Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 27-39 recibido 25-11-2020 revisado 28-01-2021 aceptado 02-02-2021

Evaluación de un modelo transformador aplicado a la tarea de generación de resúmenes en distintos dominios

Evaluation of a transformer model applied to the task of text summarization in different domains

Isabel Segura-Bedmar, Lucía Ruz, Sara Guerrero-Aspizua
Universidad Carlos III de Madrid, Leganés, Spain
isegura@inf.uc3m.es, lruz@pa.uc3m.es, sguerrer@ing.uc3m.es

Resumen: En los últimos años, las técnicas de deep learning han supuesto un gran impulso tecnológico en muchas de las tareas de Procesamiento de Lenguaje Natural (PLN). La tarea de generación de resúmenes también se ha beneficiado de estas técnicas, y en los últimos años se han implementado distintos modelos, logrando superar los resultados del estado de la cuestión. La mayoría de estos trabajos han sido evaluados en colecciones de textos periodísticos. Este artículo presenta un trabajo preliminar donde aplicamos un modelo transformador, BART, para la tarea de generación de resúmenes y lo evaluamos en varios datasets, uno de ellos formado por textos del dominio biomédico.
Palabras clave: Generación de resúmenes, Transformadores.

Abstract: In recent years, deep learning techniques have provided a significant technological advance in many Natural Language Processing (NLP) tasks. Text summarization has also benefited from these techniques. Recently, several deep learning approaches have been implemented, surpassing the previous state of the art performances. Most of these works have been evaluated on collections of journalistic texts. This article presents a preliminary work where we apply a transforming model, BART, for text summarization. The model is evaluated on several datasets, one of them consisting of texts from the biomedical domain.
Keywords: Text summarization, Transformers.

1 Introducción

En la era del Big Data, existe una gran cantidad de información en formato digital, la mayor parte de ella en texto no estructurado. Este gran volumen de datos nos aporta conocimiento, pero necesariamente primero nos enfrenta al reto de poder procesar y comprender toda esta información. Un resumen nos proporciona las ideas principales de un texto, por tanto, la generación de resúmenes puede ayudarnos a la hora de acceder a la información esencial de un texto o colección de textos. Dado que la generación de resúmenes de manera manual es una tarea compleja que conlleva mucho tiempo y recursos, las técnicas de Procesamiento de Lenguaje Natural (PLN) pueden ayudar a reducir esta carga de trabajo.

En la generación automática de resúmenes, se distinguen dos enfoques: extractivo y abstractivo. El enfoque extractivo consiste en la selección de las oraciones más relevantes del texto para formar el resumen. Por otro lado, el método abstractivo genera nuevas oraciones para la creación del resumen. En los últimos años, las técnicas de deep learning han supuesto un gran impulso tecnológico en muchas de las tareas de PLN (Beloki et al., 2020; Miranda-Escalada y Segura-Bedmar, 2020; Colón-Ruiz, Segura-Bedmar, y Martínez, 2019; Poncelas et al., 2019), en muchos casos, obteniendo mejores resultados que los algoritmos clásicos de aprendizaje automático. La generación automática de resúmenes también se ha beneficiado de esta tecnología, y en los últimos años, muchos de los sistemas han utilizado técnicas de deep learning como las redes convolucionales (Zhang et al., 2019) o las redes recurrentes (Nallapati, Zhai, y Zhou, 2017). En 2017, se propone un nuevo modelo de deep learning, Transformers(Vaswani et al.,

ISSN 1989-7553. DOI 10.26342/2021-66-2

2017), cuya principal ventaja es que no necesita procesar los textos de forma secuencial, como hacen las redes recurrentes. Esto facilita la paralelización de la tarea y reduce los tiempos de entrenamiento. Estos modelos también han sido aplicados a la tareas de generación de resúmenes (Liu y Lapata, 2019; Zhang, Wei, y Zhou, 2019), con buenos resultados. Sin embargo, la mayoría de estos trabajos han sido evaluados sobre las colecciones de artículos periodísticos, como por ejemplo el corpus CNN/Daily Mail.

En este trabajo, abordamos la tarea de generación de resúmenes utilizando el modelo BART (Lewis et al., 2020), basado en la arquitectura Sequential-to-Sequential. Además, el modelo será evaluado sobre conjuntos de datos distintos. Este artículo presenta un trabajo preliminar, cuyo principal objetivo será comprobar si el modelo BART es capaz de generar resúmenes automáticos para textos de otros estilos distintos al periodístico, como son los textos biomédicos. Además, comprobaremos si BART es capaz de generar resúmenes de calidad a partir de conjuntos de datos, cuyas características en relación al número de instancias y al tamaño medio de sus textos y resúmenes son distintas a las del corpus CNN/DailyMail, que fue utilizado para preentrenar el modelo BART.

El artículo está organizado como sigue: la sección 2 revisa los modelos más recientes aplicados a la tarea de generación de resúmenes. En la sección 3, presentamos el modelo aplicado en este trabajo. La sección 4 describe los datasets utilizados en la evaluación del modelo. Además, presenta y discute los resultados para cada uno de los datasets. Finalmente, las principales conclusiones y líneas de trabajo futuro serán descritas al detalle en la sección 5.

2 Estado de la cuestión

Como se ha indicado en la introducción, el método transformador es un modelo de aprendizaje profundo que fue creado en el año 2017 (Vaswani et al., 2017). Es un modelo diseñado para el procesamiento de datos secuenciales, y, por tanto adecuado para abordar tareas del PLN, como la clasificación de textos, la generación de resúmenes, entre otras muchas. A diferencia del modelo de redes neuronales recurrentes (RNN), esta arquitectura procesa los datos de manera paralela, sin necesidad de ser procesados en orden. Gracias a este tipo de procesamiento se disminuyen notablemente los tiempos de entrenamiento. Este menor coste computacional, sumado a los excelentes resultados obtenidos por los sistemas que utilizan el modelo, ha revolucionado el campo del PLN, reemplazando a las RNN y a otros modelos de aprendizaje automático profundo.

