Diploma Programme
Programme du diplôme
Programa del Diploma

Informática
Estudio de caso: El chatbot perfecto

Para usar en mayo y noviembre de 2025

Instrucciones para los alumnos

- Para la prueba 3 de Nivel Superior se requiere el cuadernillo del estudio de caso.

5525 – 6403

Escenario

Una empresa de seguros, *RAKT*, ha implementado recientemente un chatbot (robot de diálogo) de modelo lingüístico para atender las consultas de sus clientes. Sin embargo, el chatbot no ha funcionado bien y muchos clientes se han mostrado insatisfechos con sus respuestas.

La empresa te ha contratado como estudiante en prácticas para investigar los problemas y recomendar mejoras para el chatbot. Tu supervisor ya ha analizado las quejas de los clientes y ha identificado la causa de los problemas. Quiere que explores seis áreas clave que podrían mejorar el rendimiento del chatbot.

Problemas por resolver

1. **Latencia**: El tiempo de respuesta del chatbot es lento y perjudica la experiencia del cliente.
2. **Matices lingüísticos**: El modelo lingüístico del chatbot tiene dificultades para responder adecuadamente a enunciados ambiguos.
3. **Arquitectura**: La arquitectura del chatbot es demasiado simplista e incapaz de manejar un lenguaje complejo.
4. ***Conjunto de datos***: El conjunto de datos de entrenamiento del chatbot no es lo suficientemente diverso, lo que provoca una escasa precisión a la hora de comprender y responder a las consultas de los clientes.
5. **Potencia de procesamiento**: La capacidad de cálculo del sistema es un factor limitante.
6. **Problemas éticos**: El chatbot no siempre da consejos adecuados y es propenso a revelar información personal de su conjunto de datos de entrenamiento.

Latencia

La clientela de *RAKT* espera que el chatbot proporcione respuestas acertadas y precisas a sus consultas. Se desarrolló un complejo modelo de *procesamiento del lenguaje natural* para manejar los matices de la interacción entre humanos. Sin embargo, esto ha aumentado el tiempo que tarda el chatbot en responder, sobre todo cuando el volumen de consultas es elevado.

La inteligencia artificial (IA) conversacional necesita una amplia red de modelos de aprendizaje automático para determinar qué decir a continuación, y cada modelo resuelve una pequeña pieza del rompecabezas. Un modelo podría tener en cuenta la ubicación del usuario, mientras que otro podría analizar el historial de interacciones del usuario, incluidos los comentarios de los usuarios sobre respuestas similares. Cada modelo debe aportar mejoras, pero a costa de aumentar la latencia del sistema.

Esa latencia es producto de un algoritmo de decisión conocido como “ruta crítica”, que es la secuencia más corta y eficiente de modelos de aprendizaje automático enlazados necesaria para ir de la entrada del mensaje del usuario a la salida de la respuesta del chatbot. Se sabe que los cambios en un modelo pueden afectar a redes de aprendizaje automático más grandes con dependencias de aprendizaje automático.

Una forma de superar las dependencias es transformar el texto no estructurado en información procesable por máquinas. La *comprensión del lenguaje natural* (*NLU*, por sus siglas en inglés) es una canalización creada a través de muchos modelos de aprendizaje automático que llevan a cabo diferentes funciones para mejorar la comprensión por parte del chatbot de las entradas del usuario. Esto ayuda a identificar y filtrar los modelos innecesarios, lo que reduce la latencia.

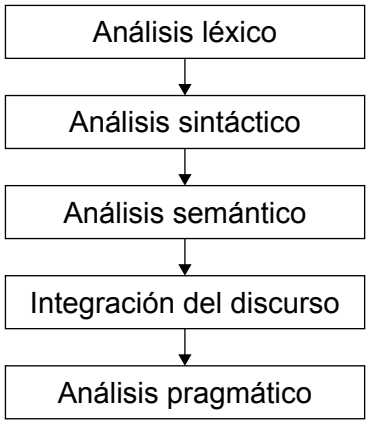
Entrenar a un chatbot es otro paso para mejorar su capacidad de comprensión de texto. Para reducir el tiempo de respuesta, el conjunto de datos de entrenamiento debe ser amplio, preciso, clasificado, legible, pertinente y específico del dominio.

