Trabajo Fin de Máster: Comparación de modelos pre-entrenados basados en Transformers aplicados a la Búsqueda de Respuestas sobre COVID-19

Trabajo Fin de Máster

Catuxa Irene Fernández Vázquez

Trabajo de investigación para el

Máster en Procesamiento del Lenguaje Natural

Universidad Nacional de Educación a Distancia

Dirigido por el

Prof. Dr. D. Anselmo Peñas

Octubre 2021

Resumen

Debido a la pandemia mundial provocada por el virus COVID-19, en el año 2020 surgen diversos desafíos en el ámbito de la investigación del procesamiento del lenguaje natural para intentar proporcionar sistemas automáticos que respondan de forma eficiente a preguntas relacionadas con la enfermedad. En este proyecto, se presenta una adaptación de un sistema existente de búsqueda de respuestas realizada para dar solución a uno de estos desafíos, el EPIC-QA. Dentro de este ámbito, el desafío propone dos tareas: dar respuesta a un conjunto de preguntas realizadas por usuarios generalistas y a un conjunto de preguntas formuladas por perfiles científicos o médicos. Para ello, el dataset utilizado incluye dos baterías de documentos distintas, una de documentos científicos y otra de textos más generalistas, que se adaptan mejor a cada una de las tareas. El objetivo del proyecto era analizar el impacto del uso de modelos pre-entrenados para el módulo de extracción de respuestas, realizando una comparativa entre tres modelos distintos. El sistema para la experimentación está preparado para responder a cualquier tipo de pregunta y consta de un módulo de recuperación de información a partir de una indexación de los documentos del dataset, seguido del módulo de extracción de respuestas para el que se implementaron las tres configuraciones distintas, con modelos basados en BERT utilizando Transformes y entrenados con baterías de datos adaptadas al ámbito científico y médico: SciBERT, ELECTRA y RoBERTa. Para analizar los resultados, se utilizó el método propuesto en EPIQ-QA basado en nuggets anotados manualmente por los evaluadores. Los resultados obtenidos en la experimentación muestran cómo una de las configuraciones propuestas se adapta mejor a los distintos escenarios planteados, obteniendo en todos los casos las mejores puntuaciones.

Abstract

Due to the global pandemic caused by the COVID-19 virus, different Natural Language Processing research challenges were proposed for providing systems capable to efficiently answer virus linked questions. In this project, an adaptation of an existing answer searching system created to solve one of the mentioned challenges (EPIC-QA) is presented. Within this scope, the challenge proposes two tasks: answering a set of questions asked by general users and a set of questions asked by scientific or medical profiles. To do this, the dataset used includes two different document batteries, one for scientific documents and the other for more generalists texts, which are better adapted to each of the tasks. The main goal of the project was to analyze the impact of the use of pre-trained models for the question answering module, making a comparison between three different models. The system for experimentation is prepared to answer any type of question and consists of an information retrieval module from an indexing of the dataset documents, followed by the question answering module for which the three different configurations were implemented with models based on BERT using Transforms and trained with datasets adapted to the scientific and medical field: SciBERT, ELECTRA and RoBERTa. To analyze the results, the method proposed in EPIQ-QA based on nuggets manually annotated by the evaluators was used. The obtained results show one of the configurations as the best for the proposed scenarios in terms of solution metrics.

Índice

Índice de Figuras

Índice de Tablas

1 Introducción

1.1 Motivación

En respuesta a la pandemia COVID-19, han surgido diversos desafíos y propuestas de investigación en múltiples disciplinas científicas para tratar de colaborar de alguna forma en la gestión de la misma. Un ejemplo de desafío lanzado en el campo del procesamiento del lenguaje natural y en concreto, en el área de la búsqueda de respuestas (*Question Answering*), ha sido el *Epidemic Question Answering* (EPIQ-QA, 2020), cuyo objetivo es que los equipos de investigación desarrollen sistemas capaces de responder automáticamente a preguntas *ad-hoc* sobre la enfermedad COVID-19, su virus causal SARS-CoV-2 y otros coronavirus relacionados.

El rápido aumento de la literatura sobre el coronavirus y las pautas en evolución sobre la respuesta de la comunidad, crean una cantidad de información desafiante no solo para las comunidades científicas y médicas, sino también para el público en general para mantenerse al día sobre los últimos desarrollos. En consecuencia, el objetivo del desafío planteado es evaluar los sistemas en función de su capacidad para proporcionar respuestas de nivel experto oportunas y precisas, tal y como lo esperan las comunidades científicas y médicas, así como respuestas en un lenguaje amigable para el consumidor o público en general.

1.2 Propuesta y objetivos

El primer objetivo del proyecto es desarrollar una adaptación de un sistema ya existente que permita responder a una serie de preguntas propuestas en el reto *Epidemic Question Answering* sobre la enfermedad COVID-19 y términos relacionados, extrayendo la información de la batería de documentos proporcionada, tanto para un perfil experto, partiendo de una batería de 236.035 documentos como *papers,* como para un consumidor general, partiendo de una batería de 117.641 documentos basados en informaciones de webs o publicaciones generalistas.

Las respuestas proporcionadas, deberán estar en forma de oraciones consecutivas extraídas de un solo contexto en un solo documento. En la Tabla 1 se muestran cuatro respuestas de ejemplo de nivel experto y amigables para el consumidor general extraídas para la pregunta: "¿Cuál es el origen de COVID 19?":

What is the origin of COVID 19?	
Consumidor general	Experto
1 COVID-19 is caused by a new coronavirus. **2** Coronaviruses are a large family of viruses that are common in people and many different species of animals, including camels, cattle, cats, and bats. Rarely, animal coronaviruses can infect people and then spread between people such as with MERS-CoV, SARS-CoV, and now with this new virus (named SARS-CoV-2). The SARS-CoV-2 virus is a betacoronavirus, like MERS-CoV and SARS-CoV. **3** All three of these viruses have their origins in bats. **4** The sequences from U.S. patients are similar to the one that China initially posted, suggesting	**1** It is improbable that SARS-CoV-2 emerged through laboratory manipulation of a related SARS-CoV-like coronavirus. As noted above, the RBD of SARS-CoV-2 is optimized for binding to human ACE2 with an efficient solution different from those previously predicted.[7,11] **2** Furthermore, if genetic manipulation had been performed, one of the several reverse-genetic systems available for betacoronaviruses would probably have been used.[19] **3** However, the genetic data irrefutably show that SARS-CoV-2 is not derived from any previously used virus backbone.[20] **4** Instead, we propose two scenarios that can plausibly explain the origin

a likely single, recent emergence of this virus from an animal reservoir.	of SARS-CoV-2: (i) natural selection in an animal host before zoonotic transfer; and (ii) natural selection in humans following zoonotic transfer.
The CDC.	Andersen et al. (2020).

Tabla 1. Ejemplo de cuatro respuestas dada para una pregunta

Para realizar esta adaptación, se partirá de uno de los múltiples desarrollos propuestos por la comunidad científica dentro del campo de la minería de datos y se modificarán los algoritmos utilizados y la batería de pruebas para intentar proporcionar las mejores respuestas posibles.

El segundo objetivo del proyecto es experimentar con varios modelos pre-entrenados para el módulo del sistema encargado de la extracción de respuestas, entrenados con distintas baterías de vocabulario para analizar cuál se adapta mejor a la tarea de este proyecto y proporciona mejores respuestas.

Para realizar esta evaluación de resultados y comparativa entre configuraciones, se utilizará el sistema de evaluación propuesto por EPIQ-QA que permite hasta 1000 respuestas por cada una de las preguntas, calculando varias métricas y permitiendo la comparación de los resultados obtenidos con un conjunto de resultados ideales.

1.3 Estructura del documento

1. Introducción. Este capítulo introduce los principales motivos que han llevado a la realización de este trabajo, así como la problemática y el estado actual de la disciplina. Por último, se presentan las diferentes contribuciones del trabajo realizado.
2. Preliminares. Este capítulo describe en mayor detalle la disciplina en la que se basa el proyecto, presentando su origen y su historia hasta el presente. Se muestran las técnicas actuales más utilizadas para resolver las tareas más relevantes del tema abordado, así como sus debilidades.
3. Arquitectura de la solución. En este capítulo se describe en profundidad el sistema de estudio propuesto.
4. Configuraciones de la experimentación para la extracción de respuestas. En este capítulo se detallan los distintos escenarios y configuraciones empleados para la experimentación.
5. Marco de evaluación. Este capítulo describe la metodología utilizada para evaluar la propuesta realizada, a la vez que presenta los resultados obtenidos al evaluar el método propuesto en diferentes tareas y sobre colecciones de evaluación de distintos dominios.
6. Resultados. Este capítulo analiza y discute en profundidad los resultados obtenidos en la evaluación presentada en el capítulo anterior.
7. Conclusiones y trabajo futuro. Este capítulo recopila las diferentes conclusiones extraídas del trabajo realizado, y propone algunas líneas de trabajo futuro.

2 Preliminares

Los seres humanos utilizamos el lenguaje natural para comunicarnos, transmitir y almacenar el conocimiento adquirido. Actualmente, la mayoría de información se encuentra almacenada digitalmente, pero es difícil aprovechar todos los volúmenes de datos tal cual están almacenados. El Procesamiento del Lenguaje Natural (*Natural Language Processing* o NLP) se encarga de procesar esta información para transformarla a una representación formal que pueda ser manipulada computacionalmente para obtener algo de ella y la vuelve a transformar en lenguaje natural si es necesario (Hernández & Gómez, 2013).

Desde la perspectiva de la inteligencia artificial (*Artificial Intelligence* o IA), el estudio del lenguaje natural tiene dos objetivos (Vásquez, Huerta, Quispe, & Huayna, 2009):

- Facilitar la comunicación con los ordenadores para que puedan acceder a ellos usuarios no especializados a través de intérpretes en un dominio restringido que hace de traductor entre el ordenador y el usuario.
- Modelar los procesos cognoscitivos que entran en juego en la comprensión del lenguaje para diseñar sistemas que realicen tareas lingüísticas complejas (recuperación y extracción de información, traducción automática, búsqueda de respuestas, generación de resúmenes automáticos, minería de datos, análisis de sentimiento, etc.).

Este proyecto está centrado en el campo de la búsqueda de respuestas (*Question Answering o QA*), que forma parte del segundo objetivo descrito.

Algunos de los primeros sistemas de inteligencia artificial eran sistemas de QA. Dos de los sistemas de QA más famosos fueron BASEBALL y LUNAR que se desarrollaron en los años 1960. BASEBALL respondía preguntas sobre los jugadores de béisbol de Estados Unidos en el periodo de un año. LUNAR, se encargaba de responder preguntas sobre análisis geológico de las rocas que trajo el Apollo en su viaje a la Luna. Ambos sistemas eran bastante efectivos, de hecho, LUNAR era capaz de responder al 90% de las preguntas correctamente.

Posteriormente surgieron SHRDU y ELIZA. SHRDLU simulaba la operación de un robot en un mundo virtual y ofrecía la posibilidad de preguntar al robot sobre el estado el mundo virtual. La potencia de este sistema fue la elección de un dominio muy específico y un mundo simple con reglas físicas que eran muy fáciles de codificar. Por otro lado, ELIZA, simulaba una conversación con un psicólogo sobre cualquier tema mediante el uso de reglas muy simples que detectaban palabras importantes en la entrada.

En las décadas de los 70 y 80 se apreció el desarrollo de las teorías de comprensión en lingüística computacional, lo cual permitió el desarrollo de proyectos de comprensión de texto y QA. Un ejemplo de estos sistemas es el Unix Consultant (UC), que respondía preguntas referentes a sistema operativo Unix. Este sistema tenía una base de datos de conocimiento comprensible del dominio.

A finales de los 90, la conferencia anual *Text Retrieval Conference* (TREC) incluía un sistema de QA que se sigue ejecutando hoy en día. Los sistemas que participan en esta competición deben responder cuestiones sobre un tema buscando un trozo de texto que varía de un año para otro. Esta competición encaminó la búsqueda y desarrollo del QA en dominio abierto. Los mejores sistemas del año 2004 lograron un 77% de las preguntas correctas. Un creciente número de sistemas incluyen la web como fuente de texto. Google y Microsoft han empezado a integrar las facilidades del QA en sus buscadores Web (Búsqueda de respuestas, 2021).

En resumen, los primeros sistemas de QA desarrollados en los años 60 eran básicamente interfaces de lenguaje natural para sistemas expertos centrados en dominios específicos. En contraste, los sistemas de QA actuales utilizan documentos de texto como base de conocimiento y combinan diversas técnicas de procesamiento del lenguaje natural.

Por lo tanto, dentro de este ámbito, pueden definirse dos paradigmas principales, centrándose este proyecto en el primero de ellos:

- La respuesta a preguntas basada en la recuperación de información (*Information Retrieval* o IR), también conocida como respuesta a preguntas de dominio abierto, basada en la gran cantidad de texto en la web o en colecciones de artículos científicos. Dada una pregunta de usuario, la recuperación de información se utiliza para encontrar pasajes relevantes. Luego, los algoritmos de comprensión de lectura neuronal leen estos pasajes recuperados y extraen una respuesta directamente de tramos de texto.
- La respuesta a preguntas basada en el conocimiento que consiste en un sistema que construye una representación semántica de la consulta y estas representaciones de significado que se utilizan luego para consultar bases de datos de hechos.

Puesto que el proceso de búsqueda de respuestas es dependiente del motor de búsqueda cuya labor consiste en encontrar los posibles documentos que contengan la respuesta, en principio, un mayor número de documentos tiene mejor rendimiento, es decir, encuentra mejores soluciones. Es posible que la respuesta se encuentre en varios sitios distintos lo que produce dos beneficios: que la respuesta correcta sea la que más veces aparece en los documentos y que se reduzca la carga del procesamiento del lenguaje.

La tarea de Recuperación de Información (*Information Retrieval* o IR), se denomina también recuperación *ad hoc*. Los usuarios plantean una consulta a un sistema de recuperación, que luego devuelve un conjunto ordenado de documentos de alguna colección. Un documento se refiere a cualquier unidad de texto que el sistema indexa y recupera (páginas web, artículos científicos, artículos de noticias o incluso pasajes más cortos como párrafos). Una colección es un conjunto de documentos que se utilizan para satisfacer las solicitudes del usuario. Un término equivale a una palabra en una colección, pero también puede incluir frases. Por último, una consulta representa la necesidad de información de un usuario expresada como un conjunto de términos (Jurafsky & Martin, 2020).

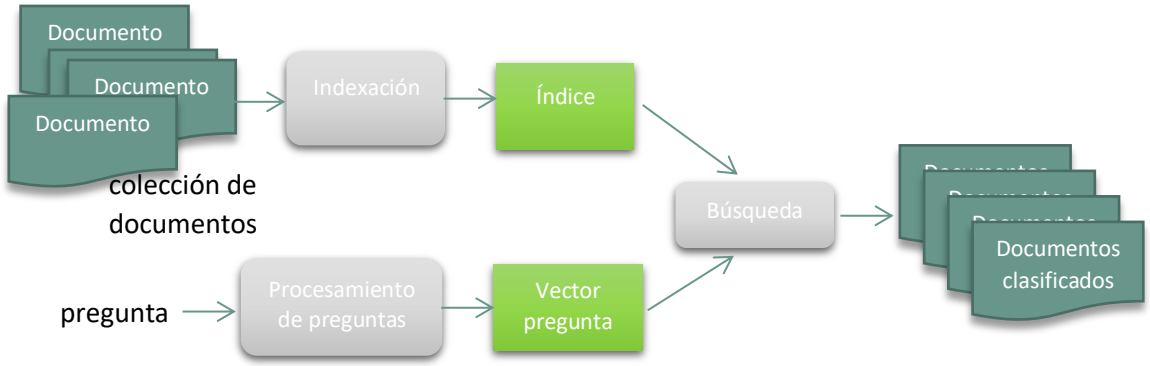


Figura 1. Arquitectura de un sistema de IR ad-hoc

Dentro de la arquitectura QA, se encuentran los siguientes módulos:

- El primero de ellos se corresponde con la clasificación de preguntas con técnicas de NLP y esta información se proporciona como entrada para el siguiente módulo.
- El segundo de los módulos es la recuperación de documentos, explicado de forma más extensa en el punto anterior.
- El tercero filtra los documentos ya obtenidos, para obtener los fragmentos concretos que contienen la respuesta. Para ello, se realiza una comparación de tipos entre la respuesta esperada a partir del primer módulo y el tipo del fragmento de texto dado.
- Por último, se realiza ejecuta el módulo de extracción, que determina dentro de las respuestas candidatas obtenidas cual es aquella que por "pistas" presenta la respuesta esperada.

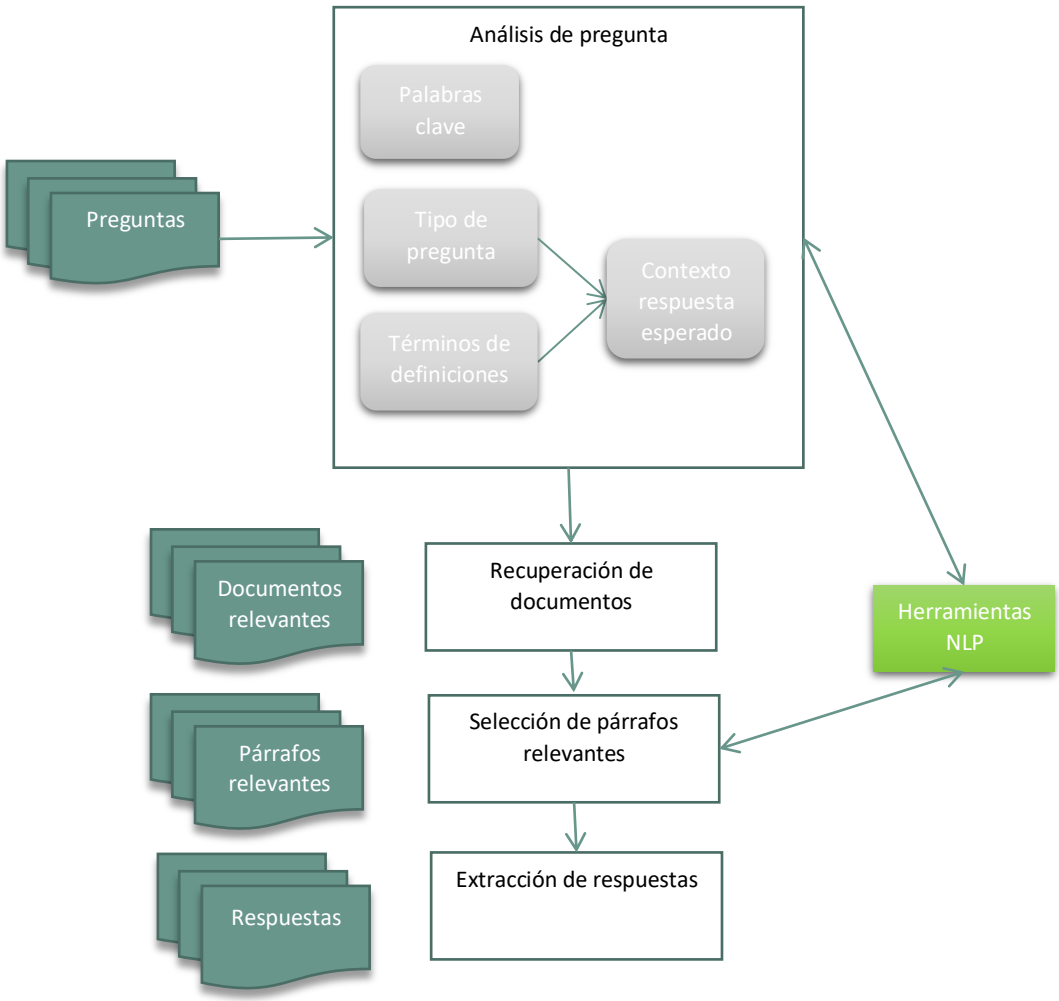


Figura 2. Arquitectura de un sistema QA

2.1 Preprocesamiento de los datos

Dentro del procesamiento del lenguaje natural, suelen aplicarse ciertos procesos a las cadenas de textos antes de comenzar a trabajar con los modelos. No todos los procesos se aplican siempre y dependiendo de la finalidad del modelo, del tipo de datos y de las necesidades de los modelos, pueden surgir otros preprocesos o no aplicarse todos. Algunos de los preprocesos más comunes son los siguientes:

Tokenización

Convierte las cadenas u oraciones en listas de palabras o *tokens*.

Eliminación de signos de puntuación

En la mayoría de los casos, se eliminan signos de puntuación como comas o puntos y coma, e incluso puntos o interrogaciones cuando no aportan nada a las necesidades del modelo.

Eliminación de stop words

Palabras frecuentes como “The”, “is”, etc. que no tienen semántica específica son eliminadas. Suelen utilizarse listas generales de *stop words* que se encuentran disponibles para los distintos idiomas.

Stemming

Las palabras se reducen a una raíz eliminando la inflexión mediante la eliminación de caracteres innecesarios, generalmente un sufijo.

Por ejemplo:

```
Biblioteca → bibliotec
Bibliotecario → bibliotec
```

Lematización

Es otro enfoque para eliminar la inflexión, determinando la parte del discurso y utilizando la base de datos detallada del idioma.

Por ejemplo:

```
Comía → comer
Comiste → comer
```

Por lo tanto, tanto *stemming* como lematización ayudan a reducir palabras como “estudios”, “estudiar” a una forma de base común o la palabra raíz “estudio”.

2.2 Modelos pre-entrenados

La idea que subyace tras los modelos de lenguaje pre-entrenados es crear una caja negra que entienda el idioma para poder realizar luego una tarea específica en ese idioma, como puede ser la de respuesta a preguntas (QA). La idea es crear una máquina equivalente a un ser humano "culto”:

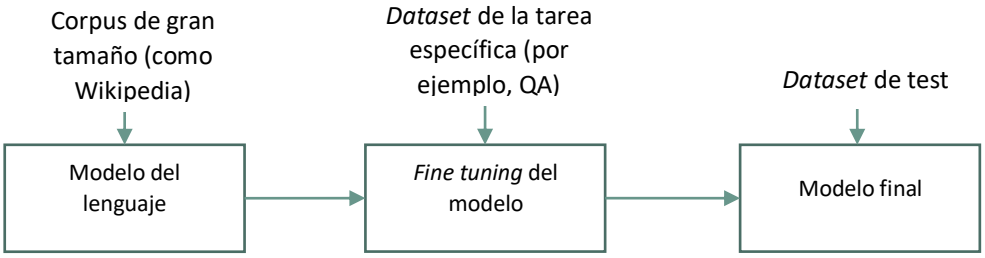


Figura 3. Estructura de funcionamiento de modelos pre-entrenados

El modelo del lenguaje se alimenta primero con una gran cantidad de datos sin anotar (por ejemplo, el volcado completo de Wikipedia). Esto permite que el modelo aprenda el uso de varias palabras y cómo se escribe el idioma en general. Este modelo se transfiere a una tarea de

NLP donde se alimenta con otro conjunto de datos más pequeño y específico de la tarea, que se utiliza para ajustar y crear el modelo final capaz de realizar la tarea específica, como QA.

