Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 63-71 recibido 15-12-2022 revisado 05-02-2023 aceptado 14-02-2023

Ajuste y evaluación del modelo DialoGPT sobre distintas colecciones de subtítulos de películas y series de televisión

Fine-tuning and evaluation of DialoGPT on several datasets of English movies and TV series subtitles

Raúl Giménez de Dios, Isabel Segura-Bedmar
Departamento de Informática, Universidad Carlos III de Madrid
raulgimenezdd@gmail.com, isegura@inf.uc3m.es

Resumen: Las nuevas plataformas de streaming han generado una proliferación de películas y series, la mayoría de ellas subtituladas. Esta proliferación proporciona una ingente cantidad de textos conversacionales, menos formales, más interactivos, que reflejan mejor la comunicación entre seres humanos. La mayoría de los modelos transformers desarrollados hasta la fecha no han sido entrenados con textos conversacionales. En este artículo, DialoGPT, un modelo GPT-2 entrenado para la tarea de diálogo sobre una colección de mensajes de Reddit, es re-entrenado y evaluado sobre distintas colecciones de subtítulos en inglés de series populares. Los experimentos muestran que DialoGPT es obtiene buenos resultados, y que el uso de los subtítulos y diálogos de películas y series es un excelente recurso para el desarrollo de chatbots.
Palabras clave: GPT-2, DialoGPT, Chatbot, Transformador.

Abstract: The new streaming platforms have generated a proliferation of movies and series, most of them subtitled. This provides a large number of conversational, less formal, more interactive texts that better reflect communication between human beings. Most of the transformative models developed to date have not been trained with conversational texts. In this article, DialoGPT, a GPT-2 model for the dialog task trained on a collection of Reddit posts, is fine-tuned and evaluated on different collections of English subtitles from popular movies and series. Experiments show that DialoGPT performs well and that English subtitles from movies and series can be an outstanding resource for chatbot development.
Keywords: GPT-2, DialoGPT, Chatbot, Transformer.

1 Introducción

Los asistentes conversacionales o chatbots son programas informáticos capaces de simular una conversación hablada o escrita, tal y como lo haría una persona (Adamopoulou y Moussiades, 2020). Durante los últimos años la creación de chatbots ha recibido gran interés tanto en la investigación científica como por parte de muchas empresas tecnológicas, consiguiendo desarrollar tecnologías cada vez más exitosas a la hora de imitar conversaciones entre seres humanos. Estas tecnologías ofrecen a los usuarios multitud de herramientas que facilitan su día a día (Fu et al., 2022), tales como Alexa y Siri, creados por Amazon y Apple, respectivamente. Estos chatbots son capaces de responder gran cantidad de cuestiones realizadas por los usuarios mediante un diálogo fluido e imitando el sistema de conversación humano con un aceptable nivel de calidad.

Los chatbots pueden ser tanto de dominio abierto o cerrado. Alexa o Siri son ejemplos de chatbots de dominio abierto porque son capaces de dar respuestas a preguntas planteadas por el usuario sobre diferentes temas, sin que la conversación esté focalizada en ningún dominio concreto. Por el contrario, los chatbots de dominio cerrado únicamente son capaces de responder cuestiones relacionadas con una temática o determinado campo de conocimiento, tal y como lo haría un técnico de atención al cliente o un apartado de preguntas frecuentes en una página web (Adamopoulou y Moussiades, 2020). Sin embargo, son poco eficaces cuando la conversación gira entorno a cualquier aspecto ajeno a su dominio.

ISSN 1135-5948 DOI 10.26342/2023-70-5

En los últimos años, la aparición de los modelos transformers ha revolucionado el campo de Procesamiento de Lenguaje Natural (PLN), logrando obtener los mejores resultados en muchas de sus aplicaciones (Wolf et al., 2020; Chernyavskiy, Ilvovsky, y Nakov, 2021). La mayoría de estos modelos (Devlin et al., 2019; Zhuang et al., 2021; Radford et al., 2019) han sido entrenados sobre grandes colecciones de texto escrito de fuentes como wikipedia, noticias, etc. Estos textos escritos posiblemente no sean capaces de representar correctamente las interacciones en un conversación humana. Recientemente, en 2022, esta situación ha comenzado a cambiar con la publicación de varios modelos, como DialoGPT (Zhang et al., 2020) o LamDA (Thoppilan et al., 2022), entrenados con textos conversaciones extraídos de redes sociales como Reddit. Más reciente aún ha sido la presentación de ChatGPT[1], por la empresa OpenAI en noviembre de 2022.