Matices lingüísticos

Los modelos lingüísticos de chatbot pueden beneficiarse de la incorporación de características más parecidas a las humanas, como la capacidad de detectar y responder a la emoción, el tono y el contexto. La mejora del modelo lingüístico permite al chatbot generar respuestas más personalizadas y adaptadas a las necesidades individuales del cliente, lo que redunda en una mejor experiencia general.

Para que las respuestas del chatbot sean personalizadas, se siguen cinco etapas de procesamiento del lenguaje natural (véase la **figura 1**).

Figura 1: Las cinco etapas de procesamiento del lenguaje natural



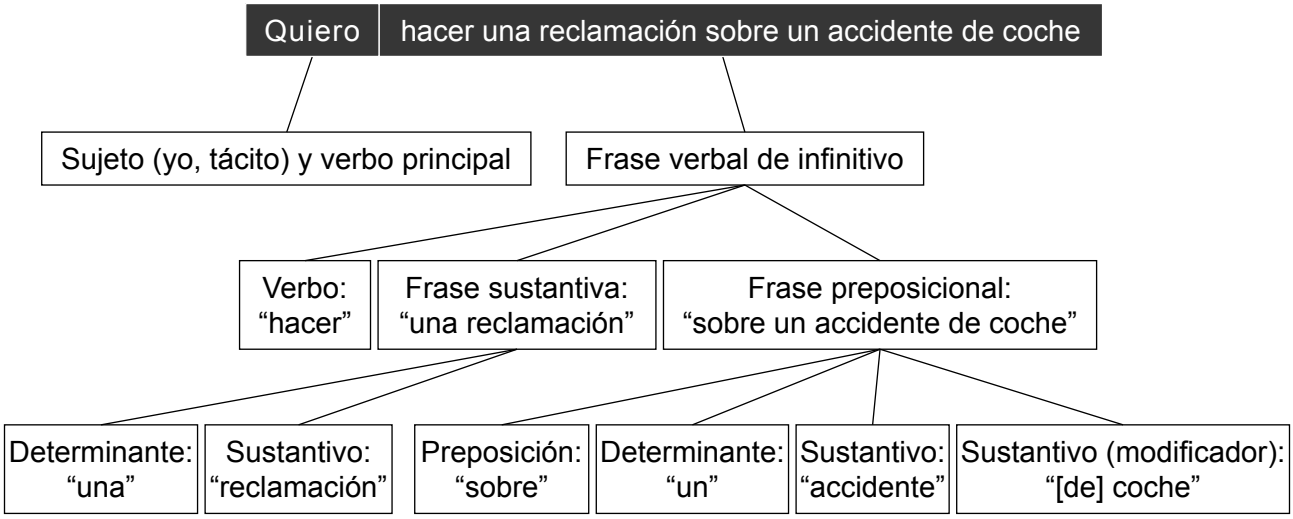
El análisis léxico y el análisis sintáctico se centran en los aspectos estructurales del lenguaje, como la descomposición del texto en palabras individuales, la identificación de las partes de la oración y el análisis de las relaciones entre palabras y frases. El análisis semántico, la integración del discurso y el análisis pragmático se centran en el significado del lenguaje, más allá de la estructura superficial.