Con este tipo de modelos, se facilita la realización de tareas de NLP, especialmente en casos en los que no se dispone de tiempo o recursos para construir modelos de procesamiento de lenguaje natural desde cero. Además, un modelo que entrena solo con el conjunto de datos específico de la tarea necesita comprender el lenguaje y la tarea utilizando un conjunto de datos comparativamente más pequeño. Sin embargo, el modelo de lenguaje ya comprende el lenguaje porque ha procesado grandes volúmenes durante la primera fase del proceso. Por lo tanto, el modelo de lenguaje puede ajustarse directamente a sí mismo para adaptarse a la tarea requerida y llegar a funcionar mejor que otros.

Existen dos aproximaciones diferentes para la fase de ajuste del modelo. La primera aproximación, es el método denominado *Word embedding,* en donde las palabras o frases del lenguaje natural son representadas como vectores de números reales. Conceptualmente implica el encaje matemático de un espacio con una dimensión por palabra a un espacio vectorial continuo con menos dimensiones.

Se han propuesto muchos algoritmos diferentes para crear estas incrustaciones (*embeddings*), entrenando previamente los modelos en un conjunto de datos más grande separado para capturar la esencia del lenguaje. Por ejemplo, las incrustaciones de Word2Vec ganaron una popularidad increíble y estas incrustaciones se utilizaron directamente para una serie de tareas en la NLP. Sin embargo, estas representaciones de palabras se aprenden en un contexto generalizado y no representan información específica de la tarea.

El uso directo de incrustaciones previamente entrenadas puede disminuir el tamaño general del modelo, pero obliga a utilizar solo representaciones de palabras generalizadas. Sin embargo, en la segunda aproximación, denominada ajuste fino del modelo de lenguaje (*fine tuning*), permite al usuario ajustar estas incorporaciones o representaciones de palabras mediante el entrenamiento en el conjunto de datos específico de la tarea.

Por ejemplo, la representación de la palabra "actual" podría tener una buena relación tanto con "noticias" como con "electricidad" en un contexto generalizado. Sin embargo, para una tarea específica que habla de circuitos eléctricos, permitir que el modelo ajuste las representaciones de palabras de modo que "corriente" y "electricidad" coincidan mejor, puede ayudar a mejorar el rendimiento del modelo.

2.3 Recuperación de Información

El proceso de Recuperación de Información consiste en comparar una *query* (consulta) del usuario con una gran colección de documentos, devolviendo una lista ordenada de documentos que mejor se ajustan a la consulta.

Las primeras aproximaciones a la recuperación de información trabajan sobre la información en bruto, siendo simplemente comparaciones de patrones con todo el texto que se haya obtenido. Pero esta solución es inmanejable, tanto por las cantidades de texto como por la pobreza de los resultados obtenidos. En general, todos los documentos que se quieran utilizar para recuperación de la información tendrán que ser categorizados e indexados.

Una vez que los documentos se categorizan e indexan, aparece el paso en el que se recupera la información. Se necesita decidir qué documentos son relevantes para la búsqueda del usuario y

ordenarlos en función de esa relevancia. Algunos de los modelos más significativos para realizar esta tarea son los siguientes (Gabriel Jaimes, 2005):

2.3.1 Modelo booleano

El modelo booleano es el más usado históricamente. Está basado en la teoría de conjuntos y el álgebra booleana. Formalmente el modelo booleano se divide en:

- D: conjunto de palabras del documento (términos de indexación)
- Q: expresión booleana (operadores: AND, OR y NOT)
- F: álgebra booleana sobre conjuntos de términos y documentos
- R: un documento es relevante para la consulta dada si satisface la expresión booleana de acuerdo con el álgebra

Se asigna un valor de similitud (número real) dado por la función $S : D \times P \rightarrow R$, donde D y P son conjuntos de términos. Por lo general, un documento es representado como un registro de que contiene ceros (0) y unos (1), dependiendo de si contiene un término o no.

Este modelo es un modelo simple que basa su efectividad en dividir los términos de la búsqueda en conjuntos, por ello es muy fácil de implementar y entender. Donde reside algo más de dificultad es en las expresiones booleanas anidadas, pero cualquier ordenador maneja las expresiones booleanas muy fácilmente. Para el modelo booleano todos los términos de una búsqueda tienen la misma importancia y relevancia, ya que no realiza ningún tipo de ranking de los términos de indexación.

2.3.2 Modelo de espacio vectorial

En el modelo de espacio vectorial, los documentos y las búsquedas se interpretan como vectores de términos, representando cada término en el vector con un peso w dentro de ese documento. La función de similitud entre el documento y una búsqueda es el coseno del ángulo entre los vectores que los representan:

$$sim(d_i, d_j) = \overrightarrow{d_\iota} \cdot \overrightarrow{d_j} = \sum_{r=1}^{k} w_{r,i} \times w_{r,j}$$

siendo $\{t_1, t_2, \ldots, t_k\}$ el conjunto de los términos, $\{d_1, d_2, \ldots, d_N\}$ el conjunto de los documentos y un documento d_i como el vector $d_i \rightarrow \sim d_i = \{w(t_1, d_i), \cdots, w(t_k, d_i)\}$, donde $(w(t_r, d_i)$ es el peso del término t_r en el documento d_i.

La funcionalidad de este modelo estriba en la elección correcta de los pesos de cada término. Para que la recuperación de información sea efectiva, se tienen que elegir unos pesos mayores para las palabras que tienen más relevancia en el documento.

Para modelizar este comportamiento, los documentos se pueden utilizar modelizando los documentos en *clusters*. En esta modelización el documento es una colección de C objetos. Las búsquedas se toman como descripciones vagas de un subconjunto A de la colección C. El objetivo es dividir C en dos subconjuntos (A y ~A) para lo que se determinan las características de los objetos que describen A de una forma más eficiente, así como las que diferencian A de ~A. En los documentos se utilizan las frecuencias de los términos en el documento, así como la frecuencia en la colección.

Un modelo para el reparto de pesos típico es el TF-IDF (ver 8.5), donde el peso w es:

$$w_{t,d} = tf_{t,d} \times idf_t$$

Es importante que aparte de este reparto de pesos se realice una normalización del tamaño de los documentos, si no los documentos más largos se verían beneficiados, gracias a que tienen más frecuencia de términos y más términos.

En resumen, las principales ventajas de estos modelos son la obtención de documentos ordenados por un ranking, el uso de términos de búsqueda con importancia baremada y la obtención de resultados de coincidencia parcial con la búsqueda.

2.3.3 Modelo probabilístico

El modelo booleano presenta ciertos problemas, ya que ofrece resultados de todo o nada (o pertenece al subconjunto de documentos con los términos de la expresión booleana, o no pertenece). Esto excluye documentos que son relevantes, pero están fuera de los subconjuntos ya que los términos de búsqueda y los términos de indexación pueden divergir.

El problema radica en que en el subconjunto R de documentos relevantes a una *query* q, la pertenencia de los documentos a R es incierta. Por ello se puede tomar una aproximación probabilística en la que los documentos se ordenen en orden decreciente de probabilidad de relevancia a la información requerida.

Este modelo plantea dos dificultades principales:

- Las evidencias para la ordenación se basan en una representación difusa: el proceso consiste en evaluar la probabilidad de relevancia basándose en las ocurrencias de los términos de la búsqueda en los documentos (similar al modelo booleano). Normalmente se empieza con una estimación y más adelante se refina a través del *feedback* de los usuarios.
- No se puede computar la probabilidad exacta: es un proceso demasiado complejo así que el modelo se basa en simplificaciones y aproximaciones.

El modelo se basa en:

- D: términos de la búsqueda, los términos se toman como ocurrencias booleanas (Presente - 1 / No presente - 0) en el documento.
- R: conjunto de documentos relevantes.
- $P(R \mid d_j)$: probabilidad de que el documento dado sea relevante
- $P(\sim R \mid d_j)$: probabilidad de que el documento dado sea relevante

En general el modelo se basa en el cálculo de una función rsv (*Retrieval Status Value*) que es la ratio entre la probabilidad de que sea relevante y la probabilidad de que no lo sea. A partir de esas ratios, se puede calcular un vector de probabilidades para todos los términos de D. Pero esto tiene un problema, R tiene un valor desconocido y difícil de calcular. Tomando valores estadísticos, se toma $R = 0.5$ y $\sim R$ como la frecuencia inversa del documento:

$$IDF = \log\left(\frac{d}{df_j}\right)$$

donde d es el número total de documentos y df_j es el número de documentos que contienen el término. En esta primera aproximación, se puede utilizar este valor IDF como pesos para hacer el ranking. A medida que el usuario utiliza el sistema, estos pesos se ajusta en función del *feedback* del usuario.

2.4 Búsqueda de Respuestas (*Question Answering*)

Los sistemas de QA son considerados de los más complejos en torno a la recuperación de información ya que además de buscar una información en una cantidad más o menos grande de documentos, se debe extraer de dichos documentos un fragmento de texto que responda a una pregunta dada en lenguaje natural.

Los sistemas actuales suelen incluir un módulo de clasificación de preguntas que se encarga de determinar el tipo de pregunta y respuesta (lugar, nombre, fecha, etc.). Tras analizar la pregunta, el sistema utiliza diversos módulos que aplican complejas técnicas de procesamiento de lenguaje natural. Tras ello, se aplica un módulo de recuperación de documentos que utiliza motores de búsqueda para identificar documentos y párrafos en el documento que puedan contener la respuesta a la pregunta. Posteriormente, se aplica un filtro que se encarga de seleccionar pequeños trozos de texto que contengan cadenas del mismo tipo al esperado. Para finalizar, el módulo de extracción de respuestas es el encargado de buscar pistas en el texto que determinen si una respuesta candidata es correcta.

2.4.1.1 BERT model

El modelo BERT (Representación de codificador bidireccional de Transformers) está basado en redes neuronales para el pre-entrenamiento de NLP. Fue creado y publicado en 2018 por Jacob Devlin y está integrado en el buscador de Google que lo utiliza para responder a las consultas de los usuarios. Obtuvo el mejor rendimiento en 11 tareas clásicas de NLP por lo que desde entonces, ha sido la fuente de múltiples modelos de lenguaje que se derivan de él.

La idea principal de BERT es realizar una fase de entrenamiento previo sin supervisión en un gran corpus de texto genérico (normalmente se utilizan artículos de Wikipedia o colecciones de libros) para aprender representaciones lingüísticas razonables. Durante una fase de ajuste fino, las representaciones aprendidas previamente se utilizan como línea de base para un entrenamiento específico del problema.

BERT utiliza Transformers, un mecanismo que aprende las relaciones contextuales entre palabras (o subpalabras) en un texto. En su forma básica, Transformer incluye dos mecanismos separados: un codificador que lee la entrada de texto y un decodificador que produce una predicción para la tarea. Dado que el objetivo de BERT es generar un modelo de lenguaje, solo es necesario el mecanismo del codificador.

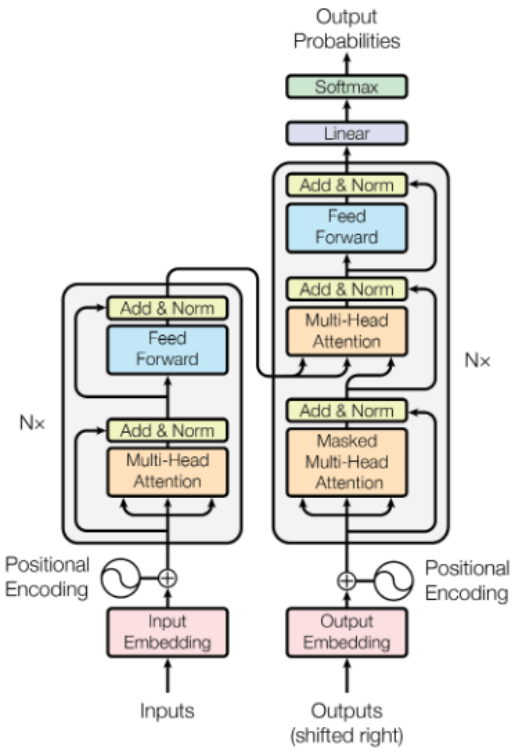


Figura 4. Esquema de la arquitectura del modelo BERT (Vaswan, 2017)

A diferencia de los modelos direccionales, que leen la entrada de texto secuencialmente (de izquierda a derecha o de derecha a izquierda), el codificador Transformer lee la secuencia completa de palabras a la vez. Por lo tanto, se considera bidireccional, aunque sería más exacto decir que no es direccional. Esta característica permite que el modelo aprenda el contexto de una palabra en función de todo su entorno (izquierda y derecha de la palabra).

La entrada es una secuencia de tokens, que primero se incrustan en vectores y luego se procesan en la red neuronal. La salida es una secuencia de vectores de tamaño H, en la que cada vector corresponde a un token de entrada con el mismo índice.

Al entrenar modelos de lenguaje, existe el desafío de definir un objetivo de predicción. Muchos modelos predicen la siguiente palabra en una secuencia (por ejemplo, "El niño llegó a casa de ___"), un enfoque direccional que limita inherentemente el aprendizaje del contexto. Para superar este desafío, BERT utiliza dos estrategias de capacitación:

Estrategia 1: LM enmascarado (MLM)

Antes de introducir secuencias de palabras en BERT, el 15% de las palabras de cada secuencia se reemplazan con un token [MASK]. Luego, el modelo intenta predecir el valor original de las palabras enmascaradas, basándose en el contexto proporcionado por las otras palabras no enmascaradas en la secuencia. En términos técnicos, la predicción de las palabras de salida requiere:

1. Agregar una capa de clasificación sobre la salida del codificador.
2. Multiplicar los vectores de salida por la matriz de incrustación, transformándolos en la dimensión de vocabulario.
3. Calcular la probabilidad de cada palabra del vocabulario con softmax. En matemáticas, la función softmax, o función exponencial normalizada, es una generalización de la

Función logística. Se emplea para "comprimir" un vector K-dimensional, z, de valores reales arbitrarios en un vector K-dimensional, $\sigma(z)$, de valores reales en el rango [0, 1]. La función está dada por:

$$\sigma\colon \mathbb{R}^K \to [0,1]^K$$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} para\ j = 1, \ldots, K$$

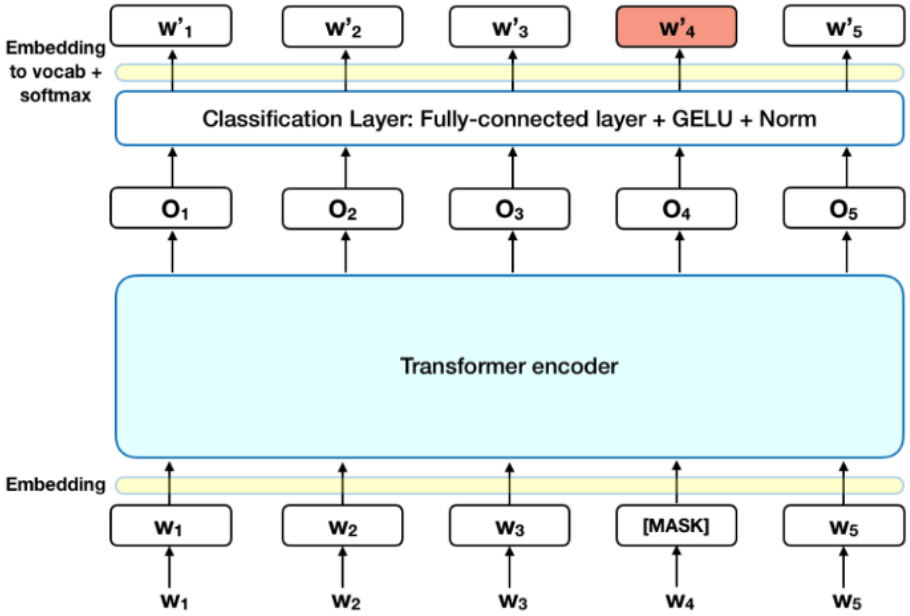


Figura 5. Estrategia de LM enmascarado (MLM) (Jacob Devlin, 2019)

La función de pérdida BERT toma en consideración solo la predicción de los valores enmascarados e ignora la predicción de las palabras no enmascaradas. Como consecuencia, el modelo converge más lento que los modelos direccionales, una característica que se ve compensada por su mayor conocimiento del contexto.

Estrategia 2: Predicción de la siguiente oración (NSP)

En el proceso de entrenamiento de BERT, el modelo recibe pares de oraciones como entrada y aprende a predecir si la segunda oración del par es la oración subsiguiente en el documento original. Durante el entrenamiento, el 50% de las entradas son un par en el que la segunda oración es la oración posterior en el documento original, mientras que en el otro 50% se elige una oración aleatoria del corpus como segunda oración. El supuesto es que la oración aleatoria se desconectará de la primera oración.

Para ayudar al modelo a distinguir entre las dos oraciones en entrenamiento, la entrada se procesa de la siguiente manera antes de ingresar al modelo:

1. Se inserta un token [CLS] al principio de la primera oración y un token [SEP] al final de cada oración.
2. A cada ficha se agrega una inserción de oración que indica la oración A o la oración B. Las incrustaciones de oraciones son similares en concepto a las incrustaciones de fichas con un vocabulario de 2.
3. Se agrega una incrustación posicional a cada token para indicar su posición en la secuencia.

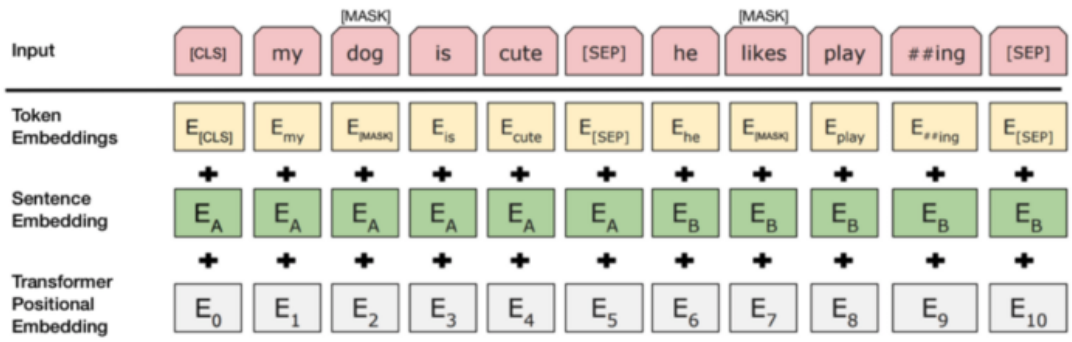


Figura 6. Estrategia de predicción de la siguiente oración (NSP) (Jacob Devlin, 2019)

Para predecir si la segunda oración está realmente conectada con la primera, se llevan a cabo los siguientes pasos:

1. Toda la secuencia de entrada pasa por el modelo Transformer.
2. La salida del token [CLS] se transforma en un vector con forma de 2 × 1, utilizando una capa de clasificación simple (matrices aprendidas de pesos y sesgos).
3. Cálculo de la probabilidad de IsNextSequence con softmax.

Al entrenar el modelo BERT, Masked LM y Next Sentence Prediction se entrenan juntos, con el objetivo de minimizar la función de pérdida combinada de las dos estrategias.

Fine tuning del modelo

BERT se puede usar para una amplia variedad de tareas de lenguaje, agregando solo una pequeña capa al modelo central:

1. Las tareas de clasificación, como el análisis de sentimientos, se realizan de manera similar a la clasificación de la siguiente oración, agregando una capa de clasificación en la parte superior de la salida de Transformer para el token [CLS].
2. En las tareas de respuesta a preguntas, el software recibe una pregunta sobre una secuencia de texto y debe marcar la respuesta en la secuencia. Con BERT, se puede entrenar un modelo de preguntas y respuestas aprendiendo dos vectores adicionales que marcan el comienzo y el final de la respuesta.
3. En el Reconocimiento de entidades nombradas (NER), el software recibe una secuencia de texto y se requiere que marque los distintos tipos de entidades (Persona, Organización, Fecha, etc.) que aparecen en el texto. Con BERT, se puede entrenar un modelo NER alimentando el vector de salida de cada token en una capa de clasificación que predice la etiqueta NER.

En el entrenamiento de ajuste fino, la mayoría de los hiperparámetros permanecen igual que en el entrenamiento BERT. Para extraer el intervalo de respuesta correcto, se obtienen los pares de índices de inicio y final de mayor probabilidad en el siguiente código. Como la longitud de entrada para el modelo BERT es fija, se aplica un enfoque de ventana deslizante para secuencias que son más largas de 384 *subtokens* (BERT Transformers, 2021).

La implementación original de BERT tiene dos tamaños de modelo: un modelo base BERT con un tamaño de capa media oculta de dimensiones $H = 768$ y un modelo grande BERT con un tamaño de capa oculta de $H = 1024$ dimensiones. Los bloques de codificador / decodificador constan de 12 y 14 capas de transformadores, respectivamente. Todas las capas contenidas en el modelo tienen el mismo tamaño (Zhang).

2.4.1.2 Conjuntos de datos

Como ya se explicó en anteriores apartados, los modelos pre-entrenados requieren de una gran cantidad de información para aprender de ella y poder realizarse luego sobre ellos el *fine tuning* que proporcione un mejor ajuste a la tarea realizada. A continuación, se detallan algunos de estos conjuntos de datos o *datasets* más conocidos y utilizados para tareas similares a las desarrolladas en este proyecto.

2.4.1.2.1 SQuAD 2.0

El conjunto de datos SQuAD (*Stanford Question Answering Dataset*) es un conjunto de datos de comprensión de lectura, que consta de preguntas formuladas por los trabajadores de Wikipedia en un conjunto de artículos de la plataforma, donde la respuesta a cada pregunta es un segmento de texto, o tramo, del pasaje de lectura correspondiente, o la pregunta puede no disponer de respuesta.

SQuAD2.0 combina las 100.000 preguntas de la versión SQuAD1.1 con más de 50.000 preguntas sin respuesta escritas de forma contradictoria por los trabajadores de la plataforma para que parezcan similares a las que se pueden responder. (SQuAD, 2021)

2.4.1.2.2 QuAC

El conjunto de datos QuAC (*Question Answering in Context*) consiste en un diálogo interactivo entre: un estudiante que plantea una secuencia de preguntas de forma libre para aprender tanto como sea posible sobre un texto oculto de Wikipedia y un maestro que responde a las preguntas proporcionando extractos breves (tramos) del texto. QuAC presenta desafíos que no se encuentran en los conjuntos de datos de comprensión de máquinas existentes: sus preguntas suelen ser más abiertas, incontestables o solo significativas dentro del contexto del diálogo.