La investigación en generación de resúmenes no ha permanecido ajena al auge de los modelos transformadores. Desde el 2015, distintas arquitecturas de redes profundas, como por ejemplo, las RNN han sido aplicadas para la generación automática de resúmenes (Nallapati, Zhai, y Zhou, 2017; Nallapati et al., 2016), estableciendo el estado de la cuestión en torno al 30 % de la media de las métricas Rouge-1, Rouge-2 y Rouge-L, sobre el corpus CNN/Daily Mail. A continuación, revisamos los trabajos más recientes sobre generación de resúmenes basados en modelos transformadores.

Zhong et al. (2019) presentan un primer estudio para la generación de resúmenes extractivos mediante redes de neuronas profundas, entre ellas la red recurrente Long short-term memory (LSTM) (Hochreiter y Schmidhuber, 1997), y BERT (Devlin et al., 2019), un modelo transformador. Los experimentos fueron realizados sobre dos datasets: CNN/Daily Mail (Hermann et al., 2015), descrito en el apartado 4.1, y Newsroom (Grusky, Naaman, y Artzi, 2018) formado por 1.3 millones de artículos y resúmenes extraídos de 38 fuentes de noticias a lo largo de los últimos 20 años. En este trabajo, los autores se centraron únicamente en las siete fuentes que proporcionaban un mayor número de ejemplos: New York Times (NYT), Washington Post, Fox News, The Guardian, New York Times Daily News, The Wall Street Journal (WSJ) y USA Today. En los experimentos, cada modelo es evaluado sobre cada fuente de forma separada.

Con respecto a los resultados obtenidos para el dataset CNN-Daily Mail, LSTM proporcionaba mejores resultados para las métricas Rouge-1 (41.56 %) y Rouge-L (37.83 %), mientras que el modelo transformador obtenía los valores más altos para Rouge-2 (18.85 %). Respecto a la evaluación con las distintas fuentes del dataset Newsroom, LSTM obtiene mejores resultados para New York Times (25.27 % de media), Washington Post (18.89 % de media) y Fox News (56.17 %

de media). El modelo transformador supera al modelo LSTM en el resto de fuentes. Para todos los datasets, no hay una diferencia significativa entre los resultados de ambos modelos.

El sistema descrito en (Liu y Lapata, 2019) utiliza el modelo pre-entrenado BERT para la generación de resúmenes extractivos y abstractivos. La principal contribución del trabajo es que la codificación de los textos se hace a nivel de documento, en lugar de a nivel de oración, como habían hecho trabajos anteriores. La evaluación es realizada sobre los datasets CNN-Daily Mail (Rouge-1=42.13 %, Rouge-2=19.60 %, Rouge-L=39.18 %), NYT (Rouge-1=49.02 %, Rouge-2=31.02 %, Rouge-L=45.55 %) y XSum (Narayan, Cohen, y Lapata, 2018) (Rouge-1=38.81 %, Rouge-2=16.50 %, Rouge-L=31.27 %).

El sistema también fue evaluado por un conjunto de usuarios. Para la evaluación humana se les aporta a los participantes el resumen generado por el sistema y por otros sistemas de generación de resúmenes (Narayan, Cohen, y Lapata, 2018; See, Liu, y Manning, 2017; Gehrmann, Deng, y Rush, 2018). A continuación, se les realiza una serie de preguntas para comprobar si el resumen contiene toda la información necesaria. A mayor número de preguntas contestadas correctamente, mejor evaluación. La evaluación basada en el usuario también muestra que el sistema proporciona mejores resúmenes que el resto de modelos comparados en el trabajo.

Zhang, Wei, y Zhou (2019) proponen un sistema para la generación de resúmenes extractivos basado en el uso del modelo transformador jerárquico (HIBERT). En lugar de utilizar alguno de los modelos de lenguajes ya proporcionados por BERT (Devlin et al., 2019), los autores entrenan su propio modelo de lenguaje. Con este objetivo crean un dataset, GIGA-CM, compuesto por 6,626,842 documentos, la mayor parte de ellos obtenidos del dataset GigaWord (Graff et al., 2003) y una pequeña parte extraídos del conjunto de entrenamiento del dataset CNN/Daily Mail. En total, el dataset contiene 2,854 millones de palabras. Los autores utilizan una técnica de enmascaramiento similar a la propuesta por los creadores de BERT (Devlin et al., 2019), pero en este caso se enmascaran todas las palabras de las oraciones seleccionadas. Una vez entrenado su propio modelo de lenguaje, en una segunda fase dedicada a adaptar el modelo para la tarea de generación de resúmenes, los autores utilizan los datasets CNN/Daily Mail y New York Times (NYT) (Xu y Durrett, 2019; Durrett, Berg-Kirkpatrick, y Klein, 2016).

A la hora de realizar la evaluación, comparan su sistema con otros enfoques propuestos por otros autores, que también han sido evaluados sobre los mismos datasets. Los modelos extractivos (Gehrmann, Deng, y Rush, 2018; Dong et al., 2018; Cheng y Lapata, 2016) con los que se comparan están basados en el modelo transformador, pero en lugar de entrenar su propio modelo de lenguaje, han reutilizado alguno de los modelos de lenguajes proporcionados por BERT. Respecto a los modelos abstractivos (See, Liu, y Manning, 2017; Celikyilmaz et al., 2018) con los que se comparan, están basados en modelos secuenciales con refuerzo. Los experimentos mostraron que HIBERT es computacionalmente más eficiente que aquellos sistemas que utilizaban modelos de lenguajes pre-entrenados con BERT. Respecto a los resultados, HIBERT es capaz de superar a todos los modelos estudiados en ambos datasets. Las métricas para el dataset CNN/Daily Mail son Rouge-1 de 43.19 %, Rouge-2 de 20.46 % y Rouge-L de 39.72 %.

