

OpenCourseWare  
**Procesamiento de Lenguaje Natural con  
Aprendizaje Profundo,**  
Máster en Ciencia y Tecnología Informática

**Tema 7.3 - Desarrollo de aplicaciones PLN basadas  
en aprendizaje profundo. Generación de resúmenes**

# Índice

- ¿Qué es generación de resúmenes?
- Tipos
- Fases
- Métricas de evaluación

# Generación de Resúmenes

- Objetivo: Producir un resumen que sintetice la información más relevante de un texto o conjunto de textos de entrada.
- La entrada podría ser un único texto o bien una colección de textos que debe ser resumida en un único texto de salida (resumen).
- Ejemplos de aplicaciones:
  - Resumen de la historia clínica de un paciente.
  - Simplificación de un texto en una versión más sencilla.
  - Resumen de emails, artículos, etc...

# Índice

- ¿Qué es generación de resúmenes?
- **Tipos**
- Fases
- Métricas de evaluación

# Tipos: extractiva y abstractiva

- **Extractiva:** el resumen es creado a partir de palabras y oraciones que aparecen en el texto de entrada.
- **Abstractiva:** el resumen generado es un texto nuevo que sintetiza la idea del texto original.
- Ejemplo:
  - **Texto original:** *Yolanda Díaz en un acto de Sumar ha afirmado que existe una elite mundial preparando planes alternativos para poder huir del mundo en el caso de que que nos vayamos al “carajo”.*
  - **Resumen extractivo:** *Yolanda Díaz ha afirmado que existe una elite mundial preparando planes para poder huir del mundo.*
  - **Resumen abstractivo:** *Según Yolanda Díaz, existe una elite mundial que planea huir si el mundo se va al carajo.*

# Índice

- ¿Qué es generación de resúmenes?
- Tipos
- **Fases**
- Métricas de evaluación

# Fases

1. Seleccionar las oraciones más relevantes (scoring).
2. Ordenar dichas oraciones para formar el resumen.
3. Simplificación de las oraciones seleccionadas.

# Fase 1: Posibles enfoques para seleccionar las oraciones

- Enfoque no supervisado (basado en tf-idf)
  - Divide el texto en oraciones.
  - Tokenización.
  - Calcula la frecuencia y tf-idf para cada palabra.
  - Calcula un score para cada oración basado en la (suma) frecuencia de las palabras.
  - Selecciona las oraciones con mayor score.



# Fase 1: Posibles enfoques para seleccionar las oraciones

- Enfoque no supervisado (basado en embeddings)
  - Divide el texto en oraciones.
  - Obtén un embedding para cada oración (BERT, word2vec).
  - Entrena el algoritmo K-means para agrupar los embeddings en K clusters (k es el número de oraciones en el resumen final).
  - Calcula el centroide de cada cluster, y selecciona la oración más próxima a dicho centroide.
  - Forma el resumen con las k oraciones.

# Fase 1: Posibles enfoques para seleccionar las oraciones

- Enfoque supervisado (solo para extractivo) :
  - Corpus de entrenamiento: texto original + resumen.
  - Cada oración del texto original se etiqueta como 1 (si aparece en el resumen) o 0 eoc.
  - Entrenar un clasificador binario (ej. SVM).
  - Selecciona las oraciones con la probabilidad mayor proporcionada por el algoritmo.

# Fase 1: Posibles enfoques para seleccionar las oraciones

- Enfoque supervisado (basado en Seq2Seq)
  - Corpus de entrenamiento: texto + resumen.
  - Entrena un modelo Seq2Seq (basado en transformers), la entrada es el texto original, la salida es el resumen.
  - Entrena el modelo con métricas como ROUGE.

# Ejemplo

## Document

London Welsh have announced a contract extension for their former England international back Olly Barkley. The 33-year-old joined the Exiles last year, having previously played for Bath, Gloucester, Racing Metro, Grenoble and the Scarlets. He won 23 Test caps, the last of which was against New Zealand in 2008. Olly Barkley has extended his stay at London Welsh after signing a new contract at the relegated-club. Welsh are set to return to the Championship next season after finishing bottom of the Aviva Premiership. ' It has not been an easy year, but I am committed to seeing my journey with London Welsh through,' Barkley said. 'I am really looking forward to working with our new team and the reinvigorated coaching regime, which is being put together, to put this season behind us and lay a strong foundation for promotion back to the Premiership for the 2016-17 season.' The fly half featured 23 times for England and previously played for Bath, Gloucester and Racing Metro .

## Summary ( $top\_k = 3$ )

London Welsh have announced a contract extension for their former England international back Olly Barkley. Welsh are set to return to the Championship next season after finishing bottom of the Aviva Premiership. ' The fly half featured 23 times for England and previously played for Bath, Gloucester and Racing Metro .