El objetivo de este artículo es utilizar el modelo DialoGPT, basado en GPT-2 y entrenado sobre una colección de 147M conversaciones extraídas de Reddit, para desarrollar un chatbot de dominio abierto. Además, el modelo será ajustado utilizando guiones y subtítulos de diferentes películas y series en inglés. Nuestra hipótesis inicial es que estas conversaciones son un excelente recurso para capturar las características del diálogo entre humanos, aún mejor que las conversaciones extraídas de una red social.

El artículo está organizado como sigue: la sección 2 revisa los trabajos más recientes en el desarrollo de asistentes conversacionales utilizando técnicas de PLN. En la sección 3, describimos las colecciones de subtítulos y el modelo DialoGPT que será re-entrenado sobre dichas colecciones. La sección 4 presenta y discute los resultados para cada una de las colecciones. Finalmente, las principales conclusiones y líneas de trabajo futuro serán descritas al detalle en la sección 5.

2 Estado de la cuestión

A continuación, presentamos los últimos avances que se han realizado en el desarrollo de chatbots basados en técnicas de PLN. Dhyani y Kumar (2021) desarrollaron un chatbot de dominio abierto, basado en una arquitectura bidireccional de red recurrente que utiliza mecanismos de atención para procesar textos más largos de forma correcta. Los autores utilizaron una colección de comentarios de usuarios de la plataforma Reddit[2], que fueron recogidos durante enero de 2015. Esta colección está formado por 3.027.254 instancias para el conjunto de entrenamiento y 5.100 para el conjunto de test. Cada instancia se compone de un comentario y la respuesta asociada al mismo. Las métricas empleadas para la evaluación del modelo generado tras el entrenamiento fueron BLEU (Papineni et al., 2002) y perplejidad (Adiwardana et al., 2020), obteniendo un 30,16 y 56,1, respectivamente.

El objetivo del trabajo presentado en (Konapur et al., 2021) fue el desarrollo de un chatbot capaz de mantener conversaciones con personas que se encuentran en una situación estresante, y responder de la misma forma que lo haría un terapeuta profesional. Para ello, los autores aplicaron distintos modelos que van desde redes de neuronas simples, redes convolucionales, varios tipos especiales de redes recurrentes como son las Long Short Term Memory (LSTM) y las Gated Recurrent Units (GRU). Además, los autores también aplicaron el primero de los modelos transformers (Vaswani et al., 2017a), que fue ajustado para la tarea de diálogo utilizando una colección de conversaciones sobre temas de salud mental entre pacientes y terapeutas, recopiladas de redes sociales como Reddit[3], Counsel chat[4] y Quora[5]. Los experimentos mostraron que el transformador obtenía una métrica BLEU de 33,5. El resto de los modelos fueron únicamente evaluados con la métrica accuracy, para medir con que precisión el modelo había generado la respuesta correcta. GRU fue el mejor modelo con una accuracy de 79 %, seguido por el modelo LSTM con una accuracy de 74 %. El modelo CNN y la red neuronal básica, obtienen peores resultados, 49 % y 56 %, respectivamente.

Adiwardana et al. (2020) presentaron Meena, un chatbot basado en un modelo generativo que fue entrenado utilizando conversaciones obtenidas de las redes sociales. En este trabajo, el objetivo es que el chatbot sea capaz de mantener conversaciones más largas y complejas con varios turnos. Para ello,

[1] https://openai.com/blog/chatgpt

[2] https://www.reddit.com

[3] https://www.reddit.com/r/mentalhealth

[4] https://counselchat.com

[5] https://www.quora.com/topic/Mental-Health

el chatbot necesita recordar el contexto durante la conversación y ser capaz de recordar qué información se ha recopilado en turnos anteriores. Así, el modelo fue entrenado considerando como contexto todas los turnos anteriores (hasta un máximo de siete), y como respuesta, el mensaje mostrado en el siguiente turno. En lugar de utilizar un modelo transformer, Meena está basado en "Evolved Transformer"(So, Le, y Liang, 2019), que utiliza técnicas "Neural architecture search (NAS)", para aprender de forma automática nuevas arquitecturas que sean más eficientes que las propuestas por los seres humanos. En concreto, utiliza un algoritmo evolutivo que copia la arquitectura original del modelo transformer, para encontrar una más óptima. El sistema obtuvo una puntuación de 10,2 en perflexity.