QuAC comparte muchos principios con SQuAD 2.0, como la evaluación basada en el intervalo y las preguntas incontestables, pero incorpora un nuevo componente de diálogo. (QuAC, 2021)

2.4.1.2.3 BioASQ 6b / 7b

Este dataset pertenece a un conjunto de datos preprocesados de un modelo que participó en el desafío BioASQ Challenge 7b. Está basado en BioBERT, un modelo de representación del lenguaje para el dominio biomédico, con algunas pequeñas modificaciones.

BioBERT se basa en conjuntos de datos como NER (*Named Entity Recognition*), el conjunto de datos de enfermedades del NCBI (*National Center for Biotechnology Information*), el conjunto de datos BC2GM (*BioCreative II Gene Mention Recognition*) y una base de datos de interacciones de proteínas llamada CHEMPROT, entre otros. (Jinhyuk Lee, 2019)

3 Arquitectura de la solución

El sistema utilizado en el proyecto se basa en un esquema realizado en código Python proporcionado como solución al *COVID-19 Open Research Dataset Challenge* propuesto en la comunidad Kaggle[1] (NQA, 2020).

Puesto que el objetivo de este proyecto es combatir la sobrecarga de información relacionada con la temática COVID19, el sistema propuesto devuelve respuestas breves y solo se responde cuando la calidad de la recuperación y las respuestas es satisfactoria. Además, el usuario puede acceder a todos los párrafos y documentos si necesita más detalles.

El sistema de partida utilizaba como conjunto de datos el CORD-19 disponible gratuitamente (ORDC, 2020), que contiene metadatos de más de 51.000 artículos científicos (el texto completo también está disponible para alrededor de 40.000 de ellos) sobre COVID-19, SARS-CoV-2 y coronavirus relacionados. Como preguntas, utilizaba las propuestas en el *COVID-19 Open Research Dataset Challenge*, agrupadas en función de la temática de las preguntas.

Una de las principales características de este sistema es que no está diseñado para preguntas específicas y se puede usar fácilmente para responder cualquier otra pregunta.

Para este proyecto, se han utilizado tanto el conjunto de datos como las preguntas proporcionados en el reto EPIC-QA (EPIQ-QA, 2020), eliminando el agrupamiento entre preguntas y utilizando directamente todas las preguntas propuestas al mismo nivel. Los detalles sobre el conjunto de datos de entrada se desarrollan con más detalle en 5.1.

En sistema proporciona de forma opcional un filtro para eliminar artículos de la batería de documentos de entrada que no estén directamente relacionados con el tema central de estudio. En este caso, podrían filtrarse artículos que tratan sobre virus distintos al COVID-19 (por ejemplo, SARS-CoV y MERS). Sin embargo, se ha optado por no realizar ningún filtro y utilizar la batería completa de artículos.

El sistema implementado está formado por tres componentes principales (Figura 7):

- El primer componente es un sistema de recuperación de información (IR), basado en el algoritmo de búsqueda clásico BM25F. Este sistema indexa resúmenes y párrafos sobre el texto completo de los trabajos.
- El segundo componente del sistema es un sistema de QA que, dada una pregunta en lenguaje natural y un párrafo, devuelve la respuesta a la pregunta en el párrafo o "No sé" si considera que no puede proporcionar una buena respuesta. El sistema implementado se basa en técnicas de redes neuronales.
- Finalmente, el tercer componente agrega los resultados del sistema de QA. Para cada tarea, ejecuta los sistemas IR y QA sobre todas las preguntas. El sistema de IR devuelve un número configurable de párrafos por respuesta.

Para combatir la sobrecarga de información, se tomaron las siguientes decisiones:

  - Descartar los párrafos donde el sistema de QA devuelve "No sé".
  - Devolver la mejor respuesta para cada uno de los mejores párrafos, es decir, un número configurable por pregunta, aunque en la salida visual generada en HTML solo se destaquen las tres mejores respuestas en colores distintos.

[1] Kaggle es una comunidad en línea de científicos de datos y profesionales del aprendizaje automático.

- Descartar las preguntas que no parecen ser adecuadas para la tecnología actual, estimadas como aquellas preguntas que reciben un porcentaje determinado de respuestas de "No sé".

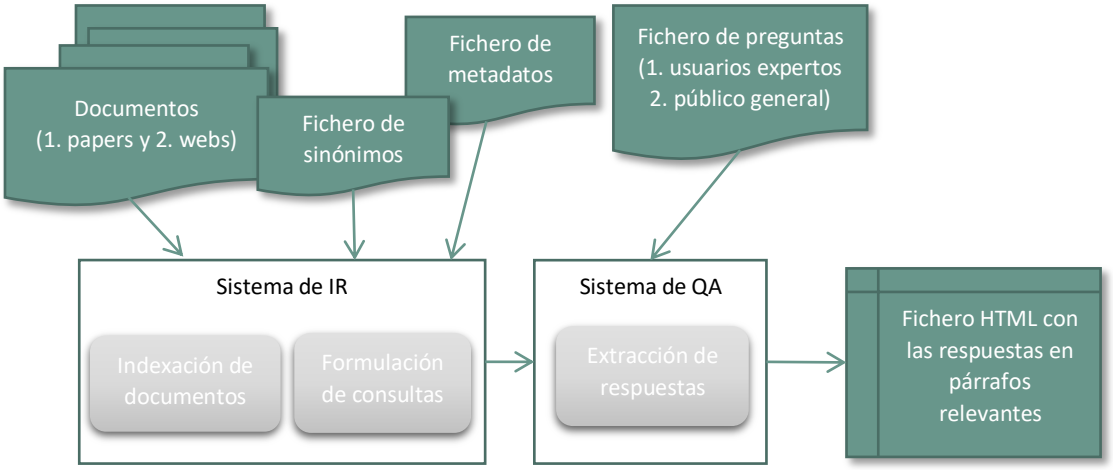


Figura 7. Esquema de módulos de la solución

A continuación, se describen con más detalle cada uno de los componentes mencionados anteriormente.

3.1 Indexación y recuperación

El sistema de recuperación de información es una herramienta que busca documentos que son relevantes para una necesidad de información a partir de una colección de documentos. Este sistema tiene dos módulos principales: el sistema de indexación y el sistema de consultas.

El primer módulo se encarga de crear la estructura de datos de entrada del sistema, que es el índice. El segundo componente es aquel con el que los usuarios interactúan enviando una consulta en función de su necesidad de información, y basándose en esta consulta y utilizando el índice, se recuperan los documentos relevantes.

El índice es una estructura de datos que permite buscar información en una colección de documentos de una manera muy eficiente: enumera, para cada palabra, todos los documentos que la contienen. Para crearlo, es necesario definir el esquema del índice que indica los campos que contiene. Un campo es una pieza de información para cada documento en el índice, por ejemplo: ruta del documento, título y texto.

En este proyecto, se indica que los campos “título” y “texto” contienen información de tipo textual sobre la que pueden realizarse búsquedas. Además, se aplica el *Stemming Analyzer* a estos campos de texto, para que se troceé en tokens todo el texto, se conviertan a minúsculas, se aplique un filtro para eliminar palabras demasiado comunes, y finalmente, se aplique un algoritmo de derivación.

Una vez construido el esquema, se crea el índice que es almacenado en el directorio de salida del espacio de trabajo. Luego, se agregan los documentos al índice, no solo los resúmenes que se encuentran en el fichero de metadatos, sino también el texto completo proporcionado en formato JSON o PDF. Como tener documentos más cortos es mejor para el sistema de respuesta

desarrollado en los siguientes puntos, no se indexa todo el texto del documento, sino que la unidad de indexación es cada uno de los párrafos del texto completo.

El siguiente paso consiste en crear una función que, dada una pregunta, un *dataframe* que contiene información de metadatos y un número máximo de documentos como entrada, utiliza esta consulta para recuperar documentos relevantes que fueron indexados en la sección anterior.

En esta función se configura el algoritmo utilizado para la puntuación (en este proyecto se utiliza el algoritmo BM25F predeterminado), y también se configura el analizador de consultas a usar, definiendo el campo predeterminado para buscar (en este caso, los campos marcados como textuales). Como último paso, se ejecuta la consulta y se obtienen los documentos más relevantes en el índice (respetando el máximo de documentos indicados en la configuración).

El resultado de la función es un *dataframe*, donde se almacenan los siguientes datos para cada párrafo relevante: "id", "fecha", "revista u origen", "título", "texto, url" (si la tiene) y "puntuación".

3.1.1.1 BM25

Okapi BM25 es una función de ranking utilizada en recuperación de información para la asignación de relevancia a los documentos en un buscador, es decir, es una función que permite ordenar por relevancia los documentos que contienen las palabras que el usuario busca.

Esta función está basada en los modelos probabilísticos de recuperación de información, concretamente en el BIR (*Binary Independent Retrieval*) desarrollado por Stephen E. Robertson y Karen Spärck Jones en los años 70. El nombre de Okapi viene del primer sistema que implementó esta función de ranking, el cual fue desarrollado por Stephen Walker en la City University de Londres. (García, 2011)

BM25 se basa en el concepto de bolsa de palabras (*bag of words*) mediante el cual se representan los documentos que se desean ordenar en función de su relevancia con una consulta dada.

Dada una consulta Q_1, que contiene las palabras clave $q_1, \dots, q_n$, el valor de relevancia asignado mediante la función BM25 para el documento D será:

$$score(D,Q) = \sum_{i=1}^{n} IDF(q_i) \frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot(1-b+b\cdot\frac{|D|}{avgdl})}$$

donde $f(q_i,D)$ es la frecuencia de aparición en el documento D de los términos que aparecen en la consulta Q, $|D|$es la longitud del documento D (en número de palabras), y $avgdl$ es la longitud media de los documentos en la colección sobre la cual se realiza la búsqueda. k_1 y b son parámetros que permiten ajustar la función a las características concretas de la colección con la que se trabaja. Aunque estos parámetros suelen depender de las características concretas de cada colección, normalmente se asignan los valores $k_1 = 2$ o $k_1 = 1.2$ y $b = 0.75$, los cuales se han establecido a partir de los experimentos que durante años se han realizado en las conferencias TREC. $IDF(q_i)$ es el peso IDF de las palabras clave que aparecen en la consulta Q. Normalmente el IDF se calcula mediante la siguiente función:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}$$

donde N es el número total de documentos en la colección, y $n(q_i)$ es el número de documentos que contienen la palabra clave q_i.

Alguno de los problemas que presenta este método son los siguientes:

- La mayoría de los modelos de recuperación de información, consideran los documentos como texto no estructurado. Sin embargo, los documentos se estructuran con frecuencia y constan de campos o secciones específicos, como etiquetas de marcado (encabezado, cuerpo, título, descripción, párrafos, divisiones, formularios, etc.). Para aplicar a estos documentos un modelo BM25 pueden concatenarse todos los campos del documento en un pseudodocumento no estructurado o tratar cada tipo de campo como colecciones de campos de documentos no estructurados. Por ejemplo, para artículos de revistas científicas se pueden construir colecciones de títulos, resúmenes y secciones. Sin embargo, ese enfoque destruye la relación de no linealidad entre los pesos de los términos y las frecuencias de los términos, restaurando la independencia temporal.
- El segundo problema es que puntuar las ponderaciones de los términos mediante la combinación de tipos de campos abre la pregunta de cómo recopilar estadísticas de campo globales como valores IDF para los campos individuales. Por ejemplo, los títulos son campos cortos y el cuerpo es un campo grande. Dado que los términos de uso frecuente en los cuerpos del texto pueden aparecer raramente en los campos de título, estos deben recibir un alto peso en la puntuación del título. Resulta que debido a un término utilizado con frecuencia se define en relación con un tipo de campo, el resultado sería un IDF muy inestable.
- El tercer problema es que no habría una manera fácil de interpretar el significado de fusionar tipos de campo ponderados uniformemente. Por ejemplo, debido a la naturaleza de no linealidad de las frecuencias de los términos, el establecimiento de todos los pesos de campo a 1 no restaura el escenario no estructurado de fusionar todos los campos de manera equivalente en un gran campo no estructurado.
- Un cuarto problema que surge al construir colecciones de tipos de campo es cómo normalizar la longitud de los campos. La normalización de la longitud en BM25 es para tener en cuenta la verbosidad y el alcance de los documentos. Pero no está claro si la normalización basada en la verbosidad y el alcance debe aplicarse a los diferentes campos, o si se debe utilizar la longitud completa del documento. Además, está la cuestión de cómo optimizar k_1 y b para cada tipo de campo.

3.1.1.2 BM25F

Una forma de solventar los problemas del método Okapi BM25, consiste en ponderar las frecuencias de los términos de acuerdo con su importancia de campo, combinándolos, y luego usando las pseudo-frecuencias resultantes.

Por ejemplo, siendo s un campo del documento, v su peso y dado un documento que se descompone en dos campos o secuencias: título y cuerpo. Si se asigna una ponderación de 6 a los términos del título y una ponderación de 2 a los términos que aparecen en el cuerpo, esto equivale a reemplazar el documento por sí mismo pero esta vez repitiendo el título seis veces y el cuerpo original dos veces.

La pseudo-frecuencia resultante de un término es una combinación lineal de estos campos ponderados:

$$\widetilde{f_{i,J}} = \sum_{s=1}^{s} v_s f_{i,s}$$

Aplicando esta fórmula a todos los términos del documento, la nueva longitud del documento es:

$$\widetilde{dl_J} = \sum_{i}^{m} \widetilde{f_{i,J}}$$

La nueva longitud promedio del documento en esta pseudo-colección es:

$$\widetilde{dl_{ave}} = \frac{\sum_{j=1}^{N} \widetilde{dl_J}}{N}$$

Esto hace que los pesos de los términos no sean lineales con las pseudo-frecuencias, preservando la dependencia estadística de los términos y sus valores IDF en la colección original.

Finalmente, el escenario no estructurado se restaura estableciendo $v = 1$ para todos los campos, sin comprometer la dependencia estadística. Ésta es esencialmente la fortaleza del modelo BM25F.

3.2 Extracción de respuestas

El segundo componente principal del sistema desarrollado es el módulo de extracción de las respuestas (QA). Dada una pregunta en lenguaje natural y un párrafo, este sistema devuelve la respuesta a la pregunta en el párrafo o "No sé" de otra manera.

En primer lugar, se define la función principal que, dada una pregunta, un *dataframe* con los párrafos relevantes, el número máximo de respuestas para extraer y la longitud máxima de la respuesta, extrae respuestas específicas de todos los párrafos relevantes. Esta función devuelve el *dataframe* con párrafos relevantes añadiéndole información. Las mejores respuestas se agregan para cada párrafo, especificando la respuesta en sí (texto), la puntuación y el índice de inicio y finalización que define la posición de la respuesta en el párrafo.

El objetivo del proyecto era adaptar un sistema existente sobre QA para experimentar con distintas configuraciones y analizar cuál obtenía mejores resultados para el problema dado, dentro del ámbito específico de la enfermedad COVID-19, dentro del dominio médico. Para ello, el módulo de extracción de respuestas se implementó con tres configuraciones distintas, basándose en tres modelos pre-entrenados (ver 2.2) que se detallan en 4.

3.3 Almacenamiento y muestra de resultados

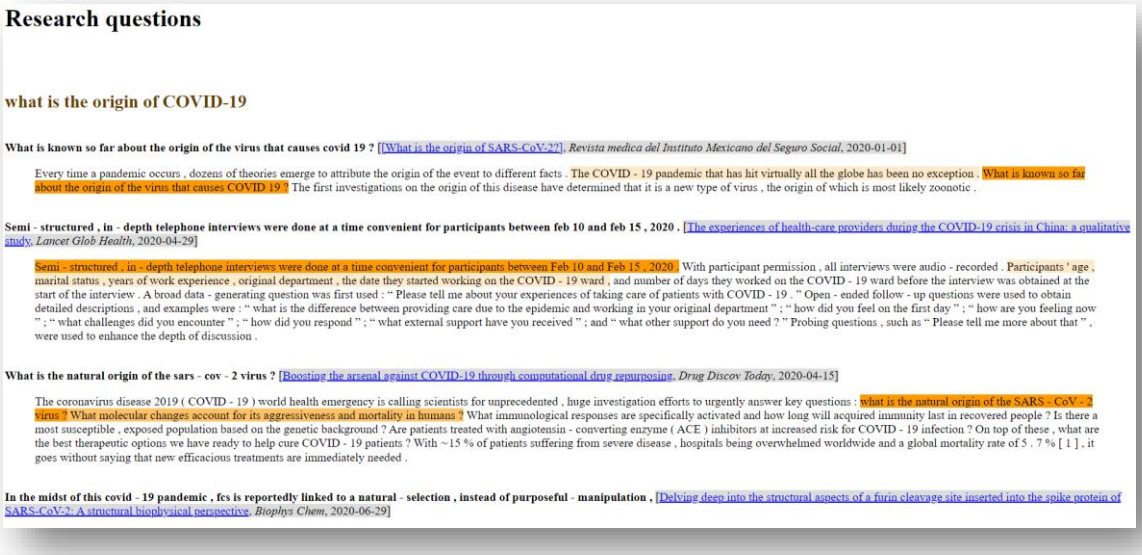
Los resultados obtenidos se almacenan en un fichero de salida con extensión *.txt* y cumplen el formato exigido para realizar la validación y evaluación propuestos en el reto (EPIQ-QA, 2020). La explicación de este fichero de salida y la forma de evaluar los resultados se describe ampliamente en 5.

Además, se genera un fichero HMTL para que puedan visualizarse de forma más intuitiva y legible los resultados. Este fichero muestra las preguntas con la selección de respuestas de entre los párrafos más significativos. Para esta visualización, se ha establecido en 20 el número máximo de párrafos que devuelve el sistema de IR, pero se descartan los párrafos en los que el sistema de QA devuelve "No sé". Además, no se muestra ningún resultado para las preguntas

que reciben más del 90% de respuestas "No sé". Para el resto de las preguntas, se muestra la mejor cadena de respuestas para cada uno de los mejores cinco párrafos, es decir, cinco respuestas específicas por pregunta.

Adicionalmente, junto a cada respuesta, se muestra alguna información extra: el título del trabajo de donde se extrajo la respuesta (con un enlace para acceder a la versión online en la web), la revista y la fecha de publicación. Además, debajo de la respuesta se muestra el párrafo del que se extrajo la respuesta. En este párrafo se resaltan las 3 mejores respuestas, usando diferente claridad de color (cuanto más oscura, mejor es la respuesta).

Para visualizar los resultados, puede abrirse en un navegador web el fichero *.html* generado (que se encuentra en la carpeta de salida del sistema con el nombre 'html').



Research questions

what is the origin of COVID-19

What is known so far about the origin of the virus that causes covid 19 ? [[What is the origin of SARS-CoV-2?], *Revista medica del Instituto Mexicano del Seguro Social*, 2020-01-01]

Every time a pandemic occurs , dozens of theories emerge to attribute the origin of the event to different facts . The COVID - 19 pandemic that has hit virtually all the globe has been no exception . What is known so far about the origin of the virus that causes COVID 19 ? The first investigations on the origin of this disease have determined that it is a new type of virus , the origin of which is most likely zoonotic .

Semi - structured , in - depth telephone interviews were done at a time convenient for participants between feb 10 and feb 15 , 2020 . [The experiences of health-care providers during the COVID-19 crisis in China: a qualitative study, *Lancet Glob Health*, 2020-04-29]

Semi - structured , in - depth telephone interviews were done at a time convenient for participants between Feb 10 and Feb 15 , 2020 . With participant permission , all interviews were audio - recorded . Participants ' age , marital status , years of work experience , original department , the date they started working on the COVID - 19 ward , and number of days they worked on the COVID - 19 ward before the interview was obtained at the start of the interview . A broad data - generating question was first used : " Please tell me about your experiences of taking care of patients with COVID - 19 . " Open - ended follow - up questions were used to obtain detailed descriptions , and examples were : " what is the difference between providing care due to the epidemic and working in your original department " ; " how did you feel on the first day " ; " how are you feeling now " ; " what challenges did you encounter " ; " how did you respond " ; " what external support have you received " ; and " what other support do you need ? " Probing questions , such as " Please tell me more about that " , were used to enhance the depth of discussion .

What is the natural origin of the sars - cov - 2 virus ? [Boosting the arsenal against COVID-19 through computational drug repurposing, *Drug Discov Today*, 2020-04-15]

The coronavirus disease 2019 (COVID - 19) world health emergency is calling scientists for unprecedented , huge investigation efforts to urgently answer key questions : what is the natural origin of the SARS - CoV - 2 virus ? What molecular changes account for its aggressiveness and mortality in humans ? What immunological responses are specifically activated and how long will acquired immunity last in recovered people ? Is there a most susceptible , exposed population based on the genetic background ? Are patients treated with angiotensin - converting enzyme (ACE) inhibitors at increased risk for COVID - 19 infection ? On top of these , what are the best therapeutic options we have ready to help cure COVID - 19 patients ? With ~15 % of patients suffering from severe disease , hospitals being overwhelmed worldwide and a global mortality rate of 5 . 7 % [1] , it goes without saying that new efficacious treatments are immediately needed .

In the midst of this covid - 19 pandemic , fcs is reportedly linked to a natural - selection , instead of purposeful - manipulation , [Delving deep into the structural aspects of a furin cleavage site inserted into the spike protein of SARS-CoV-2: A structural biophysical perspective, *Biophys Chem*, 2020-06-29]

Figura 8. Ejemplo de salida en formato .html del sistema desarrollado

4 Configuraciones de la experimentación para la extracción de respuestas

El principal objetivo del proyecto era realizar una comparativa experimentando con distintos modelos pre-entrenados para la tarea de extracción de las respuestas. A continuación, se detallan los modelos que fueron seleccionados para realizar dicha experimentación.

4.1 Configuración 1: SciBERT model

Con el uso intensivo de BERT (ver 2.4.1.1) en distintos ámbitos del NLP, se detectaron algunos problemas o debilidades. Aunque el modelo original tenía la idea correcta, estaba mal entrenado y, por lo tanto, puede considerarse que tiene potencial sin explotar.

Esta primera configuración propuesta para el proyecto utiliza el modelo de representación de lenguaje SciBERT, que es un modelo de lenguaje preentrenado basado en BERT, pero entrenado en un gran corpus de texto científico, incluyendo texto del dominio biomédico. BERT ha mostrado resultados exitosos en muchas tareas de NLP, pero SciBERT está más adaptado al dominio de este proyecto.

SciBERT está entrenado con una muestra al azar de 1.14M de artículos provenientes de Semantic Scholar. Este corpus consta de un 18% de artículos del dominio informático y de un 82% del dominio biomédico. Se utiliza el texto completo de los artículos y no solo los resúmenes. La media del tamaño de los artículos es de 154 oraciones (2.789 tokens) que da como resultado un corpus de un tamaño similar al usado en BERT. Las oraciones están divididas utilizando el algoritmo ScispaCy que está optimizado para textos científicos.