En el dataset NYT, los resultados son aún mejores (Rouge-1=49.47 %, Rouge-2=30.11 %, Rouge-L=41.63 %). Los autores también realizaron una evaluación basada en usuarios sobre una muestra de 20 documentos elegidos aleatoriamente, junto con los resúmenes generados por el modelo HIBERT y por los modelos comparados en este estudio. Los usuarios debían clasificar de peor a mejor los resúmenes generados, teniendo en cuenta si el resumen captaba la información esencial y si era correcto gramaticalmente. El resumen generado por el modelo HIBERT es seleccionado como el mejor en el 30 % de los casos.

En un trabajo posterior, Zhang et al. (2019) proponen, además de utilizar BERT en el codificador, utilizar el modelo transformador en el decodificador. Su decodificador está diseñado en dos fases. En la primera fase, genera un resumen borrador usando el algoritmo *left-context-only-decoder*, que únicamente tiene en cuenta el contexto izquierdo de cada palabra. La segunda fase tiene co-

mo objetivo refinar el resumen borrador generado en la fase anterior. Para ello, se aplica una máscara a cada palabra, conservando ahora sus respectivos contextos al completo. Además, los autores aplican el aprendizaje por refuerzo. Este tipo de aprendizaje pretende premiar a aquellas acciones óptimas y penalizar a las menos válidas. En este artículo, proponen como objetivo obtener la mayor evaluación posible mediante Rouge, es decir, a mayor puntuación mayor refuerzo positivo. El sistema es evaluado sobre el dataset CNN/Daily Mail y comparado con otros trabajos anteriores (See, Liu, y Manning, 2017; Shi et al., 2019; Chen y Bansal, 2018; Hsu et al., 2018). En el estudio, el sistema es capaz de obtener mejores resultados que dichos sistemas (Rouge-1=41.71 %, Rouge-2=19.49 % y Rouge-L=38.79 %).

Al igual que en el trabajo (Zhang et al., 2019), (Bae et al., 2019) también aplica aprendizaje por refuerzo para la selección de las oraciones más relevantes de un texto. Después de esta selección, los autores proponen una arquitectura que utiliza BERT en el codificador y una capa LSTM en el decodificador. En una segunda fase, las oraciones seleccionadas son parafraseadas mediante un modelo Sequence-to-Sequence con atención. El sistema fue evaluado sobre el dataset CNN/Daily Mail (Rouge-1=41.90 %, Rouge-2=19.08 %, Rouge-L=39.64 %), sin superar al trabajo descrito en (Zhang et al., 2019).

Los trabajos revisados muestran que los modelos transformadores han sido capaces de establecer un nuevo estado de la cuestión en la tarea de generación de resúmenes, superando a los resultados obtenidos por trabajos anteriores basados en otras arquitecturas de redes profundas (Nallapati, Zhai, y Zhou, 2017; Nallapati et al., 2016). En concreto, HIBERT (Zhang, Wei, y Zhou, 2019) es el sistema que ha obtenido mejores resultados sobre el dataset CNN/Daily Mail, con una media de Rouge de 34.5 %- El corpus CNN/Daily Mail es el dataset utilizado por la mayoría de los sistemas para su evaluación, aunque algunos trabajos también han utilizado otros datasets tales como New York Times o XSum. Todos los datasets tienen en común que son colecciones de artículos de prensa.

3 Enfoque

El objetivo de nuestro trabajo es explorar el uso del modelo BART (Lewis et al., 2020) para la tarea de generación automática de resúmenes.

BART es un modelo basado en la arquitectura Sequence-to-Sequence (Seq2Seq), cuyo principal objetivo es la transformación de una secuencia de entrada en otra secuencia de salida. La arquitectura Seq2Seq es especialmente útil en muchas tareas de PLN, como la traducción automática o la generación de resúmenes. En ambas tareas, la entrada es una secuencia de palabras. En el caso de la traducción automática, la secuencia de salida será la traducción de la secuencia de entrada a otro idioma. En el caso de la generación de resúmenes, la salida será una secuencia de palabras que constituya el resumen de la secuencia de entrada.

Los dos principales componentes de una arquitectura Seq2Seq son el codificador (encoder) y el decodificador (decoder). El primero toma la secuencia de entrada y la transforma en un vector. El decodificador transforma dicho vector en la secuencia de salida. Tradicionalmente, estos componentes han sido implementados con redes recurrentes, y en particular, con LSTM. Estas redes procesan la entrada de forma secuencial. Así, para acceder a la célula que representa la última palabra, es necesario recorrer las anteriores. Cuando la secuencia es muy larga, puede ocurrir que el modelo olvide la información contenida en las primeras células. Además, el procesamiento secuencial implica un alto coste computacional y no permite paralelizar el aprendizaje del modelo.

El modelo transformador, presentado por Vaswani et al. (2017), es una alternativa a las redes recurrentes para implementar el codificador y el decodificador de una arquitectura Seq2Seq. Un transformador está basado en la técnica de atención (attention mechanism), que permite establecer las dependencias entre las secuencias de entrada y de salida. El codificador se encarga de representar cada posición y aplicar el mecanismo de atención para conectar palabras que no son consecutivas. El mecanismo de atención asigna una puntuación a cada palabra y compara todas las puntuaciones en la secuencia. Esto permite determinar la contribución de cada una de las palabras. Este mecanismo puede ser paralelizado, acelerando así el aprendizaje. Por tanto, esta técnica es capaz de decidir qué partes de la secuencia de entrada son las más importantes en cada paso. Mientras que

el módulo LSTM lee la entrada de forma secuencial, la principal ventaja del mecanismo de atención es que es capaz de procesar de una sola vez el contexto de cada palabra y utilizarlo para asignar más peso a las partes de la secuencia de entrada más importantes. De esta forma, el decodificador sabe identificar qué partes de la secuencia son más importantes.