Patel et al. (2019) desarrollan un chatbot con la capacidad de detectar las emociones (felicidad, diversión, vergüenza, enfado, disgusto, tristeza, culpa y miedo) de los usuarios a través de los distintos turnos. En concreto, cada turno de un usuario es clasificado con un porcentaje de positividad o negatividad. Los autores evaluaron tres modelos distintos para la clasificación de estas emociones: una red convolucional (CNN), una red recurrente (RNN) y un tercer modelo que combinaba una red recuerrente y un modelo de atención. Los autores utilizaron la colección ISEAR (Satish y Punkit, 2016), formado por textos en inglés y anotados con las emociones anteriormente citadas. La evaluación mostró que CNN obtenía el mejor resultado con una accuracy de 75 %, mientras que los otros dos modelos únicamente obtenían una accuracy de 70 %.

Aunque todos los chatbots de dominio abierto han sido entrenados con conversaciones extraídas de redes sociales, es complicado dar una comparativa final porque estas colecciones son distintas. Además no siempre se han utilizado las mismas métricas de evaluación. Respecto a los enfoques utilizados, estos van desde arquitecturas de deep learning como las redes recurrentes, convolucionales y el primer modelo transformer propuesto por Vaswani et al. (2017a). Aunque dicho modelo fue ajustado con textos conversacionales de Reddit, el modelo base fue entrenado utilizando oraciones del dataset "WMT 14 English-German Sentence pairs"para la tarea de traducción automática.

3 Métodos

3.1 Datasets

Para entrenar nuestro modelo hemos utilizado distintas colecciones de subtítulos de series en inglés, que han sido obtenidos del portal Kaggle[6]. Estos datasets están formados por los diálogos entre los protagonistas en diferentes películas y series televisivas. Las colecciones seleccionados son los siguientes:

- Diálogos de la película "Pulp Fiction". Cada línea en la colección contiene un diálogo de la película. Se proporciona otra información como el número de palabras en la línea, el nombre del protagonista que dice la línea, el tiempo y lugar donde se dice el diálogo, entre otros. En nuestro caso, únicamente utilizaremos el texto del diálogo.
- Diálogos de aproximadamente 600 capítulos de la serie "The Simpson". Cada línea contiene el texto del diálogo y el protagonista que dice la línea, pero únicamente utilizaremos el texto. En total, contiene 131.551 líneas.
- Diálogos de la serie "Rick and Morty". Cada línea también contiene otra información como el número de temporada y capítulo, el nombre del episodio y el nombre del protagonista que está hablando. Sin embargo, esta información no será utilizada en nuestro sistema.
- Diálogos de la serie "Brooklyn-99". En este caso, además del texto, también se proporciona el nombre del protagonista que está hablando, pero este campo también será ignorado en nuestro sistema.
- Diálogos de la serie "The Office (US)". Otros datos como el número de temporada y episodio, título de episodio y protagonista aunque están disponibles serán omitidos por nuestro sistema.

Nuestro entorno de desarrollo para el entrenamiento y evaluación del modelo DialoGPT ha sido Google Colab, un servicio de Google que proporciona el uso gratuito de unidades GPU computacionales. Al ser gratuito, el servicio tiene ciertas limitaciones. Por ejemplo, no siempre existen unidades disponibles, no es posible ejecutar en segundo

[6] `https://www.kaggle.com/datasets`

plano y el tiempo máximo para ejecutar un proceso es de 12 horas. Por ese motivo, ha sido necesario limitar el tamaño de algunos datasets. La selección de las instancias ha sido aleatoria (aunque se ha asegurado que los diálogos fueran consecutivos) y se han eliminado todas las instancias nulas. Además, para cada dataset hemos generado dos subconjuntos con un ratio aproximado de 90:10 para entrenamiento y evaluación (ver tabla 1).

Dataset	Training	Test	Total
Pulp Fiction	1058	118	1.183
Brooklyn-99	981	110	1.098
Rick Morty	1708	190	1.905
The Office	8813	980	9.800
The Simpson	8813	980	9.800

Tabla 1: Conjuntos para el entrenamiento y evaluación del modelo.

Las colecciones han sido procesadas para que todos tengan el mismo formato: un campo denominado "respuesta", que corresponde al texto hablado por un protagonista, y un texto formado por los siete diálogos anteriores, que denominamos "contexto". Para tener un mayor conocimiento de estas colecciones, se han obtenido una serie de histogramas que muestran la distribución de tokens en cada diálogo. En estos histogramas, además se muestran otros valores estadísticos: la media de tokens por diálogo, la mediana, la desviación estándar y el percentil 90. Estos valores son útiles para definir algunos parámetros del modelo, como por ejemplo la longitud máxima de las secuencias de entrada. A continuación, se muestran los histogramas para cada uno de los colecciones que se corresponden con las figuras 1-5.