En general, este modelo tiene las mismas opciones de arquitectura, optimización e hiperparámetros que se utilizaron con BERT para la clasificación de texto (es decir, CLS y REL), se alimenta el vector BERT final para el token [CLS] en una capa de clasificación lineal. Para el etiquetado de secuencia (es decir, NER y PICO), se aplica el vector BERT final para cada token en una capa de clasificación lineal con salida softmax.

Para realizar el *fine tuning* del modelo SciBERT para QA para el desafío en Kaggle en el que se basa la primera configuración analizada para este proyecto, se probaron varios datos de entrenamiento:

1. El conjunto de datos SQuAD2.0 (ver 2.4.1.2.1).
2. La versión preprocesada de BioASQ 6b / 7b (ver 2.4.1.2.3).
3. El conjunto de datos QuAC (ver 2.4.1.2.2).

Al inspeccionar las respuestas producidas, se observó que SQuAD2.0 producía una buena calidad para las preguntas que buscaban respuestas cortas, y una combinación de SQuAD2.0 y QuAC producía una mejor calidad para las preguntas que requerían respuestas más largas; el ajuste fino de BioASQ no parecía ser efectivo. Por lo tanto, se seleccionó el modelo SciBERT ajustado primero en SQuAD 2.0 y luego en QuAC.

4.2 Configuración 2: ELECTRA model

La tercera configuración o variación propuesta para el proyecto utiliza el modelo ELECTRA y Keras para el entrenamiento con redes neuronales.

ELECTRA es un nuevo enfoque de preentrenamiento que entrena dos modelos de transformadores: el generador y el discriminador. La función del generador es reemplazar tokens en una secuencia y, por lo tanto, se entrena como un modelo de lenguaje enmascarado. El discriminador, que es el modelo que interesa, intenta identificar qué tokens fueron reemplazados por el generador en la secuencia.

ELECTRA es un método para el aprendizaje de representaciones lingüísticas auto-supervisado. Puede usarse para pre-entrenar redes de transformadores usando relativamente poca computación. Los modelos ELECTRA están entrenados para distinguir los tokens de entrada "reales" frente a los tokens de entrada "falsos" generados por otra red neuronal.

Los métodos de preentrenamiento del modelado de lenguaje enmascarado (MLM) como BERT corrompen la entrada reemplazando algunos tokens con [MASK] y luego entrenan un modelo para reconstruir los tokens originales. Si bien producen buenos resultados cuando se transfieren a tareas posteriores de NLP, generalmente requieren grandes cantidades de cálculo para ser efectivas.

Como alternativa, este método propone una tarea de preentrenamiento más eficiente para la muestra llamado detección de token reemplazado. En lugar de enmascarar la entrada, este enfoque la corrompe reemplazando algunos tokens con alternativas plausibles muestreadas de una pequeña red de generadores. Luego, en lugar de entrenar un modelo que predice las identidades originales de los tokens corruptos, se entrena un modelo discriminativo que predice si cada token en la entrada dañada fue reemplazado por una muestra del generador o no. Experimentos exhaustivos demuestran que esta nueva tarea de preentrenamiento es más eficiente que MLM porque la tarea se define sobre todos los tokens de entrada en lugar de solo el pequeño subconjunto que estaba enmascarado. Como resultado, las representaciones contextuales aprendidas por este enfoque superan sustancialmente a las aprendidas por BERT dado el mismo tamaño de modelo, datos y cálculo.

El modelo BERT original se basa, como ya se ha visto en secciones anteriores, en dos tareas previas al entrenamiento: el modelado de lenguaje enmascarado (MLM) y la predicción de la siguiente oración. En la predicción de la siguiente oración, el modelo tiene la tarea de predecir si dos secuencias de texto se suceden naturalmente o no. Se dijo que esta tarea ayudaba con ciertas tareas posteriores, como la respuesta a preguntas y la inferencia del lenguaje natural.

En el modelado de lenguaje enmascarado, BERT fue capaz de destruir puntos de referencia previamente establecidos en muchas de las tareas posteriores de NLP. Sin embargo, con ELECTRA señalan que los enfoques de MLM solo aprenden de los tokens enmascarados (típicamente el 15%) de cualquier ejemplo dado. Esto conduce a un aumento sustancial de los recursos computacionales necesarios para entrenar un modelo de lenguaje con MLM. Otra desventaja del MLM es que las fichas de máscara solo aparecen en la etapa de preentrenamiento y nunca durante el ajuste fino o el uso posterior. Esta discrepancia también contribuye a una ligera pérdida de rendimiento en modelos entrenados con MLM.

ELECTRA aporta un nuevo enfoque de preentrenamiento que tiene como objetivo igualar o superar el rendimiento posterior de un modelo pre-entrenado de MLM mientras utiliza significativamente menos recursos informáticos para la etapa de preentrenamiento. La tarea de preentrenamiento en ELECTRA se basa en la detección de tokens reemplazados en la secuencia de entrada. Esta configuración requiere dos modelos de transformador, un generador y un discriminador.

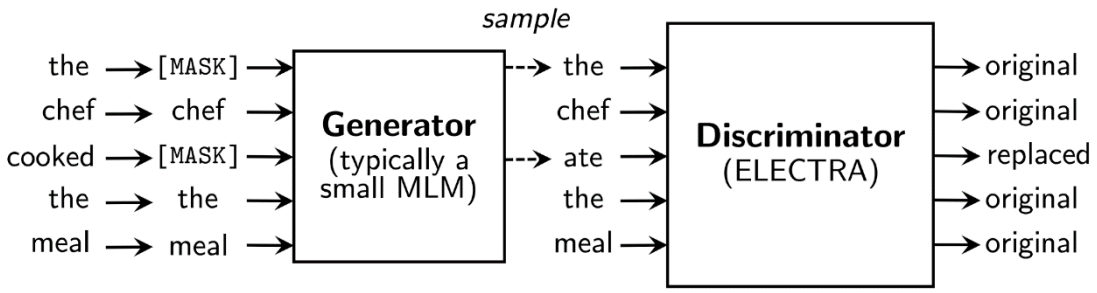


Figura 9. Enfoque de pre-entrenamiento en el modelo ELECTRA

Los principales pasos que tiene este enfoque son:

1. Para una secuencia de entrada determinada, reemplaza aleatoriamente algunos tokens con un token [MASK].
2. El generador predice los tokens originales para todos los tokens enmascarados.
3. La secuencia de entrada al discriminador se construye reemplazando los tokens [MASK] con las predicciones del generador.
4. Para cada token en la secuencia, el discriminador predice si es un original o si ha sido reemplazado por el generador.

El modelo generador está entrenado para predecir los tokens originales para tokens enmascarados, mientras que el modelo discriminador está entrenado para predecir qué tokens han sido reemplazados dada una secuencia corrupta. Esto significa que la pérdida del discriminador se puede calcular sobre todos los tokens de entrada mientras realiza la predicción en cada token. Con MLM, la pérdida del modelo solo se calcula sobre los tokens enmascarados. Se muestra que esto es una diferencia clave entre los dos enfoques y la razón principal detrás de la mayor eficiencia de ELECTRA.

Esta configuración es similar a la configuración de entrenamiento de una GAN (*Generative Adversarial Network*), excepto que el generador no está entrenado para intentar engañar al discriminador (por lo que no es un adversario en sí mismo). Además, si el generador predice correctamente el token original de un token enmascarado, ese token se considera un token original (ya que el token no se ha corrompido/ cambiado).

El modelo discriminador se utiliza para tareas posteriores y el generador se desecha después del entrenamiento previo. (Kevin Clark, 2020)

Existen varias versiones disponibles del modelo ELECTRA:

Model	Layers	Hidden Size	Params	GLUE score (test set)
ELECTRA-Small	12	256	14M	77.4
ELECTRA-Base	**12**	**768**	**110M**	**82.7**
ELECTRA-Large	24	1024	335M	85.2

Tabla 2. Parámetros de los modelos ELECTRA

Para esta configuración, se ha empleado el modelo ELECTRA-Base (Xhlulu, 2020).

4.3 Configuración 3: RoBERTa model

La segunda configuración o variación propuesta para el proyecto utiliza el modelo base RoBERTa sobre SQuAD 2.0. Las principales modificaciones que incluye con respecto al modelo BERT son las siguientes:

- Se utilizan 160GB de texto frente al dataset original usado para entrenar BERT que tiene un tamaño más pequeño.
- Se realiza un entrenamiento más largo, aumentando el número de iteraciones de 100K a 300K y luego a 500K.
- Se utilizan lotes más grandes, de 8K en lugar de 256 en el modelo base BERT original.
- El vocabulario BPE a nivel de bytes es más grande con 50K unidades de subpalabras en lugar de vocabulario BPE a nivel de carácter de tamaño 30K.
- Se elimina el siguiente objetivo de predicción de secuencia del procedimiento de entrenamiento.

El conjunto de datos incluye cinco corpus en inglés de diferentes tamaños y dominios, con un total de más de 160 GB de texto sin comprimir:

- BookCorpus además de Wikipedia en inglés que son los datos originales utilizados para entrenar a BERT.
- CC-News, recopilando la parte en inglés del conjunto de datos de CommonCrawl News. Los datos contienen 63 millones de artículos de noticias en inglés se rastrearon entre septiembre de 2016 y febrero de 2019.
- OpenWebText es texto es contenido web extraído de las URL compartidas en Reddit[2] con al menos tres votos a favor.
- Stories es un conjunto de datos contiene un subconjunto de datos de CommonCrawl filtrados para que coincidan con el estilo de historia de Esquemas de Winograd.

Se entrena previamente solo con los textos en bruto, sin etiquetarlos de forma manual de ninguna manera (por lo que puede usar muchos datos disponibles públicamente) con un proceso automático para generar entradas y etiquetas a partir de esos textos. (Yinhan Liu, 2019)

También es entrenado previamente con el objetivo de modelado de realizar lenguaje enmascarado (MLM). Al tomar una oración, el modelo enmascara al azar el 15% de las palabras en la entrada, luego ejecuta toda la oración enmascarada a través del modelo y tiene que predecir las palabras enmascaradas. Esto es diferente de las redes neuronales recurrentes tradicionales (RNN) que generalmente ven las palabras una tras otra, o de modelos autorregresivos como GPT (*Generative Pre-trained Transformer*) que enmascaran internamente los tokens futuros. Permite que el modelo aprenda una representación bidireccional de la oración.

De esta manera, el modelo aprende una representación interna del idioma inglés que luego se puede usar para extraer características útiles para tareas posteriores: si tiene un conjunto de datos de oraciones etiquetadas, por ejemplo, puede entrenar un clasificador estándar usando las características producidas por BERT modelo como entradas.

[2] Reddit es un sitio web de marcadores sociales y agregador de noticias donde los usuarios pueden añadir textos, imágenes, videos o enlaces. Los usuarios pueden votar a favor o en contra del contenido, haciendo que aparezcan en las publicaciones destacadas.

En el modelo de lenguaje enmascarado (MLM) se selecciona una muestra aleatoria de los tokens en la secuencia de entrada y se reemplaza con el token especial [MASK]. El objetivo de MLM es una pérdida de entropía cruzada al predecir los tokens enmascarados. BERT selecciona uniformemente el 15% de los tokens de entrada para un posible reemplazo. De los tokens seleccionados, el 80% se reemplaza con [MASK], el 10% se deja sin cambios y el 10% se reemplaza por un token de vocabulario seleccionado al azar.

En la implementación original, el enmascaramiento y reemplazo aleatorio se realiza una vez al principio y se guarda durante la duración del entrenamiento, aunque en la práctica, los datos se duplican para que la máscara no siempre es el mismo para todas las frases de entrenamiento.

En la predicción de la siguiente oración (NSP), los ejemplos positivos se crean tomando oraciones consecutivas del corpus de texto. Los ejemplos negativos se crean emparejando segmentos de diferentes documentos. Ejemplos positivos y negativos se muestrean con la misma probabilidad.

Los parámetros de entrenamiento del modelo RoBERTa empleado para esta configuración son los siguientes (Deepset, 2020):

```
batch_size = 24
n_epochs = 3
base_LM_model = "deepset/roberta-base-squad2"
max_seq_len = 384
learning_rate = 3e-5
lr_schedule = LinearWarmup
warmup_proportion = 0.1
doc_stride = 128
xval_folds = 5
dev_split = 0
no_ans_boost = -100
```

5 Marco de evaluación

En este capítulo se describe la metodología utilizada para evaluar el sistema propuesto.

5.1 Dataset

A continuación se detallan los datos de entrada que emplea el sistema para obtener los resultados.

5.1.1 Documentos

Los documentos proporcionados en el desafío, tanto los artículos de investigación o documentos científicos, orientados a personas de ámbito médico, como los artículos más generalistas, orientados a un consumidor, se encuentran en formato *.json* y algunos disponen, además, del texto completo en *.pdf*.

Todos los documentos siguen el mismo esquema:

```
{
	"document_id": <str>,		# sha1 de la url del documento
	"metadata": {
		"title": <str>,			# Título del documento
		"url": <str>,			# URL de la web (si es una web)
		"authors": [...]		# Autores
	},
	"contexts": [			# Lista de contexto(s) del documento
		{
			"section": <str>,	# Nombre de sección conteniendo el contexto
			"text": <str>,		# Texto completo de la sección
			"context_id": <str>,	# Identificador global único
			"sentences": [ # Lista de oraciones del contexto
				{
					"start": 0,	# Carácter inicial de la oración
					"end": 27,	# Carácter final de la oración
					"sentence_id": <str>, # Identificador global único
				},
				{	# Segunda oración del contexto
					"start": 28,
					"end": 56,
					"sentence_id": <str>,
				},
				{...}, # Tercera oración del contexto
				...
			]
		},
		{...}, # Segundo contexto del documento
		...
	]
}
```

5.1.1.1 Artículos científicos (para responder a preguntas de expertos)

La colección proporcionada en este reto para dar respuestas a usuarios de la comunidad científica es una adaptación de la colección de artículos biomédicos publicados para el COVID-19 Open Research Dataset Challenge (COVID-19 Open Research Dataset Challengue, 2020). Se utiliza una instantánea de CORD-19 del 22 de octubre de 2020.

El conjunto de datos fue creado por el Instituto Allen de IA en asociación con la Iniciativa Chan Zuckerberg, el Centro de Seguridad y Tecnología Emergente de la Universidad de Georgetown, Microsoft Research y la Biblioteca Nacional de Medicina - National Institutes of Health, en coordinación con la Oficina de Política Científica y Tecnológica de la Casa Blanca. La colección

CORD-19 incluye un subconjunto de artículos en PubMed Central (PMC), así como preimpresiones de bioRxiv. Los contextos de esta colección corresponden a los párrafos identificados automáticamente en los resúmenes de los artículos o en los textos principales.

En respuesta a la pandemia, estas instituciones prepararon el conjunto de datos con más de 400000 artículos académicos, incluidos más de 150000 con texto completo, sobre COVID-19, SARS-CoV-2 y coronavirus relacionados. Este conjunto de datos de acceso gratuito y se proporciona a la comunidad de investigación global para aplicar los avances recientes en el procesamiento del lenguaje natural y otras técnicas de inteligencia artificial para generar nuevos conocimientos en apoyo de la lucha en curso contra esta enfermedad infecciosa. (Ver 8.2.1.1).

5.1.1.2 Artículos generalistas (para responder a preguntas de consumidor)

La colección proporcionada en este reto para dar respuestas a usuarios del público general es un subconjunto de los artículos utilizados por el servicio de respuesta a preguntas de información de salud del consumidor (CHIQA) de la Biblioteca Nacional de Medicina de EE. UU. (NLM). Esta colección incluye artículos autorizados de: los Centros para el Control y la Prevención de Enfermedades (CDC); el Centro de Información sobre Enfermedades Raras y Genéticas (GARD); el Genetics Home Reference (GHR); Medline Plus; el Instituto Nacional de Alergias y Enfermedades Infecciosas (NIAID); la Organización Mundial de la Salud (OMS). Los contextos de esta colección corresponden a párrafos o secciones según lo indicado por el marcado HTML del documento.

También incluye 265 hilos de Reddit etiquetados con COVID-19, Medicina, Biología o el Cuerpo Humano, y filtrados por contenido de COVID-19.

Por último, se incluye un subconjunto del rastreo de CommonCrawl News desde el 1 de enero hasta el 30 de abril de 2020, como se usa en TREC Health Misinformation Track. Los documentos de este subconjunto se filtraron por dominio utilizando SALSA, PageRank y HITS y se filtraron aún más para el contenido de COVID-19. (Ver 8.2.1.2).

5.1.2 Preguntas

El conjunto de preguntas a responder en la versión final de la evaluación (*primary*), contiene dos conjuntos de 30 preguntas: un conjunto para preguntas de nivel de experto y otro para preguntas de nivel de consumidor.

Las preguntas consideradas de consumidor más generalista, en su mayoría fueron originadas en las interacciones de los consumidores con MedlinePlus.

Las preguntas científicas se desarrollaron basándose en discusiones grupales del grupo de interés especial de los Institutos Nacionales de Salud (NIH) sobre COVID-19, preguntas hechas por médicos de la Universidad de Ciencias de la Salud de Oregon y respuestas a una convocatoria pública de preguntas.

Las preguntas de nivel consumidor están etiquetadas como “CQ___”, siendo el rango: [CQ101 – CQ130]. Las preguntas de nivel científico o exerto están etiquetadas como “EQ___”, siendo el rango: [EQ101 – EQ130].

Todas las preguntas se encuentran en formato *.json* y siguen el mismo esquema:

```
{
       "question_id": <str>,        # Identificador único global para la pregunta
       "question": <str>,           # Texto de la pregunta
       "query"; <str>,              # Tokens preprocesados de la pregunta
```

```
        "background": <str>,          # Contexto narrativo para la pregunta
}
```

El listado completo de preguntas se encuentra disponible en 8.2.

5.1.3 Listados de nuggets por pregunta

Siguiendo la tradición TAC, al conjunto de hechos atómicos que responden a una determinada pregunta se les denomina nuggets (pepitas). En el caso del proyecto, los evaluadores del desafío han realizado este etiquetado que permite evaluar las respuestas obtenidas en base a las pepitas que contengan.

Por ejemplo, para la pregunta CQ101:

```
“If the common cold is a type of coronavirus and we're unable to find a cure, why does the medical community have confidence we will find a vaccine for COVID-19?"
```

se dispone de los siguientes nuggets:

```
"Common Cold", "Companies Efforts", "Government Agency Efforts", "Time to create a vaccine", "Human Trials", "Severity of Covid-19", "Pneumonia & Influenza", "University Efforts", "Vaccine Distribution", "Politicans", "Genetics" y "Similar Viruses".
```

5.1.4 Juicios de relevancia de los pasajes y documentos

En el desafío se proporciona la lista de respuestas generadas por humanos y las anotaciones de respuestas a nivel de oración, para las 30 preguntas evaluadas en la tarea de respuesta a preguntas del ámbito científico y las 30 preguntas evaluadas en la tarea de respuestas a preguntas de público generalista durante la evaluación preliminar.

Esta lista se utiliza como entrada en el método de evaluación propuesto en EPIC-QA que se describe con detalle en 5.2.

Los documentos se encuentran etiquetados a nivel de oración con los nuggets que contienen. Por ejemplo, el nugget identificado como CQ101-N00:

```
“Common Cold”
```

aparece en la lista de nuggets de la oración:

```
“There are a few viruses that cause common cold and inky a small proportion of those are caused by coronvirus”
```

que se encuentra en un documento perteneciente a la batería de documentos generalistas.

El fichero conjunto de juicios de relevancia y nuggets, se encuentra en formato *.json* y sigue el esquema:

```
[
    {
        "question_id": <str>  # Identificador único global para la pregunta
        "nuggets": [          # Listado de nuggets para la pregunta
            {
                "nugget_id": <str>,  # Identificador único global para el nugget
                "nugget": <str>      # Cadena de tokens del nugget
            }
        ],
        "annotations": [      # Listado de anotaciones para la pregunta
             {
                 "sentence_id": <str>,       # Identificador global único de la oración
                  "nugget_ids": [            # Listado de identificadores de los nuggets
```

```
                        <str>
                    ]
                }
            ]
        }
]
```

Una sección del documento se encuentra disponible en 8.2.3.

5.2 Metodología de evaluación

Para evaluar cada una de las tareas, se ha utilizado el método propuesto por el desafío. Para ello, se han generado listas clasificadas de respuestas para cada pregunta.

Los juicios de respuesta han sido realizados por bibliotecarios de la Biblioteca Nacional de Medicina de EE. UU. (NLM). Las respuestas se evalúan dentro del contexto en el que se extrajeron (donde el contexto se refiere a los párrafos o secciones descritos anteriormente). Dentro de un ciclo de evaluación, la evaluación de respuestas se realiza en dos rondas:

1. Generación de claves de respuesta:

En la primera ronda de evaluación, los evaluadores tienen acceso a las respuestas producidas por los participantes, así como a un motor de búsqueda *ad-hoc* sobre la colección de documentos. El objetivo de esta ronda es que los evaluadores exploren las respuestas de los equipos, así como la colección de documentos para determinar un conjunto nuggets que respondan a la pregunta. Por ejemplo, para la pregunta:

```
“What is the origin of COVID-19?”
```

se podrían generar las siguientes pepitas:

```
“Malayan pangolins”, “bats”, “Guangdon province”, etc.
```

El papel principal de esta ronda es crear una clave de respuesta para la pregunta compuesta por pepitas identificadas en las respuestas de los participantes o en la búsqueda *ad-hoc* de la colección por parte del evaluador. El motor de búsqueda se proporciona para ayudar a los evaluadores a explorar el tema e identificar las pepitas que pueden no haber sido devueltas por ningún equipo. No se espera que los evaluadores identifiquen de manera exhaustiva todas las pepitas posibles que los equipos no devolvieron; más bien, la intención es que identifiquen detalles importantes (a discreción del evaluador) que, en su opinión, deberían incluirse en la clave de respuestas en función de su comprensión del tema.

2. Anotación de pasaje:

En la segunda ronda de evaluación, la clave de respuesta (lista de pepitas) es fija. A los evaluadores se les da el mismo conjunto de pasajes de respuesta y contextos que se usaron en la primera ronda. Esta vez, se les pide que anoten oraciones en cada contexto indicando qué pepitas (si las hay) se tratan en cada oración. Por ejemplo, la oración:

```
 “Malayan pangolins (Manis javanica) illegally imported into Guangdong province contain coronaviruses similar to SARS-CoV-2”
```

podría anotarse que contiene dos pepitas:

```
“Malayan pangolins” y “Guangdon province”
```

Se asume que las oraciones sin anotaciones no contienen pepitas.