Como se ha dicho anteriormente, BART sigue una arquitectura Seq2Seq basada en transformadores. BART combina dos arquitecturas como BERT y GPT (Generative Pre-trained Transformer) (Radford et al., 2018). El codificador de BART implementa un modelo BERT, mientras que el decodificador implementa un modelo GPT. En la fase de codificación, BART primero introduce ruido en la secuencia de entrada. Para ello utiliza distintas técnicas como el enmascaramiento de palabras y de secuencias de palabras, el borrado de palabras o la permutación de oraciones, entre otras. Una vez que la secuencia de entrada ha sido transformada en una secuencia con ruido, BART trata de entrenar un modelo capaz de reconstruir la secuencia de entrada. En la fase de decodificación, BART utiliza un modelo GPT que genera los tokens de izquierda a derecha.

El modelo BART está implementado en la librería Hugging Face.[1] Este modelo ha sido preentrenado sobre el dataset CNN/Daily Mail. Con el objetivo de facilitar la replicabilidad de nuestra experimentación, nuestra implementación está disponible en el siguiente repositorio de https://github.com/isegura/sepln2021-textsummarization.

4 Evaluación

4.1 Datasets

Como se ha visto en el estado de la cuestión, la mayoría de los sistemas de generación de resúmenes han sido evaluados sobre el corpus CNN/Daily Mail. Uno de los objetivos del trabajo actual es extender la evaluación a otros datasets menos utilizados, e incluir además un nuevo dataset, como es BioMRC(Pappas et al., 2020a), cuyos textos no son artículos de periódicos, sino documentos científicos. Esto nos permitirá determinar si el modelo, que ha sido pre-entrenado con otro modelo de lenguaje generado con textos periodísticos, es capaz de obtener resultados similares sobre textos de otros dominios.

Por tanto, para evaluar nuestro modelo, los experimentos han sido realizados sobre cuatro conjuntos de datos: CNN/Daily Mail (See, Liu, y Manning, 2017) (versión no anonimizada), GigaWord (Rush, Chopra, y Weston, 2015), XSum (Narayan, Cohen, y Lapata, 2018) y BioMRC (Pappas et al., 2020b). Mientras que los tres primeros datasets constan de artículos periodísticos online, el último, BioMRC, es una colección de textos del dominio biomédico. Los cuatro conjuntos de datos están formados por textos escritos en inglés. Todos los resúmenes son abstractivos, es decir, cada resumen se ha creado a partir de oraciones nuevas y no extraídas del texto original.

Los cuatro datasets son distribuidos con la librería Hugging Face, que también usaremos para implementar nuestro enfoque. Para el dataset BioMRC, Hugging Face proporciona tres versiones dependiendo del tamaño: *large*, *small* y *tiny*. Dentro de cada tamaño, se distinguen dos tipos: A y B. La principal diferencia es que los textos de B se han limpiado para eliminar ruido. En nuestro trabajo, hemos seleccionado la versión *large_B*.

La tabla 1 muestra el número de textos de cada conjunto de datos y el tamaño medio de sus textos de entrada y sus resúmenes, que se define por el número de tokens de un texto. GigaWord, una colección de artículos periodísticos, es el dataset con mayor número de instancias. Sin embargo, su tamaño medio del texto de entrada y del resumen es el segundo más pequeño (poco más de 8 palabras por resumen). La relación entre el tamaño del texto y su resumen es aproximadamente de 3.7. El segundo dataset con mayor número de instancias es BioMRC, formado por textos del dominio biomédico. Aunque es el segundo corpus con un tamaño medio de los textos de entrada, el tamaño medio de sus resúmenes es el más pequeño, con menos de 7 palabras por resumen. En este caso, el ratio entre el tamaño medio de sus textos y sus resúmenes es de 37.7.

El tercer dataset con mayor número de instancias es CNN/Daily Mail, que como se dijo anteriormente es uno de los datasets más utilizados en la evaluación de generación de resúmenes. Sus textos y resúmenes son los que tienen un mayor tamaño, con una ratio de 13.9. La principal ventaja de este data-

[1] https://huggingface.co/

set respecto al resto, es que sus textos han sido utilizados para entrenar el modelo del lenguaje utilizado en la implementación del modelo BART, que usamos en la experimentación, y que es proporcionado por Huggging Face. XSum es el dataset con menor número de instancias. Sin embargo, también es el dataset con el segundo tamaño medio de longitud de textos de entrada. El ratio entre el tamaño medio de sus textos y resúmenes es de 35.9.

Los cuatro datasets se dividen en conjuntos de entrenamiento, validación y test.

En nuestra experimentación, hemos utilizado la configuración que por defecto proporciona la librería Hugging Face, por tanto, no ha sido necesario utilizar el conjunto de validación para ajustar los parámetros de nuestro modelo. Por este motivo, los textos del conjunto de validación también han sido incluidos en el conjunto de entrenamiento de cada dataset para ajustar el modelo a las tareas de generación de resúmenes.

4.2 Resultados

El método propuesto ha sido evaluado en diversas variantes de la métrica Rouge (Lin, 2004) y con la métrica Bleu (Papineni et al., 2002) para todos los datasets descritos en el apartado 4.1.

La métrica Rouge es la más utilizada para evaluar la generación automática de resúmenes. Podemos encontrar 5 variantes: Rouge-N, Rouge-L, Rouge-W, Rouge-S y Rouge-Su. En este trabajo, sólo utilizaremos las métricas Rouge-N (en particular, para N=1 y N=2) y Rouge-L, porque son las métricas reportadas por la mayoría de los artículos.

Rouge-N mide la superposición de N-gramas entre el resumen generado y el resumen del dataset. Así para N=1, la métrica se refiere a la superposición de palabras, mientras que para N=2, nos referimos a la superposición de bigramas. Rouge-L es una métrica que mide la subsecuencia común más larga (LCS) entre el resumen original y el resumen generado por el sistema. Una descripción detallada de estas métricas y del resto de variantes de Rouge puede encontrarse en el artículo propuesto por Lin (2004).