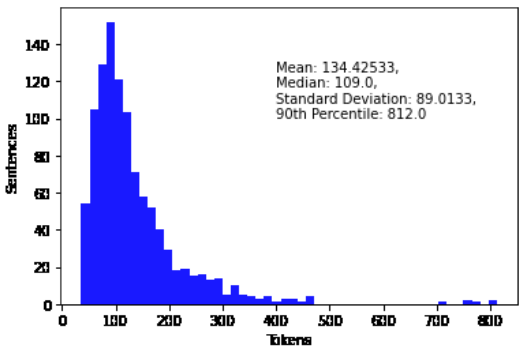

Figura 1: Histograma de la longitud de los textos (número de tokens) en el conjunto "Pulp Fiction".

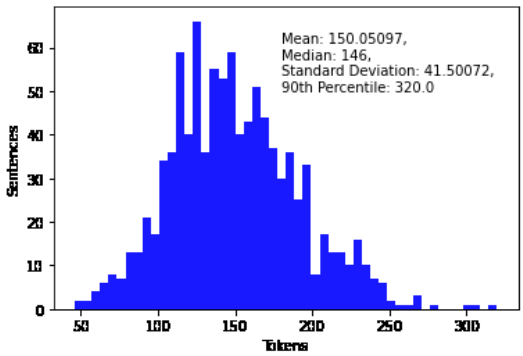

Figura 2: Histograma de la longitud de los textos (número de tokens) en el conjunto "Brooklyn-99".

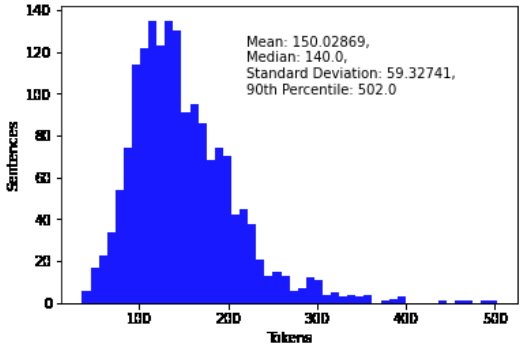

Figura 3: Histograma de la longitud de los textos (número de tokens) en el conjunto "Rick and Morty".

Se puede observar que en el conjunto "Pulp Fiction", la mayoría de las diálogos tienen una longitud máxima de 500 tokens, sin embargo, se observan ciertos casos atípicos que contienen una cantidad de tokens de entre 700 y 850. En el conjunto "Brooklyn-99", los diálogos tienen una longitud media de 150 tokens, siendo 400 tokens el tamaño del diálogo más largo. En el conjunto "Rick and Morty", practicamente el 100 % de los

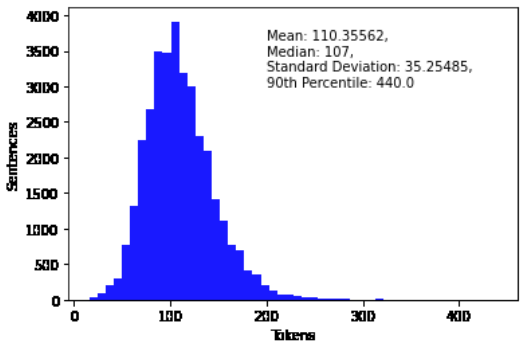

Figura 4: Histograma de la longitud de los textos (número de tokens) en el conjunto "The Simpson".

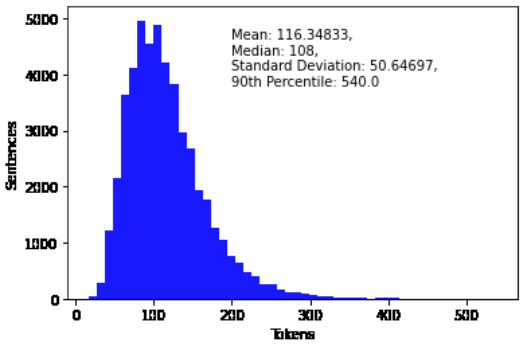

Figura 5: Histograma de la longitud de los textos (número de tokens) en el conjunto "The Office".

diálogos tiene una longitud máxima de 400 tokens, aunque existen algunos diálogos con una longitud mayor. Respecto al conjunto "The Simpson", la mayor parte de los diálogos tienen menos de 200 tokens. Por último, en el conjunto "The Office", la mayoría de los diálogos tienen una longitud máxima de 300 tokens. Para cada conjunto, se ha tomado el tamaño del percentil 90, como el tamaño máximo de la secuencia de entrada, que en ningún caso supera los 512 tokens, tamaño máximo para el modelo GPT-2.

