

La inteligencia artificial generativa en la traducción de idiomas

- Presentados los conceptos e ideas básicas que subyacen en las RNA, estamos preparados para abordar y aproximar la **traducción automática neuronal**
  - De forma intuitiva pero rigurosa
  - Conjuntando teoría con práctica
  - Mostramos nuestra propuesta en el curso de Studium 2022-23

La inteligencia artificial generativa en la traducción de idiomas

La traducción automática neuronal

- **Documentos de estudio**
  - La traducción automática neuronal
  - Diapositiva 14:
  - Diapositiva 18
  - Diapositiva 29
  - Diapositiva 48: arquitectura Transformer traduciendo una oración. Gráfico centrado en la entrada y salida
  - Diapositiva 53: arquitectura Transformer traduciendo. Gráfico centrado en los módulos de atención (extraído de Uszkoreit: 2017)
  - Diapositiva 59: evolución de la calidad de la traducción automática del motor de Google desde 2006 (inicio del motor PB-MT de Google) hasta 2020 acorde con la métrica BLEU
- **Lecturas**
  - Jay Alammar (2019): The Illustrated Word2vec. [Mensaje en un blog] Leer/comprender solo hasta apartado "Analogies" (incluido). ¡"Language Modeling" ya no!
  - Jay Alammar (2019): The Illustrated Word2vec. [Mensaje en un blog] (En PDF)
  - DeepL (2021): ¿Cómo funciona DeepL? [Mensaje en un blog]
  - Forcada, M. L. (2019): Traducción automática neural: cómo funciona y qué se puede esperar de ella. [Diapositivas]. Departamento de Lenguajes y Sistemas, Universidad de Alicante.
  - Pérez-Ortiz, J. A.; Forcada, M. L. y Sánchez-Martínez, F.: (2022): How neural machine translation works. En Kenny, D. (Ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence (pp 141-164, cap. 7). Berlin: LSP
- **Recursos**
  - Embedding projector

Teoría

Práctica 3: Adaptación a un dominio (*fine tuning*) de un modelo de OPUS-CAT

- **Documentación (consultar, informarse,...)**
  - Adaptación a un dominio de un modelo de OPUS-CAT
  - Información sobre los modelos de OPUS-CAT MT
  - Ramírez Sánchez, Gema (2022): Custom neural machine translation. En Kenny, D. (Ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence (pp 165–186, capítulo 8). Berlin: Language Science Press
  - **Formas de intentar mejorar/corregir la salida de un motor de TA**
  - OPUS CAT MT: Edit rules
  - OPUS CAT MT: Fine-tuning
- **Software**
  - Plugin OPUS-CAT para Trados 2022
  - Instalación de OPUS-CAT MT Engine
  - Entorno de ejecución: Google Colaboratory
  - Cuaderno con script en Python para fragmentar corpus en aprendizaje (de diferentes tamaños), test y desarrollo. Se asumen en formato TSV
  - Cuaderno con script en Python para evaluar con métrica BLEU, chrF y TER (actualizado 20/02/2023 a las 19:20)
- **Corpus para adaptar al dominio**
  - Corpus ONUSIDA, compilado por Jesús Torres (en TSV)
- **Descarga modelos de OPUS-CAT MT en Hugging Face**
  - Hugging Face: descarga de modelos de TA de Helsinki NLP group
- **Entrega Tarea**

Prácticas

La inteligencia artificial generativa en la traducción de idiomas

La traducción automática neuronal

1. La TAN: planteamiento del problema
2. La entrada y la salida de una RNA para TAN
3. Arquitecturas de TAN propuestas
4. La arquitectura *Transformer*
5. Conclusiones y vías de investigación en TAN
6. Bibliografía

Teoría

Este año, claro, han recibido noticias de que la arquitectura ***Transformer***, propuesta en 2017 para TA, es también la de **ChatGPT** (Chat ***Generative*** Pretrained ***Transformer***)

También han entendido de dónde viene el término IA *Generativa*

¡Buscamos un acercamiento intuitivo pero riguroso!