Supongamos que un cliente introduce la siguiente frase en el chatbot: “Quiero hacer una reclamación sobre un accidente de coche”. El procesamiento del lenguaje natural en cinco etapas interpreta esta frase de la siguiente manera:

1. **Análisis léxico**: El texto de entrada se descompone en palabras y frases; por ejemplo, [“quiero”, “hacer”, “un”, “reclamación”, “sobre”, “un”, “accidente”, “de”, “coche”].
2. **Análisis sintáctico (*parsing*)**: Interpretar la gramática y la estructura de la frase; por ejemplo, que “yo” es el sujeto tácito de la frase, “quiero” es el verbo y “reclamación” es el objeto (véase la **figura 2**).
3. **Análisis semántico**: Analizar el significado de la frase y de cada una de sus palabras; por ejemplo, la frase trata del deseo del cliente de presentar una reclamación relacionada con un accidente de coche.
4. **Integración del discurso**: Integrar el significado de la frase en el contexto más amplio de la conversación; por ejemplo, entender que el usuario es un cliente que desea informarse sobre una reclamación.
5. **Análisis pragmático**: Analizar el contexto social, jurídico y cultural de la frase; por ejemplo, entender que el cliente es un conductor de coche que ha tenido un accidente y puede o no estar herido o haber dañado el vehículo.

Véase al dorso

Figura 2: Análisis sintáctico de la gramática y la estructura de la frase por parte del chatbot



Arquitectura

El motor de procesamiento del lenguaje natural es el componente central de la arquitectura de un chatbot. Utiliza algoritmos de aprendizaje automático para determinar la intención del usuario y relacionarla con la acción realizada por el software.

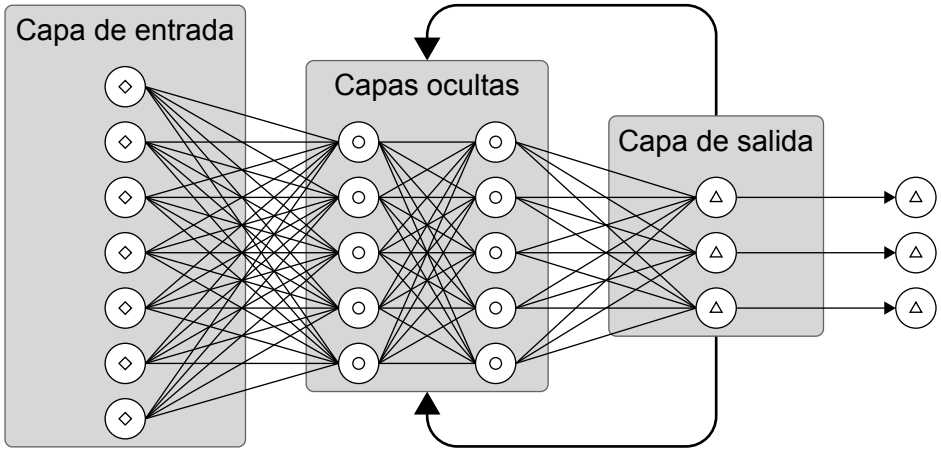
Existen dos tipos de modelos de *aprendizaje profundo* utilizados habitualmente en el procesamiento del lenguaje natural: las *redes neuronales recurrentes (RNR)* y las *redes neuronales transformadoras (RNT)*.

Redes neuronales recurrentes (RNR)

El chatbot de la aseguradora utiliza una RNR, una red neuronal artificial diseñada para procesar datos secuenciales.

Una RNR tiene tres capas. La capa de entrada procesa la secuencia de datos, y cada elemento de la secuencia se introduce como entrada en la red. Las capas ocultas se encargan de mantener la memoria de la red. La capa de salida produce el resultado final de la red, que puede ser una predicción o un resultado de clasificación (véase la **figura 3**).

Figura 3: Capa de entrada, capas ocultas y capa de salida de una red neuronal recurrente (RNR)