3.2 DialoGPT

La aparición del modelo transformer, descrito en el artículo "Attention all you need"(Vaswani et al., 2017a), ha supuesto una auténtica revolución en el campo del aprendizaje profundo, consiguiendo mejores resultados que los obtenidos por los enfoques desarrollados hasta la fecha en la mayoría de las aplicaciones de PLN (Wolf et al., 2020; Chernyavskiy, Ilvovsky, y Nakov, 2021).

El principal avance del modelo transformer es el uso del mecanismo de atención, que fue diseñado para superar uno de los principales problemas de las redes recurrentes, el procesamiento de las oraciones largas, es decir, con un gran número de palabras. Las redes recurrentes procesan la información de forma secuencial, donde cada palabra está representada con un estado de la red. La información va pasando de un estado a otro, hasta alcanzar el último estado, que será el responsable de generar un vector capaz de codificar toda la información relevante en la oración. Sin embargo, si la oración está compuesta por muchas palabras, la información de sus primeras palabras podría perderse durante el procesamiento de las siguientes palabras. Cuanto mayor sea el número de palabras de la oración, mayor será la probabilidad de que la información relevante descrita en las primeras palabras no esté presente en el último estado (Lavanya y Sasikala, 2021).

El mecanismo de atención tiene en cuenta todos los estados intermedios para generar la salida. Así por ejemplo, en un chatbot, durante la fase de generación de la respuesta para una determinada interacción en el dialogo, el decodificador podrá acceder a todos los estados generados durante la fase de codificación del texto de entrada, y seleccionar aquellos que son más relevantes para generar un elemento específico de la salida. La implementación es bastante sencilla. Para cada elemento de la salida, el decodificador calculará la suma ponderada de los estados del codificador, asignando mayor pesos a los estados que sean más relevantes para el elemento de la salida actual. El mecanismo de atención únicamente está basado en estas sumas, que pueden ser ejecutadas en paralelo, obteniendo un modelo más eficiente. La paralelización es la segunda gran ventaja que ofrecen los modelos transformers respecto a las redes recurrentes (Vaswani et al., 2017a).

Tras la aparición de este primer modelo transformer (Vaswani et al., 2017a), otros modelos transformers han sido propuestos durante los últimos años, tales como BERT (Devlin et al., 2019) o GPT-2 (Radford et al., 2019). Aunque cada uno de ellos presentan características propias, todos se basan en el uso del mecanismo de atención. Además, estos nuevos modelos han sido entrenados como modelos de lenguaje a partir de grandes colecciones de textos. Las principales diferencias entre estos modelos principalmente recaen sobre el tipo de estrategia que utilizan para entrenar sus modelos de lenguaje. Así, mientras BERT utiliza las estrategias de enmascaramiento (donde el objetivo es predecir un token enmascarado en una oración) y predecir si dos oraciones son consecutivas, GPT-2 está basado en un modelo autoregresivo donde el objetivo es dada una secuencia de palabras, predecir la siguiente palabra. Además, BERT se puede considerar un modelo bidireccional, porque cuando predice un token, puede considerar todos los tokens en el contexto del token enmascarado, mientras que GPT-2 únicamente podrá utilizar los tokens anteriores, es decir, los tokens a la izquierda del token a predecir.

Otra importante diferencia entre BERT y GPT-2 se basa en su arquitectura. Mientras BERT únicamente está compuesto por varias capas de codificadores, encargadas de aprender una representación vectorial para cada palabra de una oración, GPT-2 consiste en un bloque de decodificadores, que aplicando un modelo autoregresivo se encargan de predecir el siguiente token que podría continuar a una secuencia de palabras de entrada. Así por ejemplo, dada la entrada "El médico recetó un ", GPT-2 podría generar la palabra "antibiótico", formando así la nueva oración "El médico recetó un antibiótico". Esta oración pasaría a ser la nueva entrada del modelo, y aplicando el mismo procedimiento iterativamente se consigue generar texto nuevo. GPT-2 y BERT, ambos modelos transformers, están basados en el mecanismo de atención, aunque en GPT-2, el mecanismo de atención (denominado "masked self-atention") es ligeramente distinto, ya que únicamente puede considerar los tokens anteriores al token que se están procesando en cada momento (Radford et al., 2019).