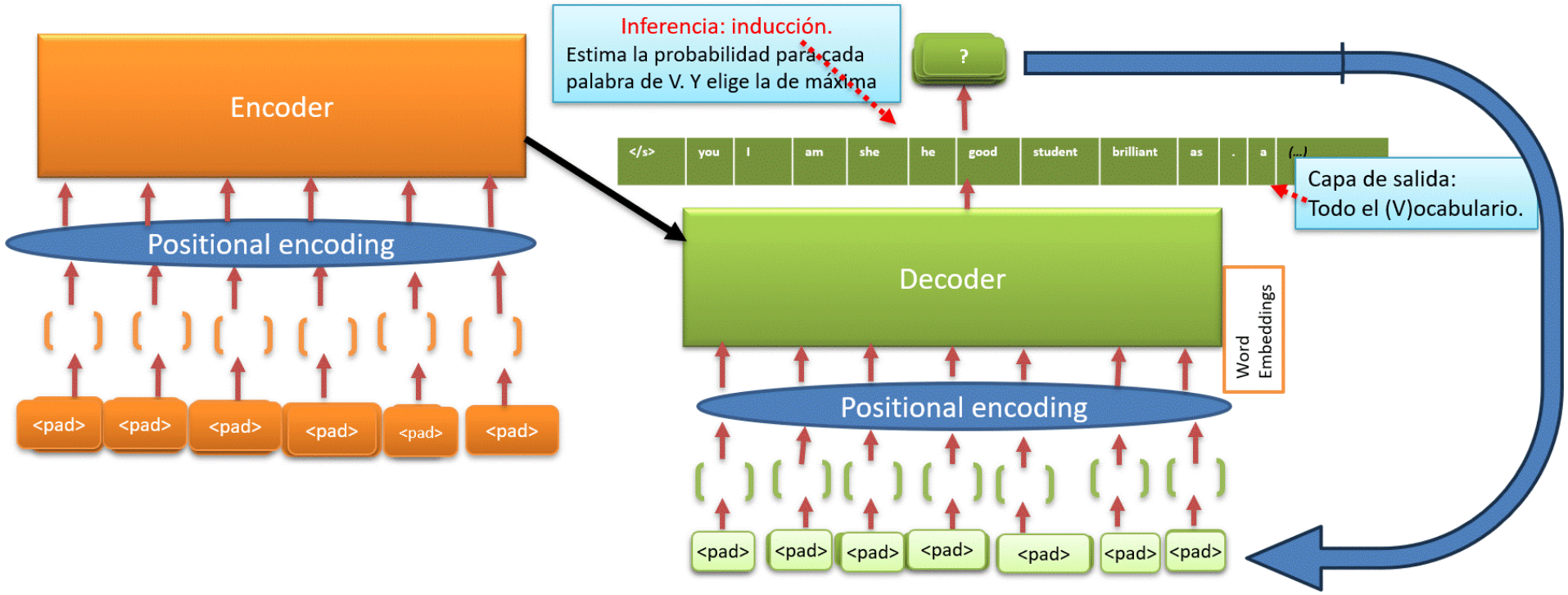
La inteligencia artificial generativa en la traducción de idiomas

4. La arquitectura *Transformer*

Teoría

La arquitectura *Transformer* traduciendo una oración

A este método en que la salida (palabra generada) pasa a ser la entrada para emitir la siguiente se le llama **autorregresivo.** Todos los motores de TAN de **todas las arquitecturas son autorregresivos. También los modelos de lenguaje**



¿Por qué **autorregresivos**?
¿Por qué IA **generativa**?

Ejemplo de material de explicación de cómo genera palabras la IA actual en una RNA ideada para traducir

4. La arquitectura *Transformer*

Teoría

La arquitectura *Transformer* traduciendo una oración destacando los módulos de autoatención del codificador y del decodificador más el módulo atención que une ambos en el decodificador

El gráfico animado intenta mostrar cómo la oración en **origen** es analizada en el módulo de **autoatención,** capturando las relaciones de cada palabra con el resto de palabras. El sistema que se muestra contiene 3 capas codificadoras.
Empieza la traducción (decodificación): El sistema también contiene 3 módulos decodificadores superpuestos. Para cada palabra a generar, el sistema primero observa las relaciones entre las palabras que lleva generadas de **destino** (**módulo de autoatención**). ¡Claro, sólo las previas! A continuación, entre en juego el **módulo de atención:** observar en qué palabras de origen ha de concentrarse para generar la siguiente palabra en destino.