Como se ha indicado anteriormente, también utilizaremos la métrica Bleu (Papineni et al., 2002). Bleu mide el número de n-gramas del resumen automático que están presentes en el resumen de referencia, mientras que Rouge contabilizael número de n-gramas del resumen de referencia que están presentes en el resumen automático. Bleu similar a Rouge, pero su principal diferencia es que introduce un nuevo parámetro, denominado penalización por brevedad (PB), encargado de sancionar aquellas predicciones cuya longitud sea menor que la del resumen original. La penalización por brevedad se calcula a través de la siguiente función:

$$PB = \begin{cases} 1 & si \quad c > r \\ e^{(1-r/c)} & si \quad c \leq r \end{cases} \tag{1}$$

En la fórmula anterior, c es la longitud de la predicción y r la longitud del resumen de referencia. De esta manera, la métrica Bleu se calcula de la siguiente manera:

$$BLEU = PB \cdot exp\left(\sum_{n=1}^{N} w_n log P_n\right) \tag{2}$$

En este caso, $w_n = 1/N$, correspondiente al peso de cada n-grama de longitud N. P es la precisión de los n-gramas, es decir, es el ratio entre el número de n-gramas comunes y el número n-gramas presentes del resumen generado.

La tabla 2 muestra los resultados de nuestro modelo respecto a las métricas anteriormente descritas y para cada uno de los datasets propuestos para nuestro estudio.

Como se ha dicho anteriormente, nuestro enfoque está basado en el uso del modelo BART, proporcionado por la librería Hugging Face. En este modelo, el modelo del lenguaje ha sido pre-entrenado con los textos del dataset CNN/Daily Mail. Posteriormente, en la segunda fase (fine tuning), el modelo es ajustado para la tarea de generación de resúmenes, utilizando para ello los conjuntos de entrenamiento de cada dataset.

Comenzamos discutiendo los resultados para la métrica Rouge-1, que mide la superposición de las palabras del resumen generado y su respectivo resumen de referencia. Aunque el modelo de lenguaje ha sido construido a partir de los textos de CNN/Daily Mail, los mejores resultados para Rouge-1 se consiguen sobre el corpus BioMRC, siendo este el único dataset compuesto por textos no periodísticos, sino de dominio biomédico. Por tanto, el estilo narrativo y el léxico de los textos no parecen afectar al modelo en

Dataset	Conjunto de entrenamiento	Conjunto de validación	Conjunto de Test	Tamaño textos	Tamaño resúmenes
CNN/Daily	287.113	13.368	11.490	781	56
GigaWord	3.803.957	189.651	1.951	31,4	8,3
XSum	204.017	11.327	11.333	431	12
BioMRC	700.000	50.000	62.707	254,01	6,72

Tabla 1: Descripción de los datasets utilizados en este trabajo.

Datasets	Rouge-1	Rouge-2	Rouge-L	BLEU
CNN/Daily	38.48	**17.74**	**37.52**	31.35
GigaWord	24.93	9.88	24.02	37.32
XSum	32.2	9.7	27.11	**66.50**
BioMRC	**39.58**	13.08	34.28	57.73

Tabla 2: Resultados.

lo que se refiere a la superposición de palabras. El segundo mejor resultado, con una diferencia únicamente de 1.1, se obtiene para el dataset CNN/Daily Mail, con una Rouge-1 de 39.58 %, más de 3.5 por debajo del estado de la cuestión en dicho dataset (Rouge-1=43.19 % en el trabajo (Zhang, Wei, y Zhou, 2019)).

Para los otros dos datasets, GigaWord y XSum, los resultados de Rouge-1 son significativamente más bajos que los obtenidos para BioMRC. Es de esperar que los resultados obtenidos con el dataset CNN/Daily Mail sean superiores a la de estos datasets, debido a que el modelo de lenguaje fue entrenado utilizando el dataset CNN/Daily Mail. Sin embargo, llama la atención que el modelo obtenga la mejor Rouge-1 para la colección de textos biomédicos, y no de estilo periodístico. Un tamaño pequeño en los resúmenes de referencia (con sólo 6.5 palabras en el caso de BioMRC) parece tener un efecto positivo sobre la métrica Rouge-1. Otra posible razón de los buenos resultados obtenidos sobre BioMRC es que, como se explicó anteriormente, el dataset BioMRC ha sido limpiado para reducir ruido, y eso podría redundar en una mejor calidad de los textos y resúmenes de referencia.

Respecto a Rouge-2, que se refiere al número de bigramas que comparten el resumen generado con el resumen de referencia, como era previsible, obtiene los mejores para el dataset CNN/Daily Mail, por lo ya explicado anteriormente. Respecto al estado de la cuestión, el modelo consigue una puntuación menor que el sistema HIBERT, con una Rouge-2 de 20.46 %. La diferencia de Rouge-2 respecto a los otros tres datasets es muy significativa, llegando alcanzar una diferencia de hasta 8 puntos en el caso del dataset XSum. El modelo pre-entrenado con textos de estilo periodístico debería proporcionar mejores resultados para los corpus GigaWord y XSum, que para el corpus BioMRC. Sin embargo, la diferencia de Rouge-2 con la obtenida sobre BioMRC es sólo de 4 puntos, muy por debajo que las diferencias respecto a GigaWord y XSum.

Respecto a los resultados de XSum, con la Rouge-2 más baja de los 4, en un principio podríamos suponer que se debe a que es el conjunto de datos más pequeño de entrenamiento. Sin embargo, ese no parece ser el motivo real porque GigaWord, el dataset con un mayor número de ejemplos, muestra una Rouge-2 muy similar a la obtenida sobre XSum.

De manera general, en Rouge-N podemos apreciar que, a medida que aumenta el número de n-gramas, el resultado de la métrica disminuye a menos de la mitad para todos los datasets. Es previsible esperar que, al aumentar el tamaño de los n-gramas, el número de coincidencias entre el resumen generado y el de referencia sea menor. Como se muestra en la tabla 2, los resultados de Rouge-N sobre el corpus CNN/Daily Mail son considerablemente superiores que los obtenidos sobre los datasets GigaWords y XSum. Esto podría deberse a dos causas distintas: 1) el modelo del lenguaje usado en el enfoque fue preentrenado sobre CNN/Daily Mail, y 2) el tamaño medio de los textos y los resúmenes en CNN/Daily Mail es considerablemente mayor al de los otros datasets, lo que podría beneficiar a la hora de encontrar co-ocurrencias entre los resúmenes generados y los de refe-

rencia.