El conocimiento codificado en estos modelos de lenguaje puede ser transferido y utilizado para el desarrollo de aplicaciones concretas de PLN. A este proceso, donde un modelo de lenguaje es re-entrenado para una tarea y dataset específico, se denomina en inglés "fine-tuning"(Vrbančič y Podgorelec, 2020). Aunque tanto BERT como GPT-2 pueden ser adaptados para cualquier tarea de PLN, GPT-2 es una elección más apropiada en aplicaciones que implican la generación de nuevo texto como es el caso de los chatbots (Budzianowski y Vulić, 2019).

En 2021, Microsoft presentó un nuevo modelo, DialoGPT (Zhang et al., 2020), que fue específicamente creado para implementar un sistema de diálogo. Su principal ventaja es que es capaz de solventar problemas que presentaban enfoques previos en tareas de diálogo como la falta de consistencia y contextualización (Huang, Zhu, y Gao, 2020). El modelo original GPT-2 fue entrenado con 40GB de páginas webs, consiguiendo un vocabulario de 50.000 tokens. Al igual que BERT, el tamaño máximo de oración fue 512 tokens. A su vez, el modelo DialoGPT fue re-entrenado para la tarea de diálogo utilizando 147M conversaciones extraídas de Reddit. Cada instancia está formada por una secuencia de mensajes consecutivos, y el siguiente mensaje que se considera la respuesta que debería generar el modelo para la secuencia de entrada. Se realizaron algunas tareas de preprocesamiento para eliminar instancias que pueden generar ruido. Por ejemplo, se eliminaron todas las instancias cuya respuesta contenía urls, carácteres especiales como "[.º "]", o lenguaje tóxico (que fue detectado usando palabras claves). También se eliminaron las instancias donde la respuesta no contenía ninguna de las 50 palabras más comunes en inglés o tenía alguna palabra que se repetía más de tres veces. Además, no se consideraron las instancias cuyo texto de entrada y respuesta superaban las 200 palabras.

Como característica especial respecto a otras tareas, en la tarea de diálogo es necesario incorpora un nuevo token especial, "[end_of_turn]", que permita identificar el final de cada interacción en el diálogo. En cada instancia, las interacciones deberán estar separadas por este token. A su vez, el modelo también debe ser capaz de generar dicho token para marcar el final del mensaje de salida. La plataforma HuggingFace proporciona el modelo DialoGPT en tres versiones diferentes: small (117M), medium (345M), y Large (762M). Dichos modelos fueron entrenados con 5, 5, y 3 epochs respectivamente. Además, las interacciones con un número de tokens similar fueron agrupadas en el mismo batch para mejorar el training.

En este trabajo, la versión small del modelo DialoGPT ha vuelto a ser entrenado para colección descrita en el apartado anterior. En cada colección, cada instancia está formada por siete interacciones consecutivas en el diálogo, y una octava interacción, que se considera como respuesta, tal y como se describió en el apartado anterior. Con el objetivo de facilitar la replicabilidad de nuestra experimentación, nuestra implementación está disponible en un repositorio de GitHub[7].

4 Evaluación

Para evaluar el chatbot, hemos utilizado una de las métricas estándar para la evaluación de modelos de lenguaje, la perplejidad (Meister y Cotterell, 2021). Es una medida estadística de la confianza con la que un modelo de lenguaje predice un nuevo de texto. Podríamos definirlo como el grado de incertidumbre o duda que un modelo tienes respecto a si un

[7] `https://github.com/isegura/DialoGPTsepln`

texto es correcto o no. La perplejidad de un texto se calcula con la siguiente ecuación:

$$Perplejidad(W) = P(w_1, ..., w_n)^{-\frac{1}{n}} \quad (1)$$

donde W es una oración, N es el número de palabras, e w_i es la i-ésima palabra de la oración.

Si la perplejidad de un texto es bajo, significa que el modelo es capaz de generar dicho texto. El objetivo deseable es obtener modelos cuya complejidad media sobre una colección de textos sea baja, mientras que una perplejidad alta indicaría que el modelo no es capaz de predecir esos textos. El cálculo de la perplejidad es sencillo y rápido, lo que es especialmente ventajoso cuando estamos evaluando un modelo sobre una gran colección de textos. En nuestro caso, hemos utilizado la implementación proporcionada por HuggingFace para el cálculo de esta métrica, donde únicamente es necesario indica el modelo a utilizar y la colección de textos.

La tabla 2 muestra el grado de perplejidad del modelo DialoGPT (versión small) ajustado para cada una de las colecciones descritas en el apartado 3.1. Además de la perplejidad, también se incluye el error (Loss). Estos valores han sido obtenidos sobre los conjuntos test que fueron creados a partir de las colecciones (ver tabla 1).