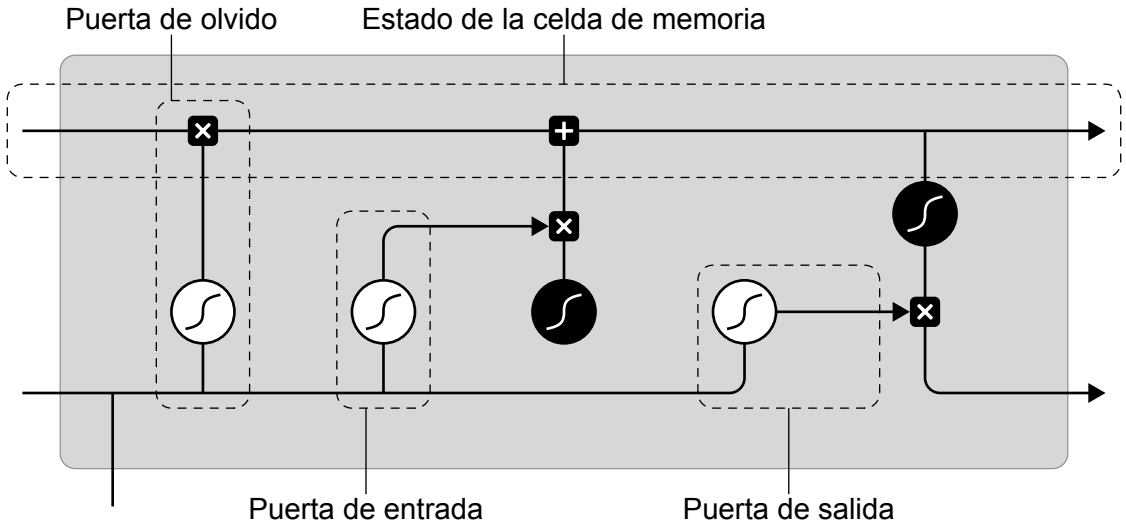
Una RNR puede procesar datos secuenciales de longitud variable manteniendo una memoria de entradas anteriores en forma de estados ocultos.

Todas las RNR deben entrenarse utilizando un conjunto de datos preprocesados adecuado dividido en conjuntos de entrenamiento, validación y prueba. Se fijan los hiperparámetros, como la tasa de aprendizaje y el número de capas ocultas. El *ajuste de hiperparámetros* consiste en encontrar los valores óptimos para estos hiperparámetros.

Una RNR ajusta los *pesos* de las capas de entrada, de salida y ocultas durante el entrenamiento mediante *retropropagación en el tiempo* (*BPTT*, por sus siglas en inglés). Esta técnica calcula los gradientes de la *función de pérdida* para cada peso en cada paso temporal, y actualiza los pesos en consecuencia. Sin embargo, entrenar las RNR utilizando BPTT puede causar el problema del *gradiente evanescente*, que dificulta el aprendizaje de dependencias a largo plazo en los datos.

La *memoria a corto plazo de larga duración* (*LSTM*, por sus siglas en inglés) es un tipo de RNR diseñado para superar el problema de los gradientes evanescentes. Lo consigue con un mecanismo de tres puertas que controla el flujo de información. El *estado de las celdas de memoria*, la puerta de entrada, la puerta de olvido y la puerta de salida permiten a las LSTM retener u olvidar información de forma selectiva a lo largo del tiempo (véase la **figura 4**).

Figura 4: La memoria a corto plazo de larga duración (LSTM)



Véase al dorso

Redes neuronales transformadoras (RNT)

Una red neuronal transformadora o de transformador (RNT) es una potente alternativa a la LSTM, cuya innovación clave es la aplicación de un *mecanismo de autoatención*. El mecanismo de autoatención capta las relaciones entre las distintas palabras de la secuencia de entrada calculando los pesos de atención para cada palabra en función de su similitud con otras palabras de la secuencia.

Un ejemplo de *gran modelo lingüístico* (*LLM*, por sus siglas en inglés) que utiliza el mecanismo de autoatención es el Generative Pretrained Transformer 3 (GPT-3) creado por Open AI.

Conjuntos de datos

Se cree que el chatbot no ha utilizado un conjunto de datos imparcial y de alta calidad. Los conjuntos de datos deben ser relevantes para el dominio en el que opera el chatbot. Puede tratarse de datos de reclamaciones de seguros, datos de atención al cliente, datos de pólizas de seguros y datos médicos. Los conjuntos de datos pueden crearse a partir de datos reales o *sintéticos*.