Para poder realizar esta evaluación utilizando el *script* disponible a tal efecto, el formato exigido para cada línea de la evaluación es el siguiente (ver todos los resultados obtenidos en 8.4):

```
QUESTION_ID Q0 START_SENTENCE_ID: END_SENTENCE_ID RANK SCORE RUN_NAME
```

Donde:

- QUESTION_ID es el ID de la pregunta del archivo *.json* de preguntas
- Q0 es una constante requerida para la compatibilidad
- START_SENTENCE_ID y END_SENTENCE_ID son ID (inclusive) que indican una serie de oraciones contiguas del mismo contexto de documento
- RANK es el rango (1-1000) de la respuesta en la lista de respuestas recuperadas para esa pregunta
- SCORE es una puntuación numérica de punto de inflexión o peso asignado a esa respuesta
- RUN_NAME es el nombre de la asociación de ejecución con este archivo de envío (de cara al envío de las respuestas para el desafío propuesto)

5.3 Métricas de evaluación

Tal y como ha explicado en el apartado anterior, la métrica de evaluación favorece la exploración del panorama de respuestas (o pepitas) en la colección, alentando a los sistemas a proporcionar una lista diversa de respuestas. Específicamente, en el desafío se clasifica a los participantes usando la métrica NDNS (*Normalized Discount Novelty Score*) que es una versión modificada de Ganancia Acumulada Descontada Normalizada (NDCG) donde:

1. Una respuesta se considera relevante si y solo si describe una pepita que no se ha incluido en ninguna de las respuestas en los rangos anteriores en la lista.
2. Las respuestas se penalizan en función de su longitud (en oraciones).

5.3.1 NDCG (Normalized Discounted Cumulative Gain)

Para comprender la métrica NDCG, primero es necesario conocer las métricas CG (*Cumulative Gain*) y DCG (*Discounted Cumulative Gain*), así como comprender las dos suposiciones que se hacen al usar DCG y sus medidas relacionadas:

1. Los documentos muy relevantes son más útiles cuando aparecen antes en la lista de resultados del motor de búsqueda.
2. Los documentos muy relevantes son más útiles que los documentos marginalmente relevantes, que son más útiles que los documentos no relevantes.

Si cada recomendación tiene una puntuación de relevancia calificada asociada, CG es la suma de los valores de relevancia calificada de todos los resultados en una lista de resultados de búsqueda:

$$CG_p = \sum_{i=1}^{p} rel_i$$

La ganancia acumulada en una posición de rango particular p, donde rel_i es la relevancia calificada del resultado en la posición i. Para demostrar, se podría crear una variable A con la puntuación de relevancia calificada de una respuesta a una consulta de búsqueda, por lo que cada calificación de relevancia calificada se asocia con un documento.

El problema con CG es que no toma en consideración el rango del conjunto de resultados al determinar la utilidad de un conjunto de resultados. En otras palabras, si fuera necesario reordenar las puntuaciones de relevancia graduadas devueltas en el conjunto A, no se obtendría una mejor idea de la utilidad del conjunto de resultados, ya que el GC no cambiaría.

DCG penaliza los documentos muy relevantes que aparecen más abajo en la búsqueda al reducir el valor de relevancia calificada logarítmicamente proporcional a la posición del resultado:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Surge un problema con DCG cuando se quiere comparar el rendimiento de los motores de búsqueda de una consulta a la siguiente porque la lista de resultados de búsqueda puede variar en longitud dependiendo de la consulta que se haya proporcionado. Por lo tanto, al normalizar la ganancia acumulada en cada posición para un valor elegido de p en las consultas, se llega a NDCG. Se realiza este proceso clasificando todos los documentos relevantes en el corpus por su relevancia relativa produciendo el máximo DCG posible a través de la posición p (también conocido como *Ideal Discounted Cumulative Gain*):

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

donde:

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{log_2(i+1)}$$

Las proporciones siempre estarán en el rango de [0, 1], siendo 1 una puntuación perfecta, lo que significa que el DCG es el mismo que el IDCG. Por lo tanto, los valores de NDCG se pueden promediar para todas las consultas para obtener una medida del rendimiento promedio de un algoritmo de clasificación de sistemas de recomendación.

5.3.2 NDNS (*Normalized Discount Novelty Score*)

La métrica NDCG tiene ciertas limitaciones, ya que no penaliza por documentos incorrectos en los resultados ni penaliza la falta de documentos en los resultados. Además, puede no ser adecuado para medir el rendimiento de consultas que a menudo pueden tener varios resultados igualmente buenos. Por ello en este caso se utiliza la métrica NDNS como variante del NDCG (Goodwin & Demner-Fushman, 2020).

Es importante destacar que, si bien la ganancia acumulada en NDCG se puede calcular para un documento independientemente de otros documentos recuperados, el NS (*Novelty Score*) mide la información en una respuesta que no se ha visto previamente en la lista clasificada. Formalmente, se define NS como:

$$NS(a) = \frac{n_a(n_a + 1)}{n_a + f_a}$$

donde n_a es el número de pepitas nuevas en la respuesta a y f_a es el factor de oración. Se considera una pepita novel si no ha estado presente en una respuesta recuperada anteriormente en la lista clasificada. Reportamos tres variantes de NDNS en el que el factor de oración, f_a, se calcula de manera diferente:

1. NDNS Exacto: las respuestas deben ser breves (es decir, deben expresar una pepita novedosa en la menor cantidad de oraciones posible) y no deben contener oraciones con solo pepitas proporcionadas en respuestas anteriores. En esta variante, el factor de oración es solo el número de oraciones en la respuesta, es decir:

$$f_a = s_a = s_0 + s_s + s_n$$

donde sa es el número de oraciones en la respuesta a, $s0$ es el número de oraciones sin pepitas, ss es el número de oraciones con pepitas vistas anteriormente, y sn es el número de oraciones con pepitas nuevas.

2. NDNS Relajado: las respuestas no se penalizan por expresar nuggets novedosos en múltiples oraciones, pero deben todavía no contiene oraciones con solo pepitas proporcionadas en respuestas anteriores, es decir:

$$f_a = s_0 + s_s + \min(s_n, 1)$$

3. NDNS Parcial: las respuestas no se penalizan por expresar nuggets novedosos en múltiples oraciones, ni por contener oraciones con solo pepitas de respuestas anteriores. En esta variante, solo penalizamos las respuestas por oraciones sin pepitas en absoluto, es decir:

$$f_a = s_0 + \min(s_n, 1)$$

Como en NDCG, se calcula el puntaje de novedad acumulativo NS de las respuestas recuperadas hasta el rango l y se descuenta la puntuación utilizando un factor de reducción logarítmico:

$$DNS(a_1, \dots, a_l) = \sum_{r=1}^{l} \frac{NS(a_r)}{log_2(r+1)}$$

Finalmente, se normaliza el NDNS de ranking $a = a1, \cdots, al$ por el NDNS del ranking óptimo o ideal de posibles respuestas que podrían haberse recuperado para esa pregunta:

$$NDNS(a) = \frac{NDNS(a)}{NDNS(\hat{a})}$$

donde $\hat{a}$ es la clasificación óptima de respuestas que se podrían haber encontrado en la colección de documentos para la pregunta dada. En la evaluación utilizada en este proyecto, se usa una búsqueda de haz con un ancho de 10 para determinar la clasificación ideal de respuestas.

6 Resultados

Los resultados que se muestran a continuación han sido obtenidos realizando la ejecución de seis escenarios distintos, basados en las siguientes variantes:

- Dos tipos de respuestas: respuestas a la comunidad científica y respuestas para el público general, descritas en 5.1.2.
- Tres configuraciones: modelos pre-entrenados distintos (SciBERT, ELECTRA y RoBERTa), implementados en el módulo de extracción de respuestas y descritos en 4.

Por lo tanto, los escenarios para dar respuesta a las preguntas de la comunidad científica son los siguientes:

	Tipo de preguntas	Configuración/ modelo
Escenario 1	Comunidad científica (expert)	Config. 1 (SciBERT)
Escenario 2	Comunidad científica (expert)	Config. 2 (ELECTRA)
Escenario 3	Comunidad científica (expert)	Config. 3 (RoBERTa)

Tabla 3. Configuración de los escenarios de experimentación para preguntas de la comunidad científica

Y los escenarios para dar respuesta a las preguntas de público general, son:

	Tipo de preguntas	Configuración/ modelo
Escenario 4	Consumidor o público generalista (consumer)	Config. 1 (SciBERT)
Escenario 5	Consumidor o público generalista (consumer)	Config. 2 (ELECTRA)
Escenario 6	Consumidor o público generalista (consumer)	Config. 3 (RoBERTa)

Tabla 4. Configuración de los escenarios de experimentación para preguntas del público generalista

Las métricas utilizadas para evaluar y comparar los resultados son las explicadas en el apartado anterior: NDNS-Exacto, NDNS-Parcial y NDNS-Relajado y han sido obtenidas realizando la evaluación proporcionada por el EPIQ-QA y descrita en 5.2.

6.1 Resultados de respuestas para comunidad científica

A continuación, se muestran los resultados de la comparativa entre los escenarios 1, 2 y 3, atendiendo a la métrica NDNS-Exact para cada una de las preguntas:

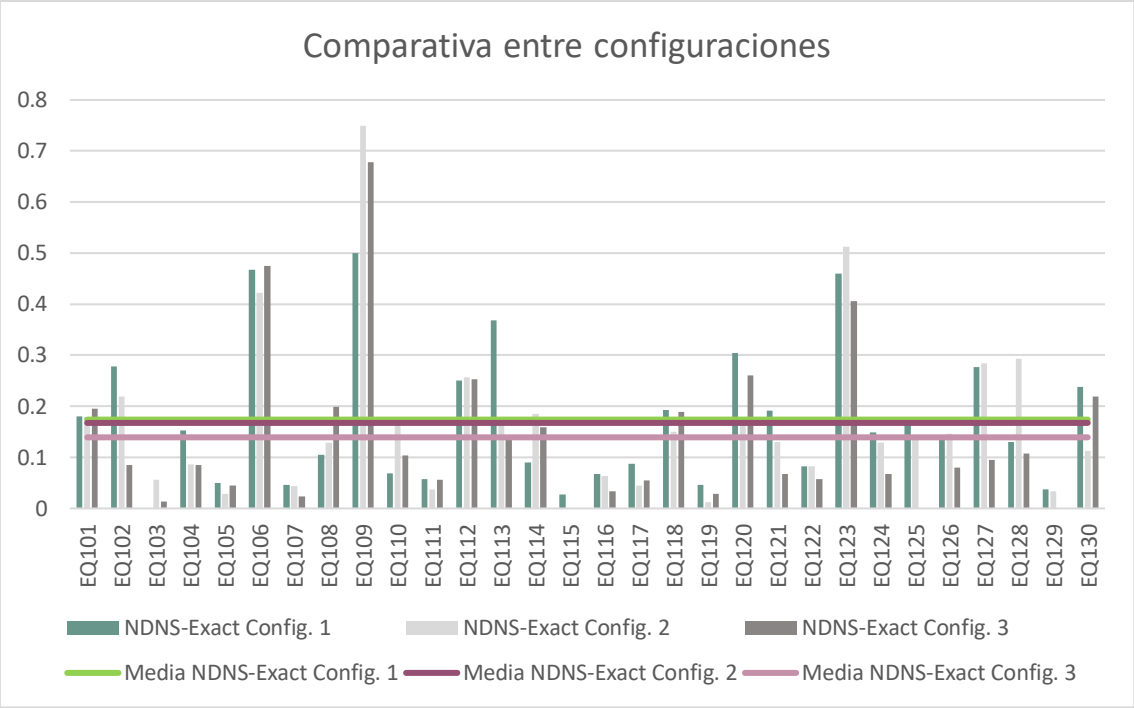


Figura 10. Gráfico de comparativa de resultados para comunidad científica

6.2 Resultados de respuestas para público general

A continuación, se muestran los resultados de la comparativa entre los escenarios 4, 5 y 6, atendiendo a la métrica NDNS-Exact para cada una de las preguntas:

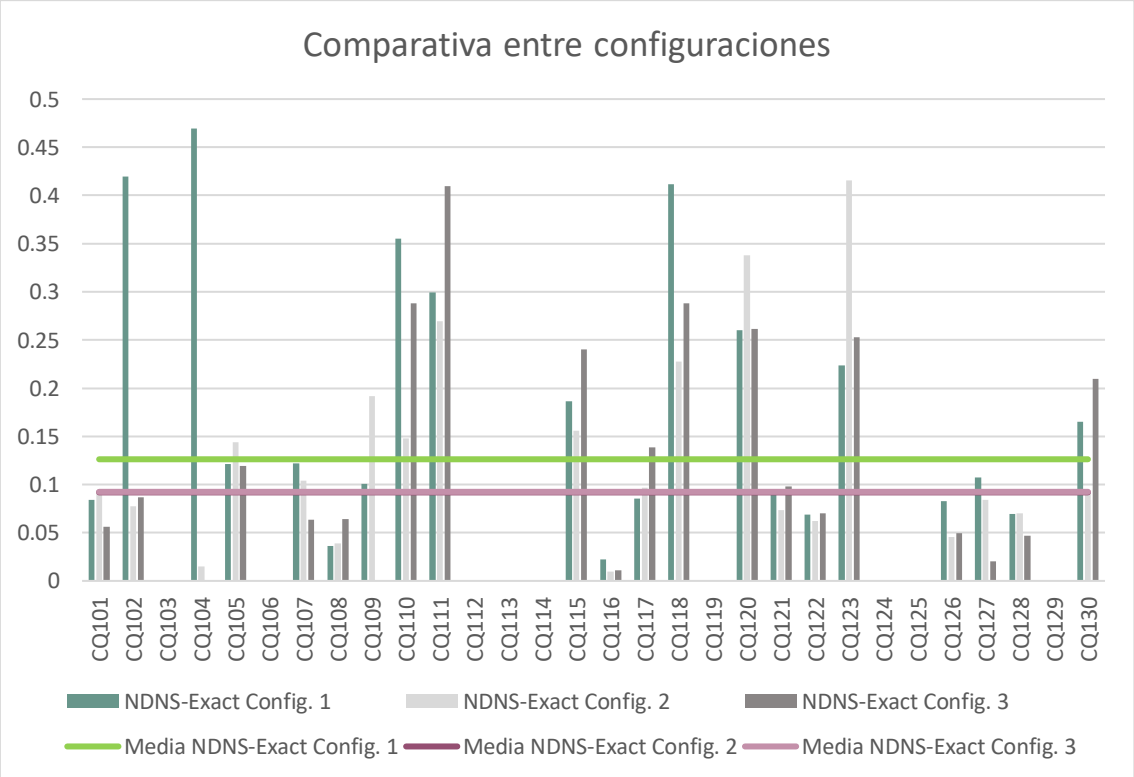


Figura 11. Gráfico de comparativa de resultados para público general

6.3 Comparativa con resultados ideales

El desafío proporciona a los participantes un listado de las puntuaciones ideales para cada una de las preguntas y es posible realizar una comparativa con los resultados obtenidos.

La comparativa obtenida para las preguntas de la comunidad científica es la siguiente:

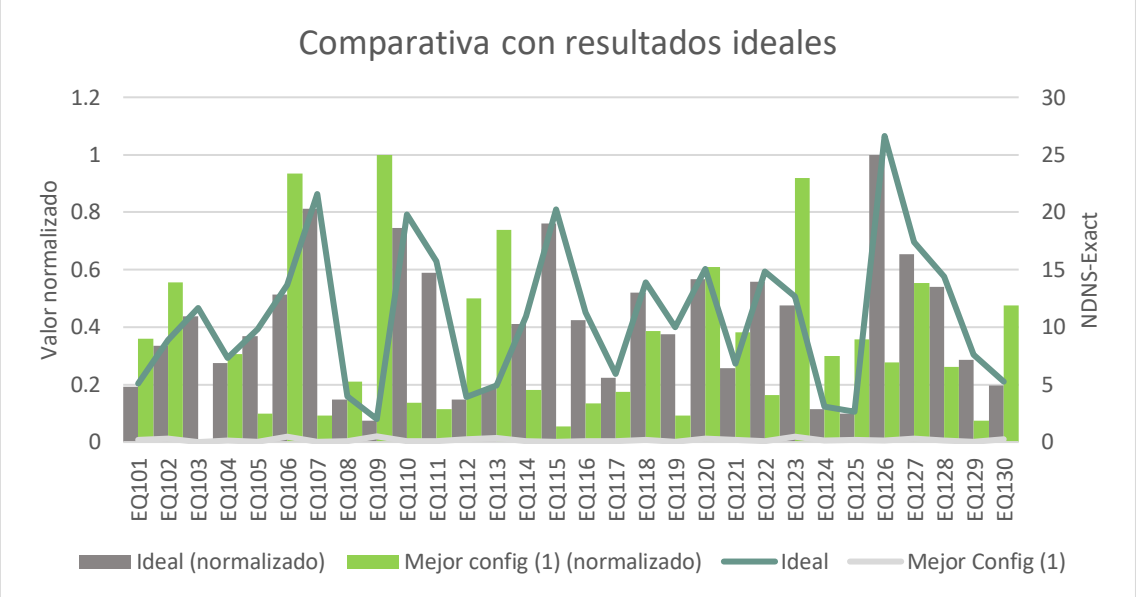


Figura 12. Gráfico de comparativa entre los resultados de la mejor configuración (escenario 1, 2 y 3) y los ideales

La comparativa obtenida para las preguntas de público general es la siguiente:

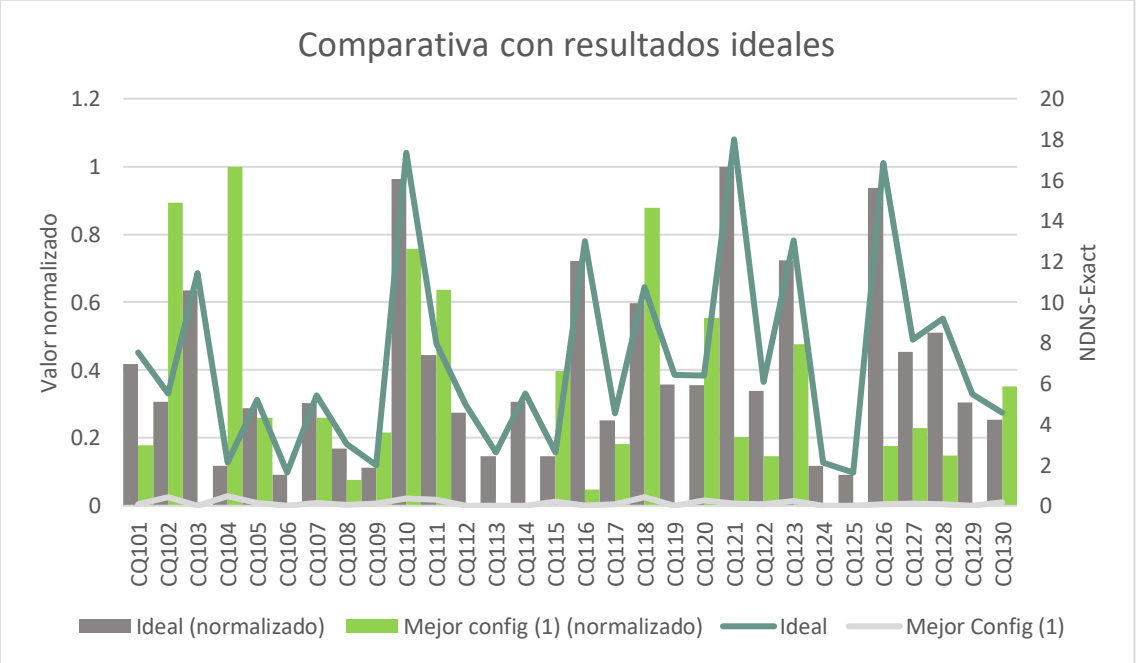


Figura 13. Gráfico de comparativa entre los resultados de la mejor configuración (escenarios 4, 5 y 6) y los ideales

6.4 Resumen de resultados

A continuación, se muestran los resultados obtenidos para los seis escenarios tomando como referencia la media para las tres métricas.

Los resultados para los tres escenarios empleados que dan respuesta a las preguntas de la comunidad científica son los siguientes:

	NDNS-Partial (avg)	NDNS-Relaxed (avg)	NDNS-Exact (avg)
Escenario 1 (expert, SciBERT)	0.187191905	0.180625196	0.173990157
Escenario 2 (expert, ELECTRA)	0.170324481	0.167377589	0.167500099
Escenario 3 (expert, RoBERTa)	*0.124889201*	*0.124521756*	*0.139217016*

Tabla 5. Resumen de medias de resultados obtenidos para los escenarios 1, 2 y 3

Y los resultados para los tres escenarios que dan respuesta a las preguntas de público general, son:

	NDNS-Partial (avg)	NDNS-Relaxed (avg)	NDNS-Exact (avg)
Escenario 4 (consumer, SciBERT)	0.126737463	0.123201337	0.126203984
Escenario 5 (consumer, ELECTRA)	0.10057127	0.098538666	0.091816653
Escenario 6 (consumer, RoBERTa)	0.08325769	0.082943689	0.092442203

Tabla 6. Resumen de medias de resultados obtenidos para los escenarios 4, 5 y 6

Los resultados medios ideales propuestos por EPIQ-QA, utilizando como referencia la media para las tres métricas, son los siguientes:

	NDNS-Partial (avg)	NDNS-Relaxed (avg)	NDNS-Exact (avg)
Comunidad científica (expert)	12.77890609	12.7313198	10.90348629
Consumidor o público generalista (consumer)	8.448441184	8.385353442	7.039330531

Tabla 7. Resultados medios propuestos por EPIQ-QA para las dos tareas

Atendiendo al rendimiento en forma de tiempo de ejecución de los escenarios, se muestra en la siguiente tabla el tiempo medio por respuesta proporcionada en cada uno de los escenarios y el tiempo total de una ejecución para la batería de preguntas dada:

	Tiempo medio por respuesta (s)	Tiempo total (m)
Escenario 1 (expert, SciBERT)	157	78
Escenario 2 (expert, ELECTRA)	160	80
Escenario 3 (expert, RoBERTa)	1028	514

Tabla 8. Tiempo de ejecución para los 1, 2 y 3

	Tiempo medio por respuesta (s)	Tiempo total (m)
Escenario 4 (consumer, SciBERT)	350	175
Escenario 5 (consumer, ELECTRA)	367	183
Escenario 6 (consumer, RoBERTa)	1042	521

Tabla 9. Tiempo de ejecución para los 4, 5 y 6

7 Conclusiones y trabajo futuro

Este capítulo recopila las diferentes conclusiones extraídas del trabajo realizado, y propone algunas líneas de trabajo futuro.

7.1 Discusión

Las tres configuraciones sobre las que se realizó la experimentación del proyecto estaban basadas en modelos pre-entrenados distintos, especialmente implementados para realizar la tarea de extracción de respuestas.

Los tres modelos partían como base del modelo BERT, realizando modificaciones a distintos niveles: variaciones en el tamaño del corpus de entrenamiento, introducción de datos de entrenamiento orientados hacia textos de ámbito más científico o médico, modificación de parámetros, etc., desglosadas en detalle en 4.