Dataset	**Error**	**Perplejidad**
Pulp Fiction	1.09	2.97
Brooklyn-99	1.55	4.28
Rick Morty	1.35	3.84
The Office	1.13	3.11
The Simpson	1.6	4.95

Tabla 2: Resultados de DialoGPT.

A la vista de los resultados obtenidos es posible afirmar que DialoGPT obtiene los mejores resultados cuando es ajustado y evaluado utilizando los diálogos de la película "Pulp Fiction", ya que tanto su error como su perplejidad son los valores más bajos de la tabla. En contraposición, DialoGPT muestra los peores resultados para la colección "The Simpson", seguido muy cerca por "Brooklyn-99". Llama nuestra atención que el tamaño de las colecciones no parece tener un efecto directo sobre el grado perplejidad. Así por ejemplo, aunque "The Simpson.^{es} una de las dos colecciones mayores, su grado perplejidad ha sido el más alto, mientras que la perplejidad obtenida en "Pulp Fiction.^{es} la más baja, aunque dicha colección es una de las más pequeñas. Una posible razón es que los diálogos de esta película contengan menos ruido que los diálogos de "The Simpson".

Aunque no es posible compararnos con los trabajos del estado de la cuestión (ver apartado 2), porque los modelos han sido evaluados sobre colecciones distintas y se han empleado métricas diferentes, sí podemos afirmar que nuestro enfoque parece obtener perplejidades significativamente más bajas que la obtenida por la red bidireccional recurrente utilizada en el trabajo (Dhyani y Kumar, 2021), con perplejidad de 56.1 en una colección de conversaciones de Reddit. Del mismo modo, DialoGPT parece obtener mejores resultados que los obtenidos por el modelo Meena (Adiwardana et al., 2020), cuya perplejidad era 10.2 sobre una colección de textos conversacionales tomados de redes sociales.

Para considerar otra referencia para los valores de perplejidad, podemos considerar el primer modelo transformador (Vaswani et al., 2017b) cuya perplejidad fue 4.33, un valor similar a los obtenidos en en nuestra experimentación. Un reciente estudio (Ngo et al., 2021), mostraba que el modelo GPT-2 tenía una perplejidad de 74.7 evaluado sobre el corpus "One Billion Word Benchmark"(Chelba et al., 2014). Por tanto, podemos considerar que nuestros resultados de perplejidad son significativamente bajos, y que el modelo DialoGPT predice correctamente el texto.

Aunque desgraciadamente no ha sido posible abordar una evaluación basada en usuarios, la ejecución del chatbot nos ha permitido observar que este es capaz de generar texto con total sentido en base a la entrada del usuario, incluso manteniendo y teniendo en cuenta la información de turnos anteriores. El código para el entrenamiento y ejecución del chatbot está disponible en un repositorio de GitHub[8].

5 Conclusiones

La principal contribución de este trabajo ha sido utilizar el modelo DialoGPT para el desarrollo de un chatbot en el dominio abierto. Aunque dicho modelo ya fue entrenado con textos conversacionales obtenidos de re-

[8] `https://github.com/isegura/DialoGPTsepln`

des sociales, en nuestro trabajo, hemos ajustado y evaluado el modelo sobre colecciones de diálogos y subtítulos de distintas películas y series de televisión en inglés.

Nuestros resultados no son directamente comparables con trabajos anteriores porque se utilizan métricas y colecciones distintos, pero basándonos en los valores de perplejidad, podemos afirmar que nuestro modelo es capaz de generar correctamente el texto. La interacción con el chatbot también nos ha permitido comprobar que es capaz de mantener una conversación fluida y coherente, incluso siendo capaz de mantener y tener en cuenta el contexto descrito en los turnos anteriores. Por tanto, DialoGPT ajustado con diálogos y subtítulos de películas y series de televisión son un enfoque adecuado para el desarrollo de chatbots en el dominio abierto.

Entre las líneas de trabajo futuro, planeamos extender nuestro estudio a otros modelos transformers que fueron también preentrenados con textos conversacionales como Lamda y ChatGPT. Aunque hemos visto que los resultados no parecen estar directamente ligados con el tamaño del conjunto entrenamiento, se tratará de trabajar con mayores capacidades de cómputo para poder procesar el mayor conjunto de datos posible. En el actual trabajo se ha demostrado que los subtítulo y diálogos de películas y series de televisión son un buen recurso para la generación de chatbots. Otra de nuestras líneas futuras será aprovechar la existencia de estos subtítulos en diferentes idiomas para entrenar modelos multilingües y aplicarlos al desarrollo de chatbots. Extender nuestra evaluación a otras métricas y abordar una evaluación basada en usuarios son algunos de los retos que nos gustaría abordar en el futuro. Otra línea de trabajo será investigar para mitigar los posibles sesgos y estereotipos de género racismo, implícitos en los datos, y que podrían generar respuestas en los chatbots que sean ofensivas o poco éticas.