Es importante garantizar que estos conjuntos de datos sean imparciales y diversos para evitar perpetuar cualquier *sesgo* existente en el sector respectivo.

Varios tipos de sesgos en los conjuntos de datos pueden afectar a la precisión y equidad de los modelos de procesamiento del lenguaje natural, como los sesgos de *confirmación*, *históricos*, de *etiquetado*, *lingüísticos*, de *muestreo* y de *selección*.

- **Sesgo de confirmación**: Esta forma de sesgo se produce cuando el conjunto de datos está sesgado hacia un punto de vista concreto, como los datos de entrenamiento que solo incluyen consultas de clientes relacionadas con determinados tipos de pólizas.
- **Sesgo histórico**: Esta forma de sesgo se produce cuando los datos de entrenamiento no reflejan los cambios a lo largo del tiempo. Por ejemplo, si el modelo de procesamiento del lenguaje natural se entrena con datos de hace varios años, puede que no sea capaz de predecir con exactitud las consultas recientes de los clientes.
- **Sesgo de etiquetado**: Esta forma de sesgo se produce cuando las etiquetas aplicadas a los datos son subjetivas, inexactas o incompletas. Por ejemplo, si las etiquetas asignadas a las consultas de los clientes son demasiado genéricas, es posible que el modelo no pueda predecir con exactitud la intención del cliente.
- **Sesgo lingüístico**: Esta forma de sesgo se produce cuando el conjunto de datos está sesgado hacia determinadas características lingüísticas, como el dialecto o el vocabulario. Por ejemplo, si un conjunto de datos se basa en un lenguaje escrito formal, es posible que el modelo no pueda interpretar con precisión el lenguaje informal.
- **Sesgo de muestreo**: Esta forma de sesgo se produce cuando el conjunto de datos de entrenamiento no es representativo de toda la población, como ocurre con los datos de entrenamiento que solo incluyen consultas de clientes de un grupo demográfico.
- **Sesgo de selección**: Esta forma de sesgo se produce cuando los datos de entrenamiento no se seleccionan al azar, sino que se eligen en función de algún criterio. Es el caso de un modelo lingüístico entrenado a partir de datos que sugieren que determinados grupos demográficos pueden ser más propensos a presentar reclamaciones de seguros sesgadas hacia las personas que pertenecen a esa categoría.

Potencia de procesamiento

Los pasos para preparar un modelo de aprendizaje automático son:

1. *Preprocesamiento* de los datos de entrada
2. Entrenamiento del modelo
3. Despliegue del modelo

Preprocesamiento de los datos de entrada

El preprocesamiento es el primer paso que se realiza sobre los datos brutos para hacerlos comprensibles a los algoritmos de aprendizaje automático. El preprocesamiento implica la depuración, selección, transformación y reducción de datos para mejorar su calidad y precisión.

En la actualidad, el modelo lingüístico del chatbot utiliza el algoritmo *de bolsa de palabras* para comprender la entrada del cliente y extraer la información relevante. Para ello, aplica una representación vectorial del texto basada en la frecuencia de las palabras. El algoritmo puede exigir una gran potencia de procesamiento cuando se trata de grandes conjuntos de datos.

Entrenamiento del modelo

El entrenamiento del modelo de aprendizaje automático es la tarea computacionalmente más intensiva, ya que se utilizan cantidades masivas de datos de texto. El proceso de entrenamiento puede llevar semanas o incluso meses y requiere hardware potente, como clústeres de *unidades de procesamiento gráfico* (*GPU*, por sus siglas en inglés) o *unidades de procesamiento tensorial* (*TPU*, por sus siglas en inglés), para acelerar los cálculos.

Despliegue del modelo

Una vez entrenado, un LLM puede desplegarse en diversas plataformas de hardware. Sin embargo, para lograr el mejor rendimiento, se recomienda el uso de hardware especializado, como las TPU. Además de los requisitos de hardware, la ejecución de los LLM también requiere una cantidad significativa de almacenamiento y memoria para mantener el modelo y sus datos asociados.