El objetivo era analizar qué configuración presentaba mejores resultados para proporcionar respuestas adecuadas sobre el virus COVID-19, tanto a personas del ámbito médico, entrenando con artículos científicos, como a personas de ámbito más generalista, basándose en informaciones publicadas en webs y otros textos. Puesto que para cada una de las dos tareas el tipo de respuesta esperado es distinto, debido al contexto y al tipo de fuente de la que se alimenta, cabría esperar que para cada una de las tareas ofrecieran mejores resultados distintas configuraciones.

Los resultados muestran que, en ambos tipos de pregunta, se comporta mejor la configuración 1 (SciBERT model), ya que tiene un valor NDNS medio superior, aunque para las preguntas del ámbito médico hay menos diferencia entre los modelos. Además, no para todas las preguntas se comporta mejor el mismo modelo.

Se observa también que en algunos casos (como la pregunta CQ104) hay una diferencia muy notable entre los resultados de las tres configuraciones, siendo las configuraciones 2 y 3 muy inferiores y cercanas a cero. Sin embargo, los valores de la configuración 1 no llegan nunca a ser tan bajos en comparación con el resto de las configuraciones.

Analizando los resultados de forma global, se aprecian 9 preguntas de público general para las que el sistema no es capaz de encontrar buenas respuestas (como CQ103). Sin embargo, no sucede en ningún caso para las preguntas de ámbito médico.

Las configuraciones 1 y 2 obtienen una puntuación similar en una misma respuesta, salvo algunos casos (como como la pregunta CQ104) en los que varía mucho el resultado. Sin embargo, la configuración 3 presenta resultados peores en ambas tareas, por lo tanto, podría decirse que, para las dos tareas, sería más beneficioso el uso de la configuración 1 ya que presenta mejores resultados en ambas, siendo bastante superior en el caso de preguntas para público general.

Las dos primeras configuraciones tienen pequeñas variaciones salvo en el corpus que es totalmente distinto, por lo que se aprecia en los resultados la importancia que tiene el texto con el que se pre-entrenan los modelos. La tercera configuración presenta otras variaciones más severas, de tipo estructural y de parametrización de valores.

Las configuraciones 1 y 3 obtienen siempre el máximo de respuestas posibles (se ha configurado como 100). Sin embargo, la configuración 2 no obtiene en casi ningún caso el número máximo

de respuestas, siendo 60 el número promedio de respuestas obtenidas. Esto no parece repercutir en los resultados ya que, aun ofreciendo el máximo de respuestas posibles, la configuración 3 se comporta peor que la configuración 2 que no llega al máximo de respuestas.

Por último, respecto al rendimiento de los tres escenarios, se observa que las configuraciones 1 y 2 presentan tiempos de ejecución muy similares. Sin embargo, la configuración 3 tiene un tiempo medio de ejecución por respuesta que prácticamente triplica al de las otras dos configuraciones.

7.2 Conclusiones

Los aspectos positivos más destacados del sistema sobre el que se ha basado la experimentación son los siguientes:

- El sistema intenta combatir la sobrecarga de información de dos formas: devuelve respuestas específicas a las preguntas y devuelve respuestas solo cuando es relevante, tratando de evitar respuestas de baja calidad.
- Dadas las preguntas, es completamente automático y no necesita ningún ajuste. Cuanto mejor sea la calidad de las preguntas, mejores serán las respuestas que proporciona el sistema.
- El sistema se puede utilizar para explorar fácilmente otras tareas y necesidades de información, ya que devuelve directamente las respuestas a la colección de documentos a través de nuevas preguntas.
- Responde a preguntas de distintos tipos o ámbitos sin necesidad de hacer ninguna adaptación. En el caso de este proyecto, se distinguieron entre dos tipos de preguntas distintos: de expertos y de consumidor generalista.
- El sistema es complementario a las técnicas de extracción de información que requieren mucha mano de obra y que intentan encontrar respuestas a tareas específicas mediante anotaciones o reglas construidas manualmente.

En cuanto a las configuraciones propuestas, se ha explorado el uso de modelos pre-entrenados para la tarea de extracción de respuesta, debido a que es un mecanismo ampliamente utilizado en el ámbito de NLP, especialmente en casos en los que no se dispone de tiempo o recursos para construir modelos de procesamiento de lenguaje natural desde cero. En este caso, se han seleccionado alternativas especialmente pensadas para el uso en el ámbito médico o científico.

Los resultados arrojan diferencias, sobre todo en la tarea de preguntas generalistas. Esto demuestra que utilizar un modelo pre-entrenado u otro tiene impacto en los resultados obtenidos a pesar de que el resto del sistema se mantenga sin cambios.

Para casos en los que se requiera de una implementación rápida y con buenos resultados, esta alternativa de métodos pre-entrenados es una buena solución, ya que están entrenados con baterías de datos muy extensas e introducirlos en los sistemas es muy sencillo. En este proyecto se han utilizado librerías externas para adaptar los modelos al proyecto y el resultado ha sido muy eficiente y positivo. Otra de las ventajas de este tipo de modelos debido a su sencillez es que permite realizar en poco tiempo una comparativa entre modelos, tal y como se ha mostrado en este proyecto.

Los resultados obtenidos para las distintas configuraciones muestran que el tipo de preguntas realizadas al sistema y, por tanto, el dataset empleado para dar respuesta a las mismas, hace que varíen los resultados a pesar de que el resto de las variables del sistema no varíen. La media

de los resultados no es significativamente distinta, pero se aprecian peores resultados en la tarea de respuesta a preguntas de usuario generalista, siendo incapaz de encontrar alguna respuesta en varias de las preguntas.

Una buena selección de un modelo pre-entrenado que se ajuste al idioma, temática del dataset y las preguntas, rendimiento necesario, etc. puede impactar significativamente en el desarrollo de un sistema de QA.

7.3 Trabajo futuro

En cuanto al sistema adaptado para la realización de las tareas propuestas, la interfaz podría ser más rica, permitiendo una exploración más profunda en los casos en los que el usuario quisiera explorar documentos y respuestas adicionales.

Actualmente, el sistema utiliza todos los ficheros del dataset que se le indiquen, creando para ellos si no existe un fichero de metadatos por lo que la incorporación de nuevos datasets o actualizaciones de los mismos es muy sencilla e implicaría la generación de un nuevo fichero de metadatos, usado por el sistema para indexar los documentos.

El sistema es lento para obtener resúmenes y documentos completos. La producción de índices más grandes y ricos puede acelerar considerablemente el sistema.

El sistema se puede mejorar fácilmente con un módulo de recuperación de información más sofisticado.

El sistema se puede mejorar fácilmente incorporando datos de desarrollo anotados específicos del dominio y un componente de aprendizaje continuo para seguir aprendiendo gracias a los comentarios de algunos usuarios expertos seleccionados a mano.

También se podría mejorar el sistema con una medida de confianza en las respuestas. Consistiría en introducir una medida de confianza mejorada que combine las puntuaciones de IR y QA en una medida unificada que evalúe automáticamente la calidad de las respuestas.

8 Anexos

8.1 Implementación

8.1.1 Archivos de entrada y configuración

Documentos de entrada

Los documentos de entrada son el conjunto de ficheros que contienen la información que se indexa y de la que se extraen las respuestas de los usuarios. En este proyecto, hay dos baterías de documentos distintas: por un lado, los *papers* y documentos científicos que dan respuesta a preguntas de usuarios expertos en la materia y por el otro, un conjunto de documentos obtenidos de información mostrada en webs y publicaciones generalistas para responder a las preguntas del público general (ver 5.1).

Archivo de metadatos

El sistema utiliza un fichero de metadatos que contiene información sobre algunas de las características de los documentos de entrada. En primer lugar, se carga la información del archivo de metadatos en un objeto *dataframe*, seleccionando solo algunas de las columnas del archivo que son de interés. Luego, se aplica un filtro para eliminar la información duplicada ya que es posible que haya documentos repetidos.

Archivo de sinónimos

El sistema dispone de un mecanismo opcional para filtrar los documentos que contengan información que no interesa al caso de estudio. En este proyecto, se dispone de un filtro opcional que elimina los artículos que tratan sobre coronavirus distintos del COVID-19 (por ejemplo, SARS-CoV y MERS). Para ello, se configura un parámetro indicando el uso de filtrado, se lee una lista de sinónimos de COVID-19 de un fichero de configuración y se verifica si aparece un sinónimo en el título o en el resumen de cada documento utilizando el fichero de metadatos. Por tanto, se descartan aquellos documentos que no incluyen ninguno de los sinónimos antes de comenzar el procesamiento de los ficheros.

Otro beneficio que aporta este filtrado en caso de aplicarse es la reducción del tamaño de los datos de entrada que es necesario indexar en el siguiente paso. Para la generación final de resultados de este trabajo, no se ha aplicado el filtro.

Archivos de preguntas

Las preguntas realizadas por los usuarios que son leídas y procesadas por el sistema, se encuentran en un fichero de entrada. Para este proyecto, se han utilizado dos baterías de preguntas distintas: preguntas de expertos y preguntas del público general (ver 5.1.2).

8.1.2 Hiperparámetros

El sistema propuesto permite al usuario la configuración de ciertos parámetros para facilitar el desarrollo y experimentación:

- **Número máximo de documentos**: número máximo de documentos que pueden extraerse como relevantes para una pregunta dada.
- **Número máximo de respuestas por párrafo**: número máximo de respuestas que pueden extraerse de un párrafo (o unidad de indexación). La combinación de este

parámetro junto con el número máximo de respuestas por párrafo debe generar un número de respuestas igual o inferior a 1000 respuestas que es el número máximo de respuestas permitidas por pregunta en el desafío.

- **Tamaño máximo de respuesta**: número máximo de caracteres, incluyendo espacios, que puede tener una respuesta proporcionada por el sistema. El límite propuesto por el desafío es a nivel de párrafo, no a nivel de carácter. Después de realizar experimentar con varios tamaños de parámetro, se optó por seleccionar un valor de 500 caracteres que es un tamaño medio para los párrafos de los documentos, pudiendo obtenerse una respuesta correspondiente a un párrafo completo.
- **Error máximo**: porcentaje de documentos mínimo necesario que deben contener una respuesta válida para considerar que es posible responder una pregunta dada. En caso de no sobrepasar el valor de error máximo, la respuesta mostrada por el sistema es "No sé".
- **Uso de filtro**: booleano que permite descartar documentos en base al contenido del fichero de sinónimos.

Los valores empleados para estos parámetros en la experimentación del proyecto para todos los escenarios propuestos han sido:

Parámetro	Valores utilizados en la experimentación
Número máximo de documentos	100
Número máximo de respuestas por párrafo	3
Tamaño máximo de respuesta	500
Error máximo	90
Uso de filtro	False

Tabla 10. Configuración de parámetros para la experimentación

8.1.3 Librerías Python utilizadas

La versión de Python utilizada para la ejecución del sistema ha sido Pyhton 3.7. A continuación, se describen las principales librerías empleadas en el código fuente desarrollado:

Nombre de la librería	Descripción
Whoosh	Contiene funciones para indexar texto y buscar en los índices creados con los documentos. (Whoosh, 2021)
Pandas	Contiene funciones para el manejo y análisis de estructuras de datos, principalmente dataframe. (Pandas, 2021)
Torch	Es una librería de aprendizaje automático de código abierto, un marco informático científico y un lenguaje de script basado en el lenguaje de programación Lua. Proporciona una amplia gama de algoritmos para *Deep learning.* (Torch, 2021)
TensorFlow	Es una librería para computación numérica rápida creada y lanzada por Google. Es una librería básica que se puede usar para crear modelos de *Deep learning* directamente o mediante el uso de bibliotecas contenedoras que simplifican el proceso. (Tensorflow, 2021)
Transformers	Proporciona arquitecturas de uso general (BERT, GPT-2, RoBERTa, XLM, DistilBert, XLNet ...) para la comprensión del lenguaje natural (NLU) y la generación del lenguaje natural (NLG) con más de 32 modelos previamente entrenados en más de 100 idiomas y una

	interoperabilidad profunda entre Jax, PyTorch y TensorFlow. (Transformers, 2021)
Marshal	Este módulo contiene funciones que pueden leer y escribir valores de Python en un formato binario. (Marshal, 2021)
Json	API para el manejo de datos en formato json para su uso en los módulos estándar de bibliotecas marshal y pickle. (Json, 2021)
Webbrowser	El módulo que provee una interfaz de alto nivel que permite desplegar documentos basados en la web. (Webbrowser, 2021)

8.1.3.1 Fichero de requerimientos

Contenido del fichero 'requirements.txt' para crear el entorno en Python con las librerías necesarias para el funcionamiento del sistema desarrollado:

```
absl-py==0.13.0
astunparse==1.6.3
backcall==0.2.0
cachetools==4.2.2
certifi==2021.5.30
charset-normalizer==2.0.4
clang==5.0
click==8.0.1
colorama==0.4.4
decorator==5.0.9
filelock==3.0.12
flatbuffers==1.12
gast==0.4.0
google-auth==1.35.0
google-auth-oauthlib==0.4.5
google-pasta==0.2.0
grpcio==1.39.0
h5py==3.1.0
huggingface-hub==0.0.12
idna==3.2
ipython==7.27.0
ipython-genutils==0.2.0
jedi==0.18.0
joblib==1.0.1
keras==2.6.0
Keras-Preprocessing==1.1.2
Markdown==3.3.4
matplotlib-inline==0.1.2
numpy==1.19.5
oauthlib==3.1.1
opt-einsum==3.3.0
packaging==21.0
pandas==1.3.2
parso==0.8.2
pickleshare==0.7.5
prompt-toolkit==3.0.20
protobuf==3.17.3
pyasn1==0.4.8
pyasn1-modules==0.2.8
Pygments==2.10.0
pyparsing==2.4.7
python-dateutil==2.8.2
pytz==2021.1
PyYAML==5.4.1
regex==2021.8.28
requests==2.26.0
requests-oauthlib==1.3.0
rsa==4.7.2
sacremoses==0.0.45
six==1.15.0
tensorboard==2.6.0
tensorboard-data-server==0.6.1
```

```
tensorboard-plugin-wit==1.8.0
tensorflow==2.6.0
tensorflow-estimator==2.6.0
termcolor==1.1.0
tokenizers==0.10.3
torch==1.9.0
tqdm==4.62.2
traitlets==5.0.5
transformers==4.9.2
typing-extensions==3.7.4.3
urllib3==1.26.6
wcwidth==0.2.5
Werkzeug==2.0.1
Whoosh==2.7.4
wrapt==1.12.1
```

8.2 Dataset

8.2.1 Colección de documentos

8.2.1.1 Ejemplo de artículo de investigación

Artículo disponible en el fichero '000ahtxm.json':


```
{
    "document_id": "000ahtxm",
    "metadata": {
        "title": "Epidemiological and clinical characteristics of 136 cases of COVID-19
in main district of Chongqing",
        "authors": "Chen, Peng; Zhang, Ying; Wen, Yongsheng; Guo, Jinjun; Jia, Jinwei;
Ma, Yu; Xu, Yi",
        "urls": [],
        "full_text_path": null
    },
    "contexts": [
        {
            "section": "Abstract",
            "context_id": "000ahtxm-C000",
            "text": "BACKGROUND: We did a comprehensive exploration of the epidemiological
and clinical characteristics of 136 patients with confirmed COVID-19 in main district of
Chongqing which was adjacent to the west of Hubei province. METHODS: This study was
conducted on 136 patients with COVID-19 in main district of Chongqing from Jan 25 to Feb
20, 2020. Data of patients included demographic, epidemiological, clinical features, chest
radiographs of imported cases, local cases, second-generation cases and third-generation
cases. Student's t-test was adopted for quantitative variables while Pearson Chi-squared
test or Fisher's exact test for categorical variables. RESULTS: The median age was 47
years and common symptoms of illness were cough (50.7%), fever (47.1%) and fatigue (14.0%).
The time from contact symptomatic case to illness was 7.7 days, and 88 patients (64.7%)
were cluster cases, radiological evidence found bilateral lung involvement was common
(57.4%).Compared with the imported cases, the local cases were significantly older, the
proportion of men is lower. There was higher proportion of cluster cases in local cases.
Unlike imported cases, which fever was the dominant symptom, the local cases have more
cough patients, with a significant higher proportion of asymptomatic patients. The third-
generation cases have a significant higher proportion of asymptomatic patients.
CONCLUSION: We concluded the epidemiological and clinical characteristics of the cases
andsuggested to take more comprehensive measures for screening patients, especially for
elderly person, avoid family gatherings, and implement more closely surveillance of
suspect patients and their close contacts.",
            "sentences": [
                {
                    "sentence_id": "000ahtxm-C000-S000",
                    "start": 0,
                    "end": 218
                },
                {
                    "sentence_id": "000ahtxm-C000-S001",
                    "start": 219,
                    "end": 341
```

```
                },
                {
                    "sentence_id": "000ahtxm-C000-S002",
                    "start": 342,
                    "end": 518
                },
                {
                    "sentence_id": "000ahtxm-C000-S003",
                    "start": 519,
                    "end": 655
                },
                {
                    "sentence_id": "000ahtxm-C000-S004",
                    "start": 656,
                    "end": 778
                },
                {
                    "sentence_id": "000ahtxm-C000-S005",
                    "start": 779,
                    "end": 1068
                },
                {
                    "sentence_id": "000ahtxm-C000-S006",
                    "start": 1069,
                    "end": 1129
                },
                {
                    "sentence_id": "000ahtxm-C000-S007",
                    "start": 1130,
                    "end": 1295
                },
                {
                    "sentence_id": "000ahtxm-C000-S008",
                    "start": 1296,
                    "end": 1385
                },
                {
                    "sentence_id": "000ahtxm-C000-S009",
                    "start": 1386,
                    "end": 1687
                }
            ]
        }
    ]
}
```


8.2.1.2 Ejemplo de artículos para el consumidor

Artículo disponible en el fichero '0000d9ae-6709-4611-a4b0-09390c095093-CC-NEWS-20200324132640-01359.json':


```
{
    "document_id": "0000d9ae-6709-4611-a4b0-09390c095093",
    "metadata": {
        "title": "Salford care home with coronavirus 'running out of masks' - BBC News",
        "url": "https://www.bbc.com/news/uk-england-manchester-52011673"
    },
    "contexts": [
        {
            "section": null,
            "context_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000",
            "text": "Homepage Accessibility links Skip to content Accessibility Help BBC
AccountNotifications Search Search the BBCSearch the BBC News BBC News Navigation Sections
Home Video World US & Canada UK selected Business Tech Science Stories Entertainment &
Arts Health In Pictures Reality Check World News TV Newsbeat Special Reports Explainers
The Reporters Have Your Say Manchester Manchester Salford care home with coronavirus
'running out of masks' 24 March 2020 Share this with Facebook Share this with Messenger
Share this with Twitter Share this with Email Share this with Facebook Share this with
```

WhatsApp Share this with Messenger Share this with Twitter Share Share this with These are external links and will open in a new window Email Share this with Email Facebook Share this with Facebook Messenger Share this with Messenger Messenger Share this with Messenger Twitter Share this with Twitter Pinterest Share this with Pinterest WhatsApp Share this with WhatsApp LinkedIn Share this with LinkedIn Copy this link Read more about sharing. These are external links and will open in a new window Close share panel Related Topics Coronavirus pandemic Image copyright Google Image caption The Broughtons care home said visiting was stopped three weeks ago in a bid to protect residents A care home where a resident developed severe coronavirus symptoms is running short of protective face masks for its staff. The resident, aged in his 80s, was taken to hospital from The Broughtons home in Salford last week. Director Bob Dhaliwal said the home had a week's supply left from a delivery of 300 masks on Thursday. The government acknowledged there had been Health Secretary Matt Hancock said a million face masks had been bought over the weekend and delivered to frontline staff in health and social care. It follows claims by NHS staff last week that the lack of protective gear was putting them at risk during the coronavirus crisis. 'Difficult time' Care home residents are considered to be at high risk of becoming seriously ill from Covid-19 infection due their age and underlying health conditions. Mr Dhaliwal said masks recommended by Public Health England (PHE) were proving impossible to source. \"We can't order them - no other suppliers have them - so we are relying on the local authority to send us more,\" the director of the Well Being Care Group said. A SIMPLE GUIDE: What are the symptoms? NEW GUIDANCE: What must I do? NEW RESTRICTIONS: What are they? LOOK-UP TOOL: Check cases in your area MAPS AND CHARTS: Visual guide to the outbreak The remaining 37 residents at The Broughtons have been told to stay in their rooms. \"It is a very difficult time,\" said Mr Dhaliwal. He said the home stopped relatives visiting three weeks ago to try to protect residents from the virus. ",
            "sentences": [
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S000",
                    "start": 0,
                    "end": 1039
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S001",
                    "start": 1040,
                    "end": 1406
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S002",
                    "start": 1407,
                    "end": 1506
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S003",
                    "start": 1507,
                    "end": 1609
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S004",
                    "start": 1610,
                    "end": 1801
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S005",
                    "start": 1802,
                    "end": 1931
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S006",
                    "start": 1932,
                    "end": 2100
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S007",
                    "start": 2101,
                    "end": 2201
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S008",
                    "start": 2202,

```
                    "end": 2362
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S009",
                    "start": 2363,
                    "end": 2401
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S010",
                    "start": 2402,
                    "end": 2431
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S011",
                    "start": 2432,
                    "end": 2464
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S012",
                    "start": 2465,
                    "end": 2633
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S013",
                    "start": 2634,
                    "end": 2682
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C000-S014",
                    "start": 2683,
                    "end": 2786
                }
            ]
        },
        {
            "section": null,
            "context_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001",
            "text": "A Department of Health and Social Care spokesperson said: \"We will give our NHS and the social care sector everything they need to tackle this outbreak. \" We have delivered millions more items of personal protective equipment to frontline staff in hospitals, ambulance trusts, GP practices, pharmacists, care homes and hospices and this urgent work continues. \" Related Topics Coronavirus pandemic Salford Share this story About sharing Email Facebook Messenger Messenger Twitter Pinterest WhatsApp LinkedIn More on this story Coronavirus: Hancock admits \u2018challenges\u2019 over NHS equipment 23 March 2020 Coronavirus: NHS staff 'at risk' over lack of protective gear 18 March 2020 Related Internet links Wellbeing Care Group Department of Health and Social Care The BBC is not responsible for the content of external Internet sites Top Stories Olympics moved to 2021 over coronavirus outbreak Organisers agree to a one-year postponement of the Olympics because of the coronavirus pandemic. 24 March 2020 Trump wants US open for business despite pandemic 24 March 2020 US-China contagion: The battle behind the scenes 24 March 2020 Features Trump wants US open for business despite pandemic BBC Future: Will hot weather kill Covid-19? Stranded abroad as coronavirus closes borders US-China contagion: The battle behind the scenes Life under coronavirus lockdown in Rome Video Teen figure skater's Special Olympics dream Antarctic seal photo wins top prize Could synthetic fish be a better catch of the day? Video Missing in Ukraine Elsewhere on the BBC Football phrases 15 sayings from around the world Full article Football phrases Why you can trust BBC News BBC News Navigation BBC News Services On your mobile On smart speakers Get news alerts Contact BBC News Explore the BBC Home News Sport Weather iPlayer Sounds CBBC CBeebies Food Bitesize Arts Taster Local Three Terms of Use About the BBC Privacy Policy Cookies Accessibility Help Parental Guidance Contact the BBC Get Personalised NewslettersCopyright \u00a9 2020 BBC. The BBC is not responsible for the content of external sites. Read about our approach to external linking.",
            "sentences": [
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S000",
                    "start": 0,
                    "end": 154
                },
                {
```

```
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S001",
                    "start": 155,
                    "end": 359
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S002",
                    "start": 360,
                    "end": 985
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S003",
                    "start": 986,
                    "end": 1229
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S004",
                    "start": 1230,
                    "end": 1501
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S005",
                    "start": 1502,
                    "end": 2018
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S006",
                    "start": 2019,
                    "end": 2080
                },
                {
                    "sentence_id": "0000d9ae-6709-4611-a4b0-09390c095093-C001-S007",
                    "start": 2081,
                    "end": 2125
                }
            ]
        }
    ]
}
```