Gráfico extraído de **Uszkoreit** (2017)

53

La inteligencia artificial generativa en la traducción de idiomas

Práctica 3: Adaptación a un dominio (*fine tuning*) de un modelo de OPUS-CAT

- **Documentación (consultar, informarse,...)**
  - Adaptación a un dominio de un modelo de OPUS-CAT
  - Información sobre los modelos de OPUS-CAT MT
  - Ramírez Sánchez, Gema (2022): Custom neural machine translation. En Kenny, D. (Ed.), Machine translation for everyone: Empowering users in the age of artificial intelligence (pp 165–186, capítulo 8). Berlin: Language Science Press
    - **Formas de intentar mejorar/corregir la salida de un motor de TA**
  - OPUS CAT MT: Edit rules
  - OPUS CAT MT: Fine-tuning
- **Software**
  - Plugin OPUS-CAT para Trados 2022
  - Instalación de OPUS-CAT MT Engine
  - Entorno de ejecución: Google Colaboratory
  - Cuaderno con script en Python para fragmentar corpus en aprendizaje (de diferentes tamaños), test y desarrollo. Se asumen en formato TSV
  - Cuaderno con script en Python para evaluar con métrica BLEU, chrF y TER (actualizado 20/02/2023 a las 19:20)
- **Corpus para adaptar al dominio**
  - Corpus ONUSIDA, compilado por Jesús Torres (en TSV)
- **Descarga modelos de OPUS-CAT MT en Hugging Face**
  - Hugging Face: descarga de modelos de TA de Helsinki NLP group
- **Entrega Tarea**

Práctica sobre traducción automática neuronal

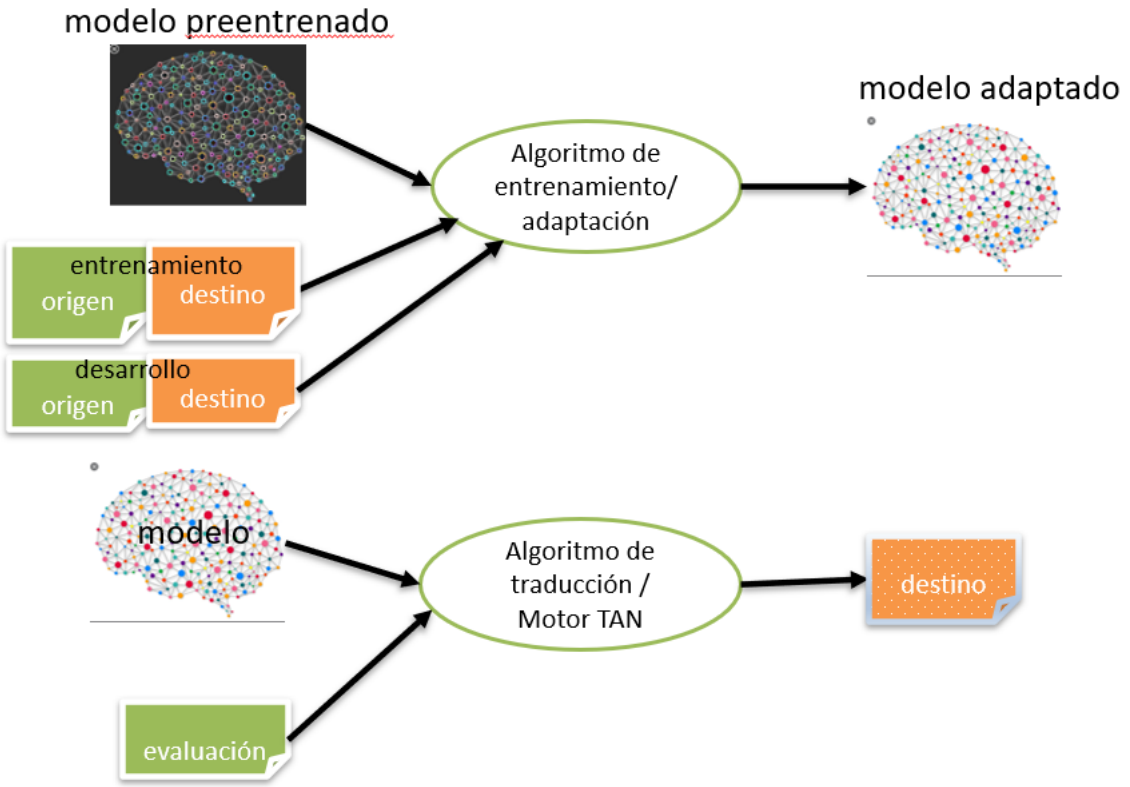
Adaptar a un ***dominio*** (*fine tunning*) un modelo (cerebro artificial) para traducir:

- Arquitectura *Transformer*
- Aprendizaje supervisado
- Inglés→ español
- Corpus: 950M de oraciones
- De la Universidad de Helsinki
- En Hugging Face (junto a los modelos de las grandes compañías)