Agradecimientos

Esta publicación es parte del proyecto de I+D+i ACCESS2MEET (PID2020-116527RB-I0) financiado por AEI/10.13039/501100011033/.

Bibliografía

Adamopoulou, E. y L. Moussiades. 2020. An overview of chatbot technology. En *IFIP International Conference on Artificial Intelligence Applications and Innovations*, páginas 373–383. Springer.

Adiwardana, D., M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, y others. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Budzianowski, P. y I. Vulić. 2019. Hello, it's GPT-2 - how can I help you? towards the use of pretrained language models for task-oriented dialogue systems. En *Proceedings of the 3rd Workshop on Neural Generation and Translation*, páginas 15–22, Hong Kong, Noviembre. Association for Computational Linguistics.

Chelba, C., T. Mikolov, M. Schuster, Q. Ge, T. Brants, P. Koehn, y T. Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. En *INTERSPEECH 2014*, páginas 2635–2639.

Chernyavskiy, A., D. Ilvovsky, y P. Nakov. 2021. Transformers:"the end of history" for natural language processing? En *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, páginas 677–693. Springer.

Devlin, J., M.-W. Chang, K. Lee, y K. Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, páginas 4171–4186, Minneapolis, Minnesota, Junio. Association for Computational Linguistics.

Dhyani, M. y R. Kumar. 2021. An intelligent chatbot using deep learning with bidirectional rnn and attention model. *Materials Today: Proceedings*, 34:817–824. 3rd International Conference on Science and Engineering in Materials.

Fu, T., S. Gao, X. Zhao, J. rong Wen, y R. Yan. 2022. Learning towards conversational ai: A survey. *AI Open*, 3:14–28.

Huang, M., X. Zhu, y J. Gao. 2020. Challenges in building intelligent open-domain

dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32.

Konapur, S. P., T. Krishna, V. G, U. R, y S. H. 2021. Design of a chatbot for people under distress using transformer model. En *2021 2nd Global Conference for Advancement in Technology (GCAT)*, páginas 1–4.

Lavanya, P. y E. Sasikala. 2021. Deep learning techniques on text classification using natural language processing (nlp) in social healthcare network: A comprehensive survey. En *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, páginas 603–609. IEEE.

Meister, C. y R. Cotterell. 2021. Language model evaluation beyond perplexity. En *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, páginas 5328–5339, Online, Agosto. Association for Computational Linguistics.

Ngo, H., J. G. Araújo, J. Hui, y N. Frosst. 2021. No news is good news: A critique of the one billion word benchmark. En *35th Conference on Neural Information Processing Systems,*.

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.

Patel, F., R. Thakore, I. Nandwani, y S. K. Bharti. 2019. Combating depression in students using an intelligent chatbot: A cognitive behavioral therapy. En *2019 IEEE 16th India Council International Conference (INDICON)*, páginas 1–4.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, y others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Satish, T. y A. Punkit. 2016. Emotion detection in text.

So, D., Q. Le, y C. Liang. 2019. The evolved transformer. En *International Conference on Machine Learning*, páginas 5877–5886. PMLR.

Thoppilan, R., D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, y others. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, y I. Polosukhin. 2017a. Attention is all you need. *Advances in neural information processing systems*, 30.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, y I. Polosukhin. 2017b. Attention is all you need. En I. Guyon U. V. Luxburg S. Bengio H. Wallach R. Fergus S. Vishwanathan, y R. Garnett, editores, *Advances in Neural Information Processing Systems*, volumen 30. Curran Associates, Inc.

Vrbančič, G. y V. Podgorelec. 2020. Transfer learning with adaptive fine-tuning. *IEEE Access*, 8:196197–196211.

Wolf, T., L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, y others. 2020. Transformers: State-of-the-art natural language processing. En *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, páginas 38–45.

Zhang, Y., S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, y B. Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, páginas 270–278, Online, Julio. Association for Computational Linguistics.

Zhuang, L., L. Wayne, S. Ya, y Z. Jun. 2021. A robustly optimized BERT pre-training approach with post-training. En *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, páginas 1218–1227, Huhhot, China, Agosto. Chinese Information Processing Society of China.