8.2.2 Preguntas

Listado de preguntas utilizadas para la evaluación del sistema propuesto en formato *.json*:

8.2.2.1 Comunidad científica


```
[
    {
        "question_id": "EQ101",
        "question": "What features of SARS-CoV2 are targeted in vaccine development? ",
        "query": "coronavirus vaccine landscape",
        "background": "Seeking information on guiding principles for the design of COVID-19 vaccine strategies and analyses of the current COVID-19 vaccine landscape and the challenges. "
    },
    {
        "question_id": "EQ102",
        "question": "How do cytokine pathways link sleep and immunity to infection and COVID-19?",
        "query": "Sleep prevent covid",
        "background": "Looking for information on bidirectional interactions between sleep, health and immunity in animal and human studies, both for COVID-19 and in general."
    },
    {
        "question_id": "EQ103",
        "question": "When is it safe to discharge a COVID-19 patient with pneumonia? ",
        "query": "COVID-19 hospital discharge",
        "background": "Looking for guidelines on discharging COVID-19 patients. The following aspects are of interest: duration of SARS-CoV-2 virus shedding in bodily fluids
```

```
after remission and in asymptomatic patients; tests to document the lack of infectivity after infection, and the longest documented transmission."
    },
    {
        "question_id": "EQ104",
        "question": "What endocrine complications are  linked to COVID-19?",
        "query": "endocrine complications covid-19",
        "background": "Looking for studies investigating whether SARS-CoV-2 may directly attack the endocrine glands, which glands could be affected and what are the short- and long-term manifestations "
    },
    {
        "question_id": "EQ105",
        "question": "What is the duration of protection and immunity to SARS-CoV-2? ",
        "query": "COVID-19 immunity duration",
        "background":  "Looking for longitudinal serological studies that follow patients\u2019 immunity to study the duration of immunity. Models of protective immunity are also of interest."
    },
    {
        "question_id": "EQ106",
        "question": "What are the safety, tolerability, and immunogenicity properties of COVID-19 vaccines?",
        "query": "adverse reactions to COVID-19 vaccines",
        "background": "Looking for studies that evaluate tolerability and immunogenicity of COVID-19 vaccine in humans. The adverse events post-vaccination are of interest. "
    },
    {
        "question_id": "EQ107",
        "question": "What are the results of remdesivir trials for the treatment of COVID-19?",
        "query": "results of remdesivir trials",
        "background": "Looking for studies that evaluate improvement in clinical course, duration and severity of COVID-19 in people treated with remdesivir. The adverse reactions to the drug are of interest."
    },
    {
        "question_id": "EQ108",
        "question": "What is the diagnostic accuracy of the available COVID-19 tests?",
        "query": "covid tests accuracy",
        "background": "Looking for studies reporting the sensitivity and specificity of the available COVID-19 diagnostic tests."
    },
    {
        "question_id": "EQ109",
        "question": "Is the association between COVID-19 and diabetes driven by the dpp4 receptor?",
        "query": "covid-19  diabetes dpp4",
        "background": "Looking for studies on the role DPP4 plays in glucose and insulin metabolism and inflammation, as well as potential effect of DPP4 inhibitors on COVID-19"
    },
    {
        "question_id": "EQ110",
        "question": "What are the pros and cons for various types of masks in children aged 6 to 12 for COVID-19 prevention? ",
        "query": "masks for children at school ",
        "background": "Looking for studies on efficiency of different mask type for COVID-19 prevention in children, including compliance and psychological and physiologic effects."
    },
    {
        "question_id": "EQ111",
        "question": "What factors influence the duration of protection and immunity to SARS-CoV-2?  ",
        "query": "Factors that influence duration of immunity to COVID-19",
        "background": "Looking for studies of lasting immunity to SARS-CoV-2, such as severity of disease. Studies of reinfection with the same seasonal coronavirus are also of interest."
    },
```

```
    {
        "question_id": "EQ112",
        "question": "What are the implications of the known COVID-19 reinfection cases
for vaccination?",
        "query": "COVID-19 reinfection",
        "background": "Looking for studies confirming that people have been infected by
the same virus a second time, the severity of infection, the frequency of reinfection and
the protectiveness of the vaccines. "
    },
    {
        "question_id": "EQ113",
        "question": "May delaying the onset of adaptive immune responses during the early
phase of infections be a potential treatment for COVID\u201019?",
        "query": "Delayed COVID-19 immune response",
        "background": "Looking for studies of the interactions between the SARS-CoV-2
virus and the host immune response, specifically, whether delayed response could be
protective or drive mortality. "
    },
    {
        "question_id": "EQ114",
        "question": "Which COVID-19 vaccine trials were paused and what were the health
safety concerns?",
        "query": "pauses for COVID-19 vaccine trials",
        "background": "Looking for information about safety considerations that lead to
recommendations to pause enrollment in the COVID-19 clinical trials, the differences
between the paused trials and other similar studies, and the next steps after a trial was
paused. Specific details about the vaccines and drugs for which the trials were paused
are of interest, and so are the characteristics of the participants whose health conditions
caused the suspension. Considerations of balancing the safety concerns and the study
design, such as blinding, are of interest."
    },
    {
        "question_id": "EQ115",
        "question": "What are the genetic and immunologic underpinnings of severe COVID-
19?",
        "query": "covid gene links",
        "background": "Looking for studies of genetic and immunologic characteristics that
determine the patients\u2019 response to COVID-19, such as studies of genes that impact
immunity. "
    },
    {
        "question_id": "EQ116",
        "question": "What are the interactions between innate and adaptive immune
responses in COVID\u201019?",
        "query": "pathways triggered by SARS-CoV-2",
        "background": "Looking for studies of the innate and adaptive immune responses
and their role in severity of COVID-19. Modeling and predictions of the disease course if
the interactions are altered are of interest.  "
    },
    {
        "question_id": "EQ117",
        "question": "What measures  could mitigate resurgence of COVID-19?",
        "query": "covid-19 resurgence",
        "background": "Looking for studies relating the duration and effectiveness of
social distancing and COVID-19 resurgence, as well as clinical trials of public health
policies. "
    },
    {
        "question_id": "EQ118",
        "question": "How efficient are herd immunity approaches to COVID-19 containment?
",
        "query": "herd immunity vs. COVID-19 containment strategies",
        "background": "Looking for studies that compare uncontrolled and controlled COVID-
19 spread and other approaches  to achieving population immunity."
    },
    {
        "question_id": "EQ119",
        "question": "When should an employee suspected or confirmed to have COVID-19
return to work?",
```

```
        "query": "Return to work after COVID-19",
        "background": "Looking for guidelines concerning test results and other indicators
for deeming the patients\u2019 return to workplace safe. "
    },
    {
        "question_id": "EQ120",
        "question": "Which interleukins and IL-inhibitors are involved in COVID-19
pathways?",
        "query": "Interleukins and interleukin inhibitors in COVID-19",
        "background": "Looking for studies of  cellular and molecular pathways of COVID-
19 and potential therapeutic Interventions, such as interleukin inhibitors."
    },
    {
        "question_id": "EQ121",
        "question": "What are the characteristics and technological platforms of the
COVID-19 vaccines that reached human clinical trials?",
        "query": "COVID-19 vaccine candidates",
        "background": "Looking for studies of the current COVID-19 vaccine candidates and
their chances of success."
    },
    {
        "question_id": "EQ122",
        "question": "What is the role of corticosteroids in COVID-19?",
        "query": "COVID-19 and steroids",
        "background": "Looking for studies evaluating the dangers of treating COVID-19
patients with corticosteroids and potential benefits of steroid treatments. Interactions
of steroids, SARS-CoV-2 and the immune system are of interest. "
    },
    {
        "question_id": "EQ123",
        "question": "Is there any scientific evidence that any of the alternative remedies
can prevent or cure COVID-19?",
        "query": "Covid-19 and alternative medicine",
        "background": "Looking for clinical trials or observational studies showing
effects of any natural and alternative treatments for COVID-19."
    },
    {
        "question_id": "EQ124",
        "question": "How do   mutations in SARS-CoV-2 impact its infectivity and
antigenicity?",
        "query": "SARS-CoV-2 mutations",
        "background": "Looking for analyses of SARS-CoV-2 mutations, their frequency,
infectivity and antigenicity. "
    },
    {
        "question_id": "EQ125",
        "question": "What computational predictions of SARS-CoV-2 mutations have been
confirmed?",
        "query": "SARS-CoV-2 mutation predictions",
        "background": "Looking for studies that correlate computational predictions of
SARS-CoV-2 mutations with population studies or observations. "
    },
    {
        "question_id": "EQ126",
        "question": "How clean Is the air in passenger aircrafts and what are the COVID-
19 risk reduction steps for airports and airlines?",
        "query": "covid-19 air travel safety",
        "background": "Looking  for studies of the air quality in airplanes and airports,
and guidelines for reducing risk of COVID-19 infection at the airports and aircrafts. "
    },
    {
        "question_id": "EQ127",
        "question": "How are telehealth services used during the Covid-19 pandemic and
what is their impact on the population health?",
        "query": "covid-19 telehealth",
        "background": "Looking for studies of the role and benefits of telehealth services
during the Covid-19 pandemic. "
    },
    {
```

```
        "question_id": "EQ128",
        "question": "What COVID-19 vaccine distribution plans are fair and  will have the
most impact?",
        "query": "covid-19 vaccine distribution plans",
        "background": "Looking for availability of  COVID-19 vaccines and guidelines  for
their distribution."
    },
    {
        "question_id": "EQ129",
        "question": "What approaches are recommended for developing  children\u2019s
social and emotional coping skills during the COVID-19 pandemic?",
        "query": "covid-19 child social skills and psychology",
        "background": "Looking for guidelines and resources for ensuring children\u2019s
well-being during the COVID-19 pandemic."
    },
    {
        "question_id": "EQ130",
        "question": "What studies are measuring risk factors  of daily life activities
during COVID-19?",
        "query": "risks of everyday activities during COVID-19",
        "background": "Looking for studies measuring rates of COVID-19 infection and risk
factors among people adhering to protective measures during daily life activities"
    }
]
```


8.2.2.2 Público general


```
[
    {
        "question_id": "CQ101",
        "question": "If the common cold is a type of coronavirus and we're unable to find
a cure, why does the medical community have confidence we will find a vaccine for COVID-
19?",
        "query": "coronavirus vaccine landscape",
        "background": "Looking for information on global efforts to support vaccine
development, such as Operation Warp Speed, as well as pathways to develop vaccines, e.g.,
the vaccines that use killed germs."
    },
    {
        "question_id": "CQ102",
        "question": "Does adequate sleep prevent COVID-19?",
        "query": "Sleep prevent covid",
        "background": "Looking for any confirmation that better sleep may prevent COVID-
19 and the amount of sleep needed."
    },
    {
        "question_id": "CQ103",
        "question": "Can I go home when I have COVID-19 and pneumonia?",
        "query": "COVID-19 hospital discharge",
        "background": "looking for information on leaving hospital after the symptoms
stopped and the temperature is normal. Also looking for guidelines on need for self-
isolation after hospital discharge. "
    },
    {
        "question_id": "CQ104",
        "question": "What endocrine complications are  linked to COVID-19?",
        "query": "endocrine complications covid-19",
        "background": "Looking for information on damage to endocrine organs, such as
thyroid, and short- and long-term endocrine diseases caused by  SARS-CoV-2"
    },
    {
        "question_id": "CQ105",
        "question": "How long do COVID-19 antibodies stay in your system?",
        "query": "COVID-19 immunity duration",
        "background": "Looking for information on how long a person will not get COVID
after having it. "
    },
    {
```

```
        "question_id": "CQ106",
        "question": "How are people coping with the COVID-19 vaccine?",
        "query": "adverse reactions to COVID-19 vaccines",
        "background": "Looking for information on complications that could be caused by
COVID-19 vaccines."
    },
    {
        "question_id": "CQ107",
        "question": "Is remdesivir recommended as treatment for COVID-19?",
        "query": "results of remdesivir trials",
        "background": "Looking for information on the pros and cons of getting remdesivir
treatments for COVID-19."
    },
    {
        "question_id": "CQ108",
        "question": "How often is a COVID-19 test result wrong?",
        "query": "covid tests accuracy",
        "background": "Looking for information on how likely the COVID-19 test will show
I don\u2019t have the virus even  if I have it or show that I have the disease even if I
don\u2019t.  "
    },
    {
        "question_id": "CQ109",
        "question": "What anti-diabetic medications are the safest during the COVID-19
pandemic?",
        "query": "covid-19  diabetes dpp4",
        "background": "Looking for information on specific risks of COVID-19 for diabetics
and links between COVID-19 and anti-diabetic medications."
    },
    {
        "question_id": "CQ110",
        "question": "Do school-age children (ages 6-12) who wear n95 masks to school
versus wearing surgical masks to school have a reduced risk of contracting COVID-19?",
        "query": "masks for children at school ",
        "background": "Looking for information on which masks will best prevent COVID-19
in my child with least harm."
    },
    {
        "question_id": "CQ111",
        "question": "If I donate plasma, could I reduce my own immunity to COVID-19?",
        "query": "Factors that influence duration of immunity to COVID-19",
        "background": "Looking for information on how quickly my antibodies will be
replenished id I donate convalescent plasma, and  plasma donation safety issues overall.
"
    },
    {
        "question_id": "CQ112",
        "question": "Why is COVID-19 more severe the second time?",
        "query": "COVID-19 reinfection",
        "background": "Looking for information about severity of COVID-19 reinfection,
how likely is it to get the same virus a second time, and how often can people get
reinfected."
    },
    {
        "question_id": "CQ113",
        "question": "What is delayed COVID-19 immune response?",
        "query": "Delayed COVID-19 immune response",
        "background": "Looking for definition of delayed immune response and its role in
the severity of COVID-19."
    },
    {
        "question_id": "CQ114",
        "question": "Why were COVID-19 vaccine trials paused?",
        "query": "pauses for COVID-19 vaccine trials",
        "background": "Looking for reasons to pause clinical trials, such as adverse
reactions, illness in the participants and other reasons. Which trials were stopped, for
what vaccines and whether the trials were resumed. How were the reasons for halting the
trials resolved?"
    },
```

```
    {
        "question_id": "CQ115",
        "question": "Could a person\u2019s DNA explain why some get hit hard by the
coronavirus?",
        "query": "covid gene links",
        "background": "Looking for information on genetic factors, such as gender, that
influence outcomes of COVID-19."
    },
    {
        "question_id": "CQ116",
        "question": "What could prevent mild COVID-19 from getting worse?",
        "query": "pathways triggered by SARS-CoV-2",
        "background": "Looking for information on   preventing  the immune system from
overreacting to  COVID-19. "
    },
    {
        "question_id": "CQ117",
        "question": "What are recommendations and advice in case of COVID-19 resurgence?",
        "query": "covid-19 resurgence",
        "background": "Looking for information on preventing getting COVID-19 as it
resurges. "
    },
    {
        "question_id": "CQ118",
        "question": "Is Swedish approach to the COVID-19 pandemic working?",
        "query": "herd immunity vs. COVID-19 containment strategies",
        "background": "Looking for information about health outcomes and economic
consequences of COVID-19 in countries that did not establish physical distancing. "
    },
    {
        "question_id": "CQ119",
        "question": "How long after I feel better from COVID-19 can I go back to work?",
        "query": "Return to work after COVID-19",
        "background": "Looking for information on returning to work after recovering from
COVID-19. "
    },
    {
        "question_id": "CQ120",
        "question": "What roles do  interleukins and IL-inhibitors play in COVID-19?",
        "query": "Interleukins and interleukin inhibitors in COVID-19",
        "background": "Looking for basic information about interleukins and their role in
COVID-19."
    },
    {
        "question_id": "CQ121",
        "question": "Which of the COVID-19 vaccines will be most effective and safe?",
        "query": "COVID-19 vaccine candidates",
        "background": "Looking information on COVID-19 vaccine safety, potential side-
effects caused by the vaccines and their ability to prevent COVID-19."
    },
    {
        "question_id": "CQ122",
        "question": "Are steroids  dangerous or protective if I have COVID-19?",
        "query": "COVID-19 and steroids",
        "background": "Looking for information on side-effects and benefits of taking
steroids in people infected with SARS-CoV-2. "
    },
    {
        "question_id": "CQ123",
        "question": "Are there alternative treatments for COVID-19?",
        "query": "Covid-19 and alternative medicine",
        "background": "Looking for information on supplements and life-style changes that
could prevent or treat COVID-19."
    },
    {
        "question_id": "CQ124",
        "question": "Is COVID-19 getting stronger as the virus mutates?",
        "query": "SARS-CoV-2 mutations",
```

```
        "background": "Looking for information about changes in the SARS-CoV-2 virus that
could make it more dangerous. "
    },
    {
        "question_id": "CQ125",
        "question": "Can science predict how coronavirus will change?",
        "query": "SARS-CoV-2 mutation predictions",
        "background": "Looking for information supporting or contradicting the predictions
that coronavirus will mutate into more infectious COVID-19."
    },
    {
        "question_id": "CQ126",
        "question": "How safe is air travel during the COVID-19 pandemic?",
        "query": "covid-19 air travel safety",
        "background": "Looking for information about COVID-19 infection spread related to
air travel. "
    },
    {
        "question_id": "CQ127",
        "question": "When is it safe to visit my doctor virtually and when do I have to
see the doctor in-person?",
        "query": "covid-19 telehealth",
        "background": "Looking for information on when visiting my doctor virtually is
enough and will not reduce the quality of healthcare. "
    },
    {
        "question_id": "CQ128",
        "question": "Who will get COVID--19 vaccine first and why?",
        "query": "covid-19 vaccine distribution plans",
        "background": "Looking for information on when different populations and
professions will get COVID-19 vaccines."
    },
    {
        "question_id": "CQ129",
        "question": "How to build my child\u2019s social skills and prevent psychological
harm during COVID-19?",
        "query": "covid-19 child social skills and psychology",
        "background": "Looking for information on supporting my child\u2019s development
and reducing harm from the COVID-19 pandemic."
    },
    {
        "question_id": "CQ130",
        "question": "How risky are everyday life activities during COVID-19?",
        "query": "risks of everyday activities during COVID-19",
        "background": "Looking for risks of getting COVID-19 while shopping, recreation
and other everyday life activities."
    }
]
```


8.2.3 Juicios de relevancia de los pasajes y nuggets por pregunta

Ejemplo de una parte del fichero que contiene tanto los juicios de relevancia de los pasajes, como los *nuggets* por pregunta:


```
[
    {
        "question_id": "CQ101",
        "nuggets": [
            {
                "nugget_id": "CQ101-N00",
                "nugget": "Common Cold"
            },
            {
                "nugget_id": "CQ101-N01",
                "nugget": "Companies Efforts"
            },
            {
                "nugget_id": "CQ101-N02",
```

```
                "nugget": "Government Agency Efforts"
            },

                                                ...
        ],
        "annotations": [
            {
                "sentence_id": "0821f151476049f5d3602c2d67947c929e6b1d3a-C000-S000",
                "nugget_ids": [
                    "CQ101-N00"
                ]
            },
            {
                "sentence_id": "0821f151476049f5d3602c2d67947c929e6b1d3a-C000-S002",
                "nugget_ids": [
                    "CQ101-N00"
                ]
            },
            {
                "sentence_id": "0821f151476049f5d3602c2d67947c929e6b1d3a-C000-S001",
                "nugget_ids": [
                    "CQ101-N10"
                ]
            },

                                                    ...
        ]
    }
]
```


8.3 Evaluación

8.3.1 Fichero de resultados

Ejemplo de un fichero de entrada con los resultados utilizado para la evaluación automática proporcionada por EPIQ-QA:

```
CQ001                                Q0     0a0b55c3563642cae734c7d60427d93dd6eabb9c-C000-
S000:0a0b55c3563642cae734c7d60427d93dd6eabb9c-C000-S000 1     0.999 EXAMPLE_RUN
CQ001                                Q0     0ae18966b18bc877fc498926e7025a22983605ad-C004-
S000:0ae18966b18bc877fc498926e7025a22983605ad-C004-S002 2     0.998 EXAMPLE_RUN
```

8.3.2 Fichero de cálculo de resultados

Fichero *.bat* utilizado para ejecutar la evaluación de resultados que se realiza en el fichero 'epiq_eval.py', proporcionado por EPIQ-QA:

```
python   epic_eval.py   primary_judgments.json   results_Research_questions_config1.txt
expert_ideal_ranking_scores.tsv --task expert > results_Research_questions_config1.eval
```

8.4 Tablas de resultados

Tablas de resultados obtenidas en la evaluación de las respuestas utilizando los dos bloques de preguntas y las tres configuraciones implementadas en el sistema. Además, tabla de datos de las respuestas proporcionadas como ideales por el sistema y utilizadas para las comparativas de resultados:

8.4.1 Resultados de respuestas para comunidad científica

8.4.1.1 Configuración 1

Configuración 1	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
EQ101	0.24454386	0.224335621	0.18026715

EQ102	0.283431114	0.283431114	0.27748844
EQ103	0	0	0
EQ104	0.156990728	0.156990728	0.1532644
EQ105	0.053784922	0.053784922	0.04929293
EQ106	0.452762669	0.426858093	0.46706192
EQ107	0.04709275	0.04709275	0.04626161
EQ108	0.117914774	0.101852171	0.10493288
EQ109	0.6	0.6	0.5
EQ110	0.052994762	0.052461336	0.06816305
EQ111	0.05131334	0.049046389	0.05713637
EQ112	0.260714459	0.240967164	0.24997905
EQ113	0.469393197	0.457186932	0.36876223
EQ114	0.087489603	0.087489603	0.09054495
EQ115	0.025446197	0.025530367	0.02670582
EQ116	0.082455958	0.070202513	0.06738365
EQ117	0.086684045	0.086684045	0.08683466
EQ118	0.193071028	0.169365052	0.19332726
EQ119	0.034965041	0.034965041	0.04591777
EQ120	0.347780614	0.349123243	0.30498284
EQ121	0.245123216	0.230218721	0.19101566
EQ122	0.086900935	0.073755036	0.08189987
EQ123	0.435660304	0.420177174	0.45933766
EQ124	0.128325957	0.128325957	0.14928325
EQ125	0.175504597	0.175504597	0.17903288
EQ126	0.122295244	0.117497048	0.13806806
EQ127	0.315953981	0.302152618	0.27731526
EQ128	0.139428742	0.136022525	0.13049358
EQ129	0.040838532	0.040838532	0.03678995
EQ130	0.276896585	0.276896585	0.23816156
MEAN	0.187191905	0.180625196	0.17399016