Con la métrica Rouge-L, que tiene en cuenta la subsecuencia común más larga (LCS) entre el resumen original y el resumen generado por el sistema, el modelo vuelve a obtener los mejores resultados sobre el dataset CNN/Daily Mail. Este dataset contiene el mayor tamaño medio del resumen referencia (56), y por tanto es más probable que las subsecuencias comunes sean más largas. También, como se puede comprobar, el tamaño medio de los textos es más del doble que en el resto de datasets. Como consecuencia, la extensión de los resúmenes propuestos será mayor. Al ser ambos resúmenes más extensos que en el resto de datasets, es de esperar que la subsecuencia común sea más larga. El modelo obtiene una Rouge-L de 37.52 %, 2.2 puntos por debajo del estado de la cuestión reportado en el artículo de Zhang, Wei, y Zhou (2019), con una Rouge-2 de 39.72 %. De nuevo, el modelo obtiene los mejores resultados para el dataset BioMRC, a pesar de pertenecer a un dominio distinto.

Como se explicó anteriormente, la métrica BLEU introduce un parámetro de penalización para aquellas predicciones más cortas que el resumen de referencia. El modelo consigue la mejor puntuación sobre el dataset XSum (Bleu=66.50), seguido de la puntuación obtenida en BioMRC (Bleu=57.73 %). Una posible causa podría estar relacionada con el ratio entre el tamaño medio de los textos y el tamaño medio de los resúmenes. Ambos datasets muestran un ratio considerablemente alto (por encima de 35), mientras que los otros datasets presentan ratios mucho menores. Así la relación entre el tamaño resumen del texto y del resumen es sólo de 3.9 en en GigaWord y 13.9 en CNN/Daily Mail.

En nuestra evaluación, también hemos querido estudiar el tiempo de entrenamiento para cada uno de los datasets, que son mostrados en la tabla 3. Como era de esperar, el tiempo de entrenamiento está directamente relacionado con el tamaño del corpus. Así, el dataset GigaWord tiene un tiempo de entrenamiento mucho mayor que el resto de datasets, porque su tamaño también es mucho mayor. Para una colección de casi 4 millones de instancias de entrenamiento, que su tiempo de entrenamiento sea de poco más de una hora parece bastante razonable y demuestra que los modelos transformers son más eficientes en relación a otras arquitecturas profundas como las RNN. Por la misma razón, XSum, el dataset más pequeño, requiere sólo 8 minutos de entrenamiento. En todos los datasets, el tiempo medio para generar un nuevo resumen es aproximadamente de 15 segundos.

	Tiempo entrenamiento
CNN/Daily	14 minutos
GigaWord	1 hora y 13 minutos
XSum	8 minutos
BioMRC	23 minutos

Tabla 3: Tiempo de entrenamiento para cada conjunto de datos.

La tabla 4 muestra un ejemplo de resumen automático del corpus BioMRC, junto con su correspondiente texto original y resumen de referencia. Este corpus contiene un gran número de conceptos del dominio biomédico, que probablemente no están representados en el corpus CNN/Daily Mail, utilizado para pre-entrenar el modelo BART. A pesar de eso, podemos ver que el resumen generado es bastante similar al resumen de referencia, tanto en contenido semántico como en longitud. Por tanto, podemos concluir que el uso del modelo BART es capaz de generar resúmenes para textos del dominio biomédico.

La tabla 5 muestra un ejemplo tomado del corpus GigaWord. Como se muestra en la tabla 1, este es el corpus con el tamaño más pequeño de textos (un tamaño medio de 31.4 palabras) y de resúmenes, con sólo 8.3 palabras de media. Hemos observado que las predicciones generadas por el modelo automático suelen ser demasiado largas en comparación a los resúmenes de referencia, incluso a veces llegando a superar el tamaño del texto original. En el ejemplo mostrado, consideramos que el resumen generado automáticamente ofrece más información que el resumen original, aunque contiene algunas incoherencias y errores gramaticales (President's invitation of China's President Hu Jintao's).

En el caso del corpus XSum (tabla 6), el resumen automático es completamente correcto desde el punto de vista sintáctico, y también puede ser considerado correcto en cuanto a su significado, aunque difiera del contenido del resumen de referencia. Mientras que el resumen automático pone el foco en que la visita fue realizada por los duques de Cambridge, el resumen de referencia des-

Texto original
The case is reported of a sixty-four-year-old @entity0 with DeBakey type I aortic dissection in whom postoperative extensive intra-aortic balloon pumping was applied. Surgical repair involved replacing the ascending aorta with a Medtronic Hall valved conduit. After surgery severe @entity2 occurred. Despite the use of high-dose inotropic drugs the @entity0 could not be hemodynamically stabilized. An intra-aortic balloon pump was finally applied as a therapeutical last resort. Within three days, under counterpulsation, the @entity0 reached a stable hemodynamic condition. After twenty-one days in the intensive care unit, he could be transferred to a normal ward. The @entity0 was discharged on the fifty-fourth postoperative day. During counterpulsation there were no balloon- or catheter-induced complications. Follow-up at five months showed the @entity0 in good general health: echocardiography did not identify any lesions of the thoracic aorta which could be linked to counter-pulsation. It is concluded that the postoperative use of intra-aortic balloon pump in the event of DeBakey type I dissecting @entity1 of the aorta, and adversely affected @entity0 hemodynamics, is a justifiable therapeutical alternative.
Resumen original
The use of intra-aortic balloon pump after surgical treatment of DeBakey type I dissecting XXXX of the aorta.
Resumen automático
Postoperative use of intra-aortic balloon pumping in the event of DeBakey type I dissecting XXXX of the aorta.