# Ejemplo

## Document

London Welsh have announced a contract extension for their former England international back Olly Barkley. The 33-year-old joined the Exiles last year, having previously played for Bath, Gloucester, Racing Metro, Grenoble and the Scarlets. He won 23 Test caps, the last of which was against New Zealand in 2008. Olly Barkley has extended his stay at London Welsh after signing a new contract at the relegated-club. Welsh are set to return to the Championship next season after finishing bottom of the Aviva Premiership. 'It has not been an easy year, but I am committed to seeing my journey with London Welsh through,' Barkley said. 'I am really looking forward to working with our new team and the reinvigorated coaching regime, which is being put together, to put this season behind us and lay a strong foundation for promotion back to the Premiership for the 2016-17 season.' The fly half featured 23 times for England and previously played for Bath, Gloucester and Racing Metro.

Olly Barkley renews his contract with London Welsh. After the failure in the Aviva Premiership, Welsh will return to the Championship next season. Barkley has played 23 times for England.

# Fase 2: ¿Cómo ordenar las oraciones?

- Si es un único texto, respetar el orden en el texto original.
- Si es multi-documento:
  - orden cronológico (fecha de cada documento)
  - usar distancia coseno para agrupar oraciones similares.
  - usar distancia coseno para agrupar oraciones que discuten una misma entidad.

# Fase 3: ¿Cómo simplificar las oraciones?

- Eliminar estructuras apositivas:
  - *Ana Obregón, ~~una de las mujeres más queridas y reconocidas de nuestro país,~~ nos ofrece un desgarrador testimonio sobre la pérdida de su hijo.*
- Eliminar cláusulas de atribución:
  - *Ana podría estar pensando en tener más hijos, ~~según dijeron el martes varios periodistas en el programa de Ana Rosa.~~*

# Fase 3: ¿Cómo simplificar las oraciones?

- Eliminar adverbios de inicio:
  - ~~Por otro lado,~~ *la artista sigue feliz con su hija en Miami*
- Eliminar sintagmas preposicionales sin entidades:
  - *Ana charló sobre su libro,* ~~con varios periodistas.~~



# Fase 3: ¿Cómo simplificar las oraciones?

- Reemplazar palabras por sinónimos (para enfoque abstractivo):
  - *Ana Obregón nos ofrece un desgarrador testimonio sobre la pérdida de su hijo -> Ana Obregón nos da un **doloroso relato** sobre la **muerte** de su hijo.*
- Paráfrasis (para abstractivo):
  - *Ana Obregón y Alejandro Lequio perdieron a su hijo en 2022 -> El hijo de Ana Obregón y Lequio falleció en 2002.*

# Índice

- ¿Qué es generación de resúmenes?
- Tipos
- Fases
- **Métricas de evaluación**

# Métricas de evaluación: rouge

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation).
- Indica cuánto texto del resumen de referencia es capturado en el resumen generado por el sistema. Se calcula como
  - número de palabras en común que tienen el resumen generado y resumen de referencia, entre el número total de palabras en el resumen de referencia.

# Métricas de evaluación: Rouge

- Veamos un ejemplo:
  - Resumen de referencia: *The cat was under the bed.*
  - Resumen generado: *The cat was found under the bed.*
  - El valor de ROUGE es  $6 / 6 = 1$ .

# Métricas de evaluación: rouge precisión

- También vamos a calcular su **precisión**, para determinar cuánta de la información del resumen generado es relevante o necesaria. Se calcula:
  - número de palabras en común en el resumen generado y de referencia dividido por el número de palabras en el resumen generado.

# Métricas de evaluación: rouge precisión

- Veamos un nuevo ejemplo:
  - Resumen de referencia: *The cat was under the bed*
  - Resumen generado: *The tiny little cat was found under the big bed*
  - Recall =  $6/6 = 1$
  - La precisión será:  $6 / 9 = 0.66 \Rightarrow$  nos indica que el resumen contiene muchas palabras que no son relevantes.
  - **Solución:** usar F1 para elegir el mejor resumen generado.

# Métricas de evaluación: rouge

- Y qué pasaría con el siguiente ejemplo?:
  - Resumen de referencia: *The cat was under the bed.*
  - Resumen generado: *The bed was under the cat.*
  - Recall = 6/6
  - Precision = 6/6
- **Rouge tiene limitaciones.** No puede medir la coherencia semántica, únicamente está basada en el número de palabras en común.

# Métricas de evaluación: variaciones de Rouge

- ROUGE-N: mide cuantos unigramas, bigramas, trigramas y n-gramas están presentes en el resumen de referencia y el generado.
- ROUGE-L — computa la secuencia más larga de palabras que se solapan en el resumen de referencia y generado.
- ROUGE-S: dado un n-grama, genera todos sus posibles bigramas (aunque no sean consecutivos). Por ejemplo: “*cat in the hat*”, los skip-bigramas serían “*cat in, cat the, cat hat, in the, in hat, the hat*”.



# Métricas de evaluación: librerías para Rouge

- <https://pypi.org/project/rouge/>
- <https://huggingface.co/spaces/evaluate-metric/rouge>

OpenCourseWare  
Procesamiento de Lenguaje Natural con  
Aprendizaje Profundo,

**Gracias!!!**

<https://github.com/iseaura>