8.4.1.2 Configuración 2

Configuración 1	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
EQ101	0.202082384	0.202082384	0.179031871
EQ102	0.200374367	0.200374367	0.219332686
EQ103	0.055695178	0.055695178	0.056272493
EQ104	0.080467121	0.080467121	0.086318322
EQ105	0.027181413	0.027181413	0.028756724
EQ106	0.381713731	0.381713731	0.422428044
EQ107	0.041685843	0.041344304	0.043012872
EQ108	0.124713523	0.124713523	0.129293443
EQ109	0.75	0.75	0.75
EQ110	0.155108858	0.153547587	0.172932634
EQ111	0.031721036	0.032604859	0.036677374
EQ112	0.235457501	0.227264341	0.256876426
EQ113	0.169200484	0.169200484	0.164319232
EQ114	0.181556298	0.181556298	0.185205964

EQ115	0	0	0
EQ116	0.069463203	0.06663793	0.063373046
EQ117	0.045281903	0.045281903	0.045360581
EQ118	0.146435038	0.141287998	0.149741659
EQ119	0.009194569	0.009194569	0.01207475
EQ120	0.157213225	0.157750725	0.162557577
EQ121	0.154965303	0.154965303	0.130452447
EQ122	0.080042955	0.080558956	0.082194807
EQ123	0.464802619	0.446284119	0.513020121
EQ124	0.110617662	0.110617662	0.12868296
EQ125	0.200885395	0.200885395	0.137294189
EQ126	0.165090806	0.157494271	0.146605504
EQ127	0.311213912	0.295066603	0.284132257
EQ128	0.317035073	0.287021615	0.292556758
EQ129	0.037040564	0.037040564	0.033368502
EQ130	0.203494475	0.203494475	0.113129719
MEAN	0.170324481	0.167377589	0.167500099

8.4.1.3 Configuración 3

Configuración 1	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
EQ101	0.161665907	0.161665907	0.195307496
EQ102	0.077771372	0.077771372	0.085129671
EQ103	0.011102585	0.011102585	0.01335645
EQ104	0.079662193	0.079662193	0.085454863
EQ105	0.052873944	0.052873944	0.044750622
EQ106	0.404629275	0.394737091	0.474769607
EQ107	0.026259632	0.026259632	0.024012512
EQ108	0.170749714	0.170749714	0.199464544
EQ109	0.678103594	0.678103594	0.678103594
EQ110	0.080560596	0.079749702	0.103618842
EQ111	0.047262935	0.048579791	0.05578946
EQ112	0.219221023	0.219221023	0.253263354
EQ113	0.134598404	0.134598404	0.138390403
EQ114	0.152957256	0.152957256	0.158298883
EQ115	0	0	0
EQ116	0.029718076	0.029718076	0.033594841
EQ117	0.054532321	0.054532321	0.054627072
EQ118	0.145718603	0.145718603	0.188705559
EQ119	0.021299872	0.021299872	0.027972015
EQ120	0.246099936	0.249085003	0.260997405
EQ121	0.066314105	0.066314105	0.067759566
EQ122	0.050871553	0.04817985	0.057361425
EQ123	0.314737089	0.315483845	0.406396397
EQ124	0.058404714	0.058404714	0.067942961
EQ125	0	0	0
EQ126	0.064878732	0.06521519	0.08031986
EQ127	0.087059235	0.08325919	0.094709879

EQ128	0.090748487	0.091534826	0.107538327
EQ129	0	0	0
EQ130	0.218874879	0.218874879	0.218874879
MEAN	0.124889201	0.124521756	0.139217016

8.4.1.4 Resultados ideales

Ideal	NDNS-Partial Ideal	NDNS-Relaxed Ideal	NDNS-Exact Ideal
EQ101	6.18559606	6.18559606	5.12013118
EQ102	9.745688041	9.745688041	8.90330624
EQ103	14.0157806	14.0157806	11.6506548
EQ104	7.830535705	7.830535705	7.29973255
EQ105	10.39278926	10.39278926	9.82346582
EQ106	16.84831827	16.84831827	13.646741
EQ107	23.43118214	23.43118214	21.5998795
EQ108	4.630929754	4.630929754	3.96426309
EQ109	2	2	2
EQ110	25.49668516	25.75593526	19.8229213
EQ111	20.03736776	19.49421313	15.7184083
EQ112	4.561606312	4.561606312	3.94845912
EQ113	5.07464507	5.07464507	4.93559606
EQ114	11.33904339	11.33904339	10.9564194
EQ115	21.23996373	21.16993858	20.2381482
EQ116	12.75337032	12.75337032	11.2816618
EQ117	5.959024215	5.959024215	5.94868823
EQ118	17.97217813	17.97217813	13.8781321
EQ119	13.09599145	13.09599145	9.97221497
EQ120	15.99795491	15.80623331	15.0848077
EQ121	6.978250351	6.978250351	6.82938887
EQ122	17.78184984	17.66795264	14.8399259
EQ123	16.34031271	16.30163489	12.6548919
EQ124	3.561606312	3.561606312	3.06160631
EQ125	3.130929754	3.130929754	2.63092975
EQ126	32.97355778	32.80344069	26.6345413
EQ127	20.18106094	19.78848243	17.3959995
EQ128	19.40104003	19.23437336	14.3845525
EQ129	9.140995184	9.140995184	7.61019203
EQ130	5.268929393	5.268929393	5.26892939
MEAN	12.77890609	12.7313198	10.9034863

8.4.2 Resultados de respuestas para público general

8.4.2.1 Configuración 1

Configuración 1	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
CQ101	0.089752484	0.097497051	0.08391218
CQ102	0.363759106	0.355208301	0.419589692
CQ103	0	0	0
CQ104	0.469278726	0.469278726	0.469278726

CQ105	0.158246429	0.138851783	0.121461532
CQ106	0	0	0
CQ107	0.104238761	0.104238761	0.121782946
CQ108	0.030057619	0.030057619	0.036007797
CQ109	0.121109452	0.121109452	0.100924543
CQ110	0.319710515	0.32163652	0.35530375
CQ111	0.351784746	0.317925563	0.299055593
CQ112	0	0	0
CQ113	0	0	0
CQ114	0	0	0
CQ115	0.195460615	0.186345493	0.186345493
CQ116	0.021157017	0.019301223	0.022445332
CQ117	0.085697721	0.085697721	0.085697721
CQ118	0.309825584	0.302888584	0.411950908
CQ119	0	0	0
CQ120	0.288373673	0.279220197	0.260222806
CQ121	0.095854355	0.09772938	0.094700306
CQ122	0.096703736	0.085489892	0.068626017
CQ123	0.217837238	0.201701366	0.223908124
CQ124	0	0	0
CQ125	0	0	0
CQ126	0.089961363	0.089243387	0.082853826
CQ127	0.118808231	0.118112568	0.107310656
CQ128	0.067842569	0.067842569	0.069410418
CQ129	0	0	0
CQ130	0.206663958	0.206663958	0.165331167
MEAN	0.126737463	0.123201337	0.126203984

8.4.2.2 Configuración 2

Configuración 1	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
CQ101	0.083630097	0.090846375	0.093825821
CQ102	0.067326837	0.06938158	0.077196176
CQ103	0	0	0
CQ104	0.015211813	0.015211813	0.015211813
CQ105	0.164228142	0.160022521	0.143956513
CQ106	0	0	0
CQ107	0.084602342	0.084602342	0.103961357
CQ108	0.037340989	0.037340989	0.039141354
CQ109	0.19195872	0.19195872	0.19195872
CQ110	0.121528788	0.11839167	0.147550334
CQ111	0.452750083	0.432123359	0.26964539
CQ112	0	0	0
CQ113	0	0	0
CQ114	0	0	0
CQ115	0.163613904	0.155707399	0.155707399
CQ116	0.007930109	0.007930109	0.009617259
CQ117	0.096446791	0.096446791	0.096446791

CQ118	0.192483026	0.168881999	0.227362336
CQ119	0	0	0
CQ120	0.313298218	0.313298218	0.337802249
CQ121	0.101411088	0.102541147	0.073661934
CQ122	0.095895844	0.088896915	0.062283756
CQ123	0.484841594	0.482977904	0.415726763
CQ124	0	0	0
CQ125	0	0	0
CQ126	0.050953948	0.047914377	0.045569123
CQ127	0.09084093	0.09084093	0.084323577
CQ128	0.085356448	0.085356448	0.070230393
CQ129	0	0	0
CQ130	0.115488384	0.115488384	0.093320546
MEAN	0.10057127	0.098538666	0.091816653

8.4.2.3 Configuración 3

Configuración 1	NDNS-Partial	NDNS-Relaxed	NDNS-Exact
CQ101	0.049862491	0.054165028	0.055941453
CQ102	0.070732487	0.072891167	0.086553465
CQ103	0	0	0
CQ104	0	0	0
CQ105	0.111372868	0.111372868	0.119288925
CQ106	0	0	0
CQ107	0.051770742	0.051770742	0.063617111
CQ108	0.053271866	0.053271866	0.063817515
CQ109	0	0	0
CQ110	0.211019653	0.21140578	0.288319235
CQ111	0.379640016	0.373059613	0.40940161
CQ112	0	0	0
CQ113	0	0	0
CQ114	0	0	0
CQ115	0.239978842	0.239978842	0.239978842
CQ116	0.008997863	0.008997863	0.010912181
CQ117	0.138862444	0.138862444	0.138862444
CQ118	0.201985063	0.188716056	0.28805488
CQ119	0	0	0
CQ120	0.242515223	0.242515223	0.261483096
CQ121	0.096361108	0.098687771	0.098146049
CQ122	0.085593773	0.085593773	0.070436918
CQ123	0.189806681	0.191062038	0.252518177
CQ124	0	0	0
CQ125	0	0	0
CQ126	0.041634921	0.041634921	0.049398645
CQ127	0.017351362	0.017351362	0.02001266
CQ128	0.042678224	0.042678224	0.046820262
CQ129	0	0	0
CQ130	0.264295073	0.264295073	0.209702636

MEAN	0.08325769	0.082943689	0.092442203

8.4.2.4 Resultados ideales

Ideal	NDNS-Partial Ideal	NDNS-Relaxed Ideal	NDNS-Exact Ideal
CQ101	8.43559606	7.765524184	7.518929393
CQ102	6.679388872	6.679388872	5.43559606
CQ103	24.36496739	23.99589715	17.35414273
CQ104	17.32991057	17.21604588	13.02612294
CQ105	2.630929754	2.630929754	2.630929754
CQ106	15.7627089	15.7627089	12.99746593
CQ107	6.892789261	6.892789261	6.392789261
CQ108	23.52324506	22.96866112	18.00098914
CQ109	10.09124838	10.09124838	9.198508131
CQ110	6.761859507	6.561606312	5.525869399
CQ111	6.079388872	6.079388872	6.079388872
CQ112	13.46829536	13.46829536	11.42566856
CQ113	2.130929754	2.130929754	2.130929754
CQ114	5.561606312	5.561606312	5.192536065
CQ115	2	2	1.630929754
CQ116	6.412722206	6.412722206	6.43559606
CQ117	3.630929754	3.630929754	3.030929754
CQ118	2	2	2
CQ119	8.761859507	8.761859507	7.984081729
CQ120	5.679388872	5.679388872	4.93559606
CQ121	4.579388872	4.579388872	4.579388872
CQ122	3.130929754	3.130929754	2.630929754
CQ123	5.823465819	5.823465819	5.514231409
CQ124	4.543559338	4.543559338	4.543559338
CQ125	16.86185951	16.87707119	10.73889005
CQ126	9.399859146	9.399859146	8.149859146
CQ127	2.630929754	2.630929754	2.130929754
CQ128	1.630929754	1.630929754	1.630929754
CQ129	19.99775003	19.99775003	16.85480962
CQ130	6.656799152	6.656799152	5.479388872
MEAN	8.448441184	8.385353442	7.039330531

8.5 Métricas en recuperación de información

Hay muchas medidas propuestas para evaluar el rendimiento de los sistemas de recuperación de información. Las medidas necesitan una colección de documentos y una consulta. A continuación, son descritas algunas medidas comunes, las cuales asumen que cada documento se sabe que es relevante o no relevante para una consulta particular:

Precisión (Precision)

La precisión es la fracción de documentos recuperados que son relevantes para la necesidad de información del usuario:

$$Precision = \frac{|\{documentos\ relevantes\} \cap documentos\ recuperados|}{|\{documentos\ recuperados\}|}$$

La precisión tiene en cuenta todos los documentos recuperados. También puede ser evaluada en un corte determinado del ranking, considerando solamente los primeros resultados obtenidos del sistema.

Exhaustividad (Recall)

La exhaustividad es la fracción de documentos relevantes para una consulta que fueron recuperados:

$$\frac{|\{documentos\ relevantes\} \cap \{documentos\ recuperados\}|}{|\{documentos\ relevantes\}|}$$

Resulta trivial obtener un 100% de exhaustividad si se toman como respuesta para cualquier consulta todos los documentos de la colección. Por lo tanto, la exhaustividad sola no es suficiente, sino que se necesita también medir el número de documentos no relevantes, por ejemplo, con el cálculo de la precisión.

Proposición de fallo (Fall-out)

La proposición de fallo es la proporción de documentos no relevantes que son recuperados, fuera de todos los documentos relevantes disponibles:

$$fall - out = \frac{|\{documentos\ no\ relevantes\} \cap \{documentos\ recuperados\}|}{|\{documentos\ no\ relevantes\}|}$$

Resulta trivial obtener un 0% de proposición de fallo si no se devuelve ningún documento de la colección para cualquier consulta.

Medida F (F-score)

La medida F es un balance de la precisión y la exhaustividad:

$$F = 2\frac{Precision \cdot Exhaustividad}{Precision + Exhaustividad}$$

Ésta es conocida también como la medida F_1, ya que la precisión y la exhaustividad toman pesos uniformes.

La fórmula general para el parámetro real no negativo β es:

$$F_\beta = \frac{(1+\beta^2) \cdot (Precision \cdot Exhaustividad)}{\beta^2 \cdot Precision + Exhaustividad}$$

Si β es igual a uno, se está dando la misma ponderación (o importancia) a la precisión que a la exhaustividad, si β es mayor que uno, implica que se da más importancia a la exhaustividad, mientras que si es menor que uno, se le da más importancia a la precisión.

Otras dos medidas F ampliamente utilizadas son la medida F_2, que pondera la exhaustividad dos veces por encima de la precisión, y la medida $F_{0.5}$, que pesa la precisión dos veces por encima de la exhaustividad.

Precisión Promedio

La precisión y la exhaustividad son métricas basadas en toda la lista de documentos devuelta por el sistema dada una consulta. Para sistemas que hacen ranking a los documentos devueltos para una consulta es deseable considerar además el orden en que los documentos devueltos son presentados. Si se miden la precisión y exhaustividad en cada posición de la secuencia de documentos con ranking, se obtiene la suma finita sobre todas las posiciones en la secuencia de documentos con ranking:

$$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{número\ de\ documentos\ relevantes}$$

donde $rel(k)$ es un indicador igual a 1 si el ítem en la posición k del ranking es relevante al documento, y cero en otro caso.

Media de la precisión promedio (Mean Average Precision)

La media de la precisión promedio, para un conjunto de consultas es el promedio de las puntuaciones medias de precisión para cada consulta:

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$$

donde Q es el número de consultas que se están evaluando.

Término Frecuencia de frecuencia de documento inversa (TF – IDF)

El propósito de esta métrica es convertir el documento en un modelo vectorial basado en la frecuencia con la que aparecen las palabras en el documento (independientemente del orden estricto). Siendo N el número de documentos, para cada documento D, se define:

- Término Frecuencia (tf): para un término t, su tf en D es el número de veces que aparece t en D.
- Frecuencia de documento inversa (IDF): para un término, IDF se define como el logaritmo de una relación del número de todos los documentos en el corpus y el número de documentos que contienen el término t.

La fórmula de TF-IDF da la importancia relativa de un término en un corpus (una colección de documentos):

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

8.6 Términos de interés

Corpus

Un corpus es un conjunto de textos naturales o fragmentos de textos, almacenados en formato electrónico, representativos en su conjunto de una variedad lingüística, en alguno de sus componentes o en su totalidad, y reunidos con el propósito de facilitar su estudio científico.

Tipos de preguntas

Diferentes tipos de preguntas requieren el uso de diferentes estrategias para encontrar la respuesta. Los tipos de cuestiones se organizan jerárquicamente en taxonomías.

Taxonomías

Una taxonomía es un conjunto organizado de palabras o frases que se utiliza para organizar la información con la finalidad de facilitar la búsqueda.

Procesamiento de preguntas

La misma pregunta puede ser expresada de varias formas (interrogativa o asertivamente). Un modelo semántico que entienda estos tipos de preguntas es necesario para reconocer cuestiones equivalentes. Este modelo permite la transición de una cuestión compleja a varias cuestiones simples.

Contexto de QA

Las preguntas son usualmente formuladas con un contexto y respondidas con ese mismo contexto. El contexto se puede usar para clarificar una cuestión, resolver ambigüedades.

Recursos de datos

Antes de que una pregunta pueda ser contestada, se debe saber que recursos de sabiduría están disponibles. Si la respuesta a una pregunta no está en esos recursos, no importa cuán bien procesemos la pregunta, pues no obtendremos una respuesta correcta.

Extracción de la respuesta

La extracción de la respuesta depende de la complejidad de la pregunta, en el tipo de respuesta seleccionada en el procesamiento de cuestiones, en los datos que disponemos y en el método de búsqueda.

Formulación de la respuesta

En resultado del sistema de QA debe ser presentado en un lenguaje tan natural como sea posible. En algunos casos, la extracción simple es insuficiente. Por ejemplo, cuanto la clasificación de una cuestión indica que la pregunta es el tipo nombre, una cantidad, o una fecha, la extracción del dato es suficiente. Para otros casos, la presentación de la respuesta puede requerir el uso de otros tipos de técnicas.

QA en tiempo real

La respuesta en tiempo real es muy necesaria. Estos sistemas tienen que ser capaces de procesar grandes cantidades de datos en un tiempo reducido.

Tagging Part of Speech (PoS)

Clasificar las oraciones en verbo, sustantivo, adjetivo preposición, etc.

Shallow parsing / Chunks

Sirve para entender la gramática en las oraciones. Se hace un parseo de los *tokens* y a partir de su PoS se arma un árbol de la estructura.

Pragmatic Analysis

Este tipo de análisis se realiza RNN "clásica" de extremo a extremo. Esta operación se visualiza para detectar con qué intencionalidad se dicen las cosas: ironía, sarcasmo, intencionalidad, etc.

Bag of words

Es una manera de representar el vocabulario que se utiliza en los modelos de NLP y consiste en una matriz en la que cada columna es un *token* contabilizando la cantidad de veces que aparece ese *token* en cada oración (representadas en cada fila).

word2vec

Es una técnica que aprende de leer enormes cantidades de textos y memorizar qué palabras parecen ser similares en diversos contextos. Después de entrenar suficientes datos, se generan vectores de 300 dimensiones para cada palabra conformando un nuevo vocabulario en donde las palabras “similares” se ubican cercanas unas de otras. Utilizando vectores pre-entrenados, se logra tener muchísima riqueza de información para comprender el significado semántico de los textos.

Bibliografía

(2020). Obtenido de COVID-19 Open Research Dataset Challengue: https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

BERT Transformers. (2021). Obtenido de https://huggingface.co/transformers/v3.4.0/model_doc/bert.html?highlight=bertforquestionanswering

Búsqueda de respuestas. (2021). Obtenido de Wikipedia: https://es.wikipedia.org/wiki/B%C3%BAsqueda_de_respuestas

Deepset. (2020). Obtenido de https://github.com/deepset-ai/COVID-QA

EPIQ-QA. (2020). Obtenido de Epidemic Question Answering: https://bionlp.nlm.nih.gov/epic_qa/#

Gabriel Jaimes, L. &. (2005). Modelos clásicos de recuperación de la información. *Revista Integración*.

García, E. (2011). *A Tutorial on the BM25F Model.* Obtenido de http://www.miislita.com

Goodwin, T., & Demner-Fushman, D. (2020). Overview of the 2020 Epidemic Question Answering Track. *Text Analysis Conference (TAC).*

Hernández, M., & Gómez, J. (2013). Aplicaciones de Procesamiento de Lenguaje Natural. *Revista Politécnica, 32*.

Iz Beltagy, K. L. (2019). SciBERT: A Pretrained Languaje Model for Scientific Text. *arXiv:1903.10676*.

Jacob Devlin, M.-W. C. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805*.

Jinhyuk Lee, W. Y. (2019). BioBERT: a pre-trained biomedical language. *Bioinformatics, 2019, 1–7*.

Json. (2021). Obtenido de https://docs.python.org/3/library/json.html

Jurafsky, D., & Martin, J. (2020). Speech and Language Processing.

Kevin Clark, M.-T. L. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *ICLR*.

Marshal. (2021). Obtenido de https://docs.python.org/es/3/library/marshal.html.

NQA. (2020). Obtenido de Neural Question Answering for CORD19 (task8): https://www.kaggle.com/aotegi/neural-question-answering-for-cord19-task8

ORDC. (2020). Obtenido de COVID-19 Open Research Dataset Challenge (CORD-19): https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

Pandas. (2021). Obtenido de https://pandas.pydata.org/

QuAC. (2021). Obtenido de https://quac.ai/

SQuAD. (2021). Obtenido de https://rajpurkar.github.io/SQuAD-explorer/

Tensorflow. (2021). Obtenido de https://www.tensorflow.org/?hl=es-419

Torch. (2021). Obtenido de https://pytorch.org/

Transformers. (2021). Obtenido de https://huggingface.co/transformers/

Vásquez, A., Huerta, H., Quispe, J., & Huayna, A. (2009). Procesamiento de lenguaje natural. *Revista de investigación de Sistemas e Informática*.

Vaswan, A. (2017). Attention Is All You Need.

Webbrowser. (2021). Obtenido de https://docs.python.org/es/3/library/webbrowser.html

Whoosh. (2021). Obtenido de https://whoosh.readthedocs.io/en/latest/intro.html

Xhlulu. (2020). Obtenido de https://github.com/xhlulu/covid-qa

Yinhan Liu, M. O. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692*.

Zhang, Y. &. (s.f.). BERT for Question Answering on SQuAD 2.0.