Tabla 4: Ejemplo tomado del corpus BioMRC.

Texto original
US President George W. Gush arrived here saturday evening for a three-day visit to china at the invitation of chinese President Hu Jintao
Resumen original
US President arrives in Beijing
Resumen automático
Bush arrives in China for three-day visit at President's invitation of China's President Hu Jintao's

Tabla 5: Ejemplo tomado del corpus GigaWord.

Texto original
The White Garden, at Kensington Palace, was planted to mark 20 years since Princess Diana died in a car crash.The Duchess of Cambridge joined the princes on the garden tour... Members of the public have been leaving tributes and flowers at the gates of the palace to mark the anniversary of Diana's death. ...It is the fourth London memorial created in tribute to Diana - the others are the Diana Memorial Playground at Kensington Palace, the Diana Memorial Fountain in Hyde Park, and the Diana Memorial Walk at St James's Palace.
Resumen original
Prince William and Prince Harry have visited a London memorial garden for their mother on the eve of the 20th anniversary of her death.
Resumen automático
The Duke and Duchess of Cambridge have visited the White Garden at Kensington Palace to mark the 20th anniversary of Princess Diana's death.

Tabla 6: Ejemplo tomado del corpus XSum.

taca que la visita fue realizada por los príncipes, William y Harry. Ambos resúmenes son correctos.

5 Conclusiones

Este artículo presenta un trabajo preliminar donde exploramos el uso de un modelo trans-

formador para abordar la tarea de generación de resúmenes. Dicho modelo está basado en la arquitectura BART y su implementación es proporcionada por la librería Hugging Face. Mientras que la mayoría de los trabajos anteriores se han centrado en evaluar sus enfoques sobre el dataset CNN/Daily Mail, en nuestro estudio ampliamos la evaluación a cuatro datasets con distintas características. Uno de los datasets está formado por textos biomédicos, mientras que los otros tres son colecciones de artículos periodísticos. Además, los cuatro datasets presentan características distintas en relación al número de instancias y al tamaño medio de sus textos y resúmenes. El modelo ha sido evaluado por las métricas Rouge y Bleu.

Como era de esperar la mejor media de Rouge (31.2 %) sobre el dataset CNN/Daily Mail. La principal razón es porque el modelo de lenguaje usado ha sido pre-entrenado sobre esta misma colección de textos. Esta media es similar a la obtenida por otros trabajos previos basados también en transformadores(Bae et al., 2019; Zhang et al., 2019), pero no llega a superar el estado de la cuestión actual conseguido por el trabajo descrito en (Zhang, Wei, y Zhou, 2019).

Debemos destacar que la segunda mejor media (28.98) en Rouge es obtenida sobre el dataset BioMRC, compuesto por textos biomédicos. Esta media es significativamente superior a las obtenidas en los otros dos datasets, XSum y GigaWord, que sí están formados por artículos periodísticos como el dataset CNN/Daily Mail. Esto podría decirnos que el estilo narrativo y el vocabulario de los textos no parece asegurar unos mejores resultados. También nos permite concluir que el modelo BART puede ofrecer buenos resultados en la generación de resúmenes automáticos en dominios como el biomédico.

Un mayor número de instancias en el corpus de entrenamiento, como es el caso de GigaWord, no parece garantizar la obtención de mejores resultados. De hecho, el modelo obtiene la peor media sobre dicho dataset, a pesar de ser el dataset que tiene mayor número de textos. Por el contrario, la relación entre el tamaño de los textos y el de sus resúmenes podría estar afectando de alguna manera a los resultados, ya que se ha observado que a mayor ratio entre estos tamaños, también es mayor la puntuación obtenida en Rouge-L.

Como trabajo futuro, nos planteamos explorar otros enfoques recientes de transfer learning, tales como T5 (Raffel et al., 2020) y estudiar su adaptación a la tarea de generación de resúmenes. Además, debido a que encontrar las causas de por qué un modelo funciona mejor sobre un determinado dataset no siempre es una tarea trivial, como hemos visto en nuestra experimentación preliminar, también nos gustaría realizar una experimentación con usuarios. Este tipo de evaluación nos podría arrojar más luces sobre las ventajas y las desventajas de cada modelo, y la posible relación entre las características de un conjunto de datos y los resultados obtenidos por un modelo.

La investigación descrita en este artículo forma parte del proyecto NLP4Rare (NLP4Rare-CM-UC3M). El proyecto está financiado por la Comunidad de Madrid y la Universidad Carlos III de Madrid, que pretenden estimular la investigación de naturaleza interdisciplinar de jóvenes doctores. El proyecto NLP4Rare tiene como principal objetivo aumentar el conocimiento sobre las enfermedades raras mediante el uso de técnicas de PLN. Para ello, entre los objetivos específicos persigue la generación automática de resúmenes, tanto para médicos como pacientes, que faciliten la comprensión de toda la información publicada al respecto a una enfermedad o conjunto de enfermedades raras. Debido a que en la actualidad no existe ningún dataset para este dominio, otro de nuestros principales retos será la creación de una dataset para un dominio distinto al periodístico, como es el de las enfermedades raras. Nuestro objetivo es que incluya no sólo textos en inglés, sino también de otras lenguas como el español. Esto nos permitirá evaluar nuestros enfoques de generación de resúmenes para otros idiomas distintos al inglés.

Agradecimientos

Este trabajo ha sido financiado por el Programa de Investigación del Ministerio de Economía y Competitividad del Gobierno de España, (Proyecto DeepEMR TIN2017-87548-C2-1-R) y por el Programa para proyectos interdisciplinares para jóvenes doctores en la Universidad Carlos III de Madrid financiado por la Comunidad de Madrid (Proyecto NLP4Rare-CM-UC3M).