La inteligencia artificial generativa en la traducción de idiomas

Adaptación y traducción con un modelo: entrada y salida

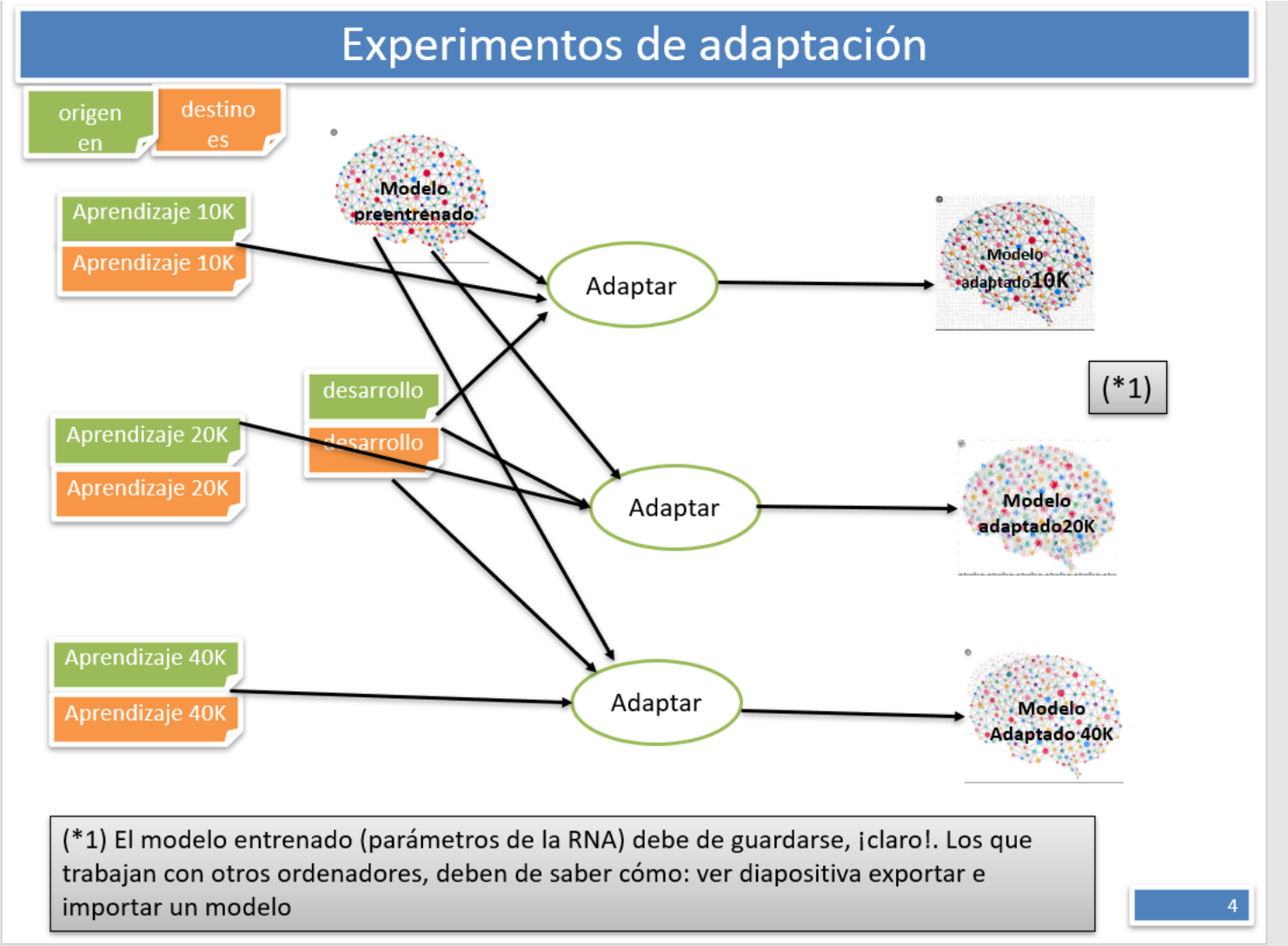
- Algoritmo entrenando/adaptando modelo y motor de TAN traduciendo



3

Práctica sobre traducción automática neuronal

La inteligencia artificial generativa en la traducción de idiomas



Práctica sobre traducción automática neuronal

La inteligencia artificial generativa en la traducción de idiomas

Cómo se explica y cómo se aplica en el Grado en Traducción e Interpretación

¡Muchas gracias!

Emilio Rodríguez Vázquez de Aldana

aldana@usal.es