Problemas éticos

El equipo de *RAKT* ha identificado varios problemas éticos que tendrán que investigar. Entre estos figuran:

- Privacidad y seguridad de los datos: proteger los datos de los usuarios contra usos indebidos.
- Sesgos e imparcialidad: evitar respuestas discriminatorias.
- Rendición de cuentas y responsabilidad: asignar responsabilidades por el asesoramiento prestado.
- Transparencia: explicar claramente la toma de decisiones.
- Desinformación y manipulación: prevención de la difusión de información falsa.

Desafíos

Los siguientes problemas deberían estudiarse con más detalle:

- Evaluación de métodos para reducir la latencia.
- Comprensión de las cinco etapas del procesamiento del lenguaje natural.
- Determinación de si una RNT podría mejorar el rendimiento de una RNN para el procesamiento del lenguaje natural.
- Comprensión de lo que se necesita para crear un gran conjunto de datos imparciales y de alta calidad.
- Análisis de los requisitos de potencia de procesamiento de un modelo de procesamiento del lenguaje natural.
- Exploración de la ética de usar un chatbot para asesorar.

No es necesario que los candidatos comprendan las características lingüísticas de las oraciones, como en el modelo sujeto, verbo, objeto (SVO).

Véase al dorso

Terminología adicional

Ajuste de hiperparámetros (*hyperparameter tuning*)
Aprendizaje profundo (*deep learning*)
Bolsa de palabras (*bag-of-words*)
Comprensión del lenguaje natural (NLU, *natural language understanding*)
Conjunto de datos (*dataset*)
Datos sintéticos (*synthetic data*)
Estado de la celda de memoria (*memory cell state*)
Función de pérdida (*loss function*)
Gradiente evanescente (*vanishing gradient*)
Gran modelo lingüístico (LLM, *large language model*)
Latencia (*latency*)
Mecanismo de autoatención (*self-attention mechanism*)
Memoria a corto plazo de larga duración (LSTM, *long short-term memory*)
Pesos (*weights*)
Preprocesamiento (*pre-processing*)
Procesamiento del lenguaje natural (*natural language processing*)
- Análisis léxico (*lexical analysis*)
- Análisis pragmático (*pragmatic analysis*)
- Análisis semántico (*semantic analysis*)
- Análisis sintáctico (*syntactical analysis [parsing]*)
- Integración del discurso (*discourse integration*)

Red neuronal transformadora o de transformador (RNT, *transformer NN*)
Red neuronal recurrente (RNR, *recurrent neural network*)
Retropropagación en el tiempo (BPTT, *backpropagation through time*)
Sesgos (*biases*)
- De confirmación (*confirmation*)
- De etiquetado (*labelling*)
- Histórico (*historical*)
- Lingüístico (*linguistic*)
- De muestreo (*sampling*)
- De selección (*selection*)

Unidad de procesamiento gráfico (GPU, *graphical processing unit*)
Unidad de procesamiento tensorial (TPU, *tensor processing unit*)

Algunas empresas, productos o personas nombradas en este estudio de caso son ficticios y cualquier semejanza con entidades reales es solamente una coincidencia.

Advertencia:

Los contenidos usados en las evaluaciones del IB provienen de fuentes externas auténticas. Las opiniones expresadas en ellos pertenecen a sus autores y/o editores, y no reflejan necesariamente las del IB.

Referencias:

Figura 3 Con autorización de Vinod Sharma. https://vinodsblog.com/2019/01/07/deep-learning-introduction-to-recurrent-neural-networks/.

Figura 4 Rengasamy, D., Jafari, M., Rothwell, B., Chen, X. y Figueredo, G. P., 2020. Deep Learning with Dynamically Weighted Loss Function for Sensor-Based Prognostics and Health Management. *Sensors*, 20(723), páginas 1–21. doi:10.3390/s20030723. Artículo de acceso abierto. Material original adaptado.