Bibliografía

Bae, S., T. Kim, J. Kim, y S.-g. Lee. 2019. Summary level training of sentence rewriting for abstractive summarization. En *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, páginas 10–20, Hong Kong, China, Noviembre. Association for Computational Linguistics.

Beloki, Z., X. Saralegi, K. Ceberio, y A. Corral. 2020. Grammatical error correction for basque through a seq2seq neural architecture and synthetic examples. *Procesamiento del Lenguaje Natural*, 65:13–20.

Celikyilmaz, A., A. Bosselut, X. He, y Y. Choi. 2018. Deep communicating agents for abstractive summarization. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 1662–1675, New Orleans, Louisiana, Junio. Association for Computational Linguistics.

Chen, Y.-C. y M. Bansal. 2018. Fast abstractive summarization with reinforce-selected sentence rewriting. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 675–686.

Cheng, J. y M. Lapata. 2016. Neural summarization by extracting sentences and words. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume1: Long Papers)*, páginas 484–494.

Colón-Ruiz, C., I. Segura-Bedmar, y P. Martínez. 2019. Análisis de sentimiento en el dominio salud: analizando comentarios sobre fármacos. *Proces. del Leng. Natural*, 63:15–22.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. En *NAACL-HLT (1)*.

Dong, Y., Y. Shen, E. Crawford, H. van Hoof, y J. C. K. Cheung. 2018. BanditSum: Extractive summarization as a contextual bandit. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 3739–3748, Brussels, Belgium, Octubre-Noviembre. Association for Computational Linguistics.

Durrett, G., T. Berg-Kirkpatrick, y D. Klein. 2016. Learning-based single-document summarization with compression and anaphoricity constraints. En *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1998–2008.

Gehrmann, S., Y. Deng, y A. Rush. 2018. Bottom-up abstractive summarization. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 4098–4109, Brussels, Belgium, Octubre-Noviembre. Association for Computational Linguistics.

Graff, D., J. Kong, K. Chen, y K. Maeda. 2003. English gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Grusky, M., M. Naaman, y Y. Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. En *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, páginas 708–719.

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, y P. Blunsom. 2015. Teaching machines to read and comprehend. En *Advances in neural information processing systems*, páginas 1693–1701.

Hochreiter, S. y J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hsu, W.-T., C.-K. Lin, M.-Y. Lee, K. Min, J. Tang, y M. Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 132–141.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, y L. Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, páginas 7871–7880, Online, Julio. Association for Computational Linguistics.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. En *Text Summarization Branches Out*, páginas 74–81, Barcelona, Spain, Julio. Association for Computational Linguistics.

Liu, Y. y M. Lapata. 2019. Text summarization with pretrained encoders. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 3730–3740, Hong Kong, China, Noviembre. Association for Computational Linguistics.

Miranda-Escalada, A. y I. Segura-Bedmar. 2020. One stage versus two stages deep learning approaches for the extraction of drug-drug interactions from texts. *Proces. del Leng. Natural*, 64:69–76.

Nallapati, R., F. Zhai, y B. Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. En S. P. Singh y S. Markovitch, editores, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, páginas 3075–3081. AAAI Press.

Nallapati, R., B. Zhou, C. dos Santos, Ç. GuÌ‡lçehre, y B. Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. En *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, páginas 280–290, Berlin, Germany, Agosto. Association for Computational Linguistics.

Narayan, S., S. B. Cohen, y M. Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. En *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, páginas 1797–1807, Brussels, Belgium, Octubre-Noviembre. Association for Computational Linguistics.

Papineni, K., S. Roukos, T. Ward, y W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. En *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, páginas 311–318.

Pappas, D., P. Stavropoulos, I. Androutsopoulos, y R. McDonald. 2020a. Biomrc: A dataset for biomedical machine reading comprehension. En *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, páginas 140–149.

Pappas, D., P. Stavropoulos, I. Androutsopoulos, y R. McDonald. 2020b. BioMRC: A dataset for biomedical machine reading comprehension. En *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, páginas 140–149, Online, Julio. Association for Computational Linguistics.

Poncelas, A., K. Sarasola, M. Dowling, A. Way, G. Labaka, y I. Alegria. 2019. Adapting NMT to caption translation inwikimedia commons for low-resource languages. *Proces. del Leng. Natural*, 63:33–40.

Radford, A., K. Narasimhan, T. Salimans, y I. Sutskever. 2018. Improving language understanding by generative pre-training.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, y P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Rush, A. M., S. Chopra, y J. Weston. 2015. A neural attention model for abstractive sentence summarization. En *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, páginas 379–389, Lisbon, Portugal, Septiembre. Association for Computational Linguistics.

See, A., P. J. Liu, y C. D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. En *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, páginas 1073–1083, Vancouver, Canada, Julio. Association for Computational Linguistics.

Shi, J., C. Liang, L. Hou, J. Li, Z. Liu, y H. Zhang. 2019. Deepchannel: Salience estimation by contrastive learning for extractive document summarization. En *Proceedings of the AAAI Conference on Artificial Intelligence*, volumen 33, páginas 6999–7006.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017. Attention is all you need. En *Advances in neural information processing systems*, páginas 5998–6008.

Xu, J. y G. Durrett. 2019. Neural extractive text summarization with syntactic compression. En *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, páginas 3283–3294.

Zhang, H., J. Cai, J. Xu, y J. Wang. 2019. Pretraining-based natural language generation for text summarization. En *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, páginas 789–797, Hong Kong, China, Noviembre. Association for Computational Linguistics.

Zhang, X., F. Wei, y M. Zhou. 2019. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 5059–5069, Florence, Italy, Julio. Association for Computational Linguistics.

Zhang, Y., D. Li, Y. Wang, Y. Fang, y W. Xiao. 2019. Abstract text summarization with a convolutional seq2seq model. *Applied Sciences*, 9(8):1665.

Zhong, M., P. Liu, D. Wang, X. Qiu, y X. Huang. 2019. Searching for effective neural extractive summarization: What works and what's next. En *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, páginas 1049–1058, Florence, Italy, Julio. Association for Computational Linguistics.