Respuesta a preguntas

Cálamo Educación S.L.
María Teresa Tijeras Pascual
Cálamo Educación S.L.
María Teresa Tijeras Pascual

Indice

Respuesta a preguntas

El objetivo de esta unidad es presentar los conceptos, fundamentos y técnicas más importantes de la tarea de respuesta a preguntas.

0. Objetivos de la unidad

El objetivo de esta unidad es presentar los conceptos, fundamentos y técnicas más importantes de la tarea de respuesta a preguntas. Esta tarea está relacionada con la recuperación de información y con parte de las tareas de procesamiento del lenguaje natural dedicadas a la extracción de información. Los sistemas de respuesta a preguntas son una vuelta de tuerca más a la recuperación de información, ya que, en vez de devolver la lista de documentos que contienen la información solicitada, directamente devuelven la respuesta a la consulta realizada.

El **apartado 1** presenta una introducción a los conceptos y la historia de esta tarea.

El **apartado 2** se centra en describir el proceso típico de respuesta a preguntas, en tres fases: análisis de la pregunta, recuperación de los contextos de información y extracción de la respuesta.

En el **apartado 3** se describe la tarea específica de respuesta a preguntas denominada *extractive question answering*.

Por último, el **apartado 4** presenta las tareas habituales del lingüista en este tipo de proyectos.

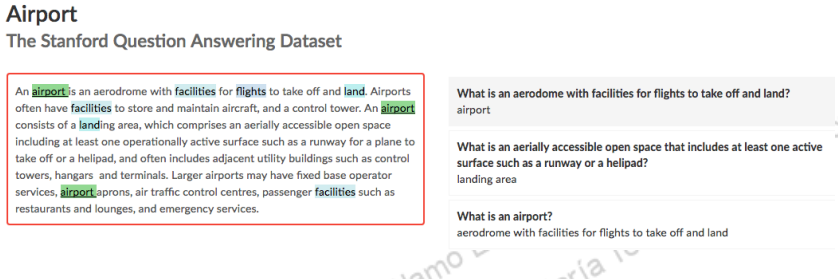
1. Introducción

1.1. Conceptos y fundamentos

La respuesta a preguntas (del inglés *question answering*, QA) es un área relacionada con la recuperación de información y el procesamiento del lenguaje natural que se ocupa de construir sistemas que respondan automáticamente a las preguntas planteadas por los seres humanos en lenguaje natural.

A diferencia de un sistema de recuperación de información, que devolvería la lista de documentos donde estaría la respuesta a la consulta realizada, un sistema de respuesta a preguntas devuelve directamente la respuesta solicitada.

Aunque un sistema de respuesta a preguntas podría construir sus respuestas consultando una base de datos estructurada con los conocimientos del sistema, normalmente la tarea consiste en encontrar las respuestas directamente de una colección no estructurada de documentos en lenguaje natural (textos).



Ejemplo de respuesta a preguntas en el corpus SQuAD de Stanford

Los sistemas utilizan una combinación de técnicas de lingüística computacional, recuperación de información y representación del conocimiento para encontrar respuestas.

1.2. Historia

Dos de los primeros sistemas de respuesta a preguntas en los años sesenta fueron BASEBALL, capaz de responder a preguntas sobre las Grandes Ligas de béisbol de EE. UU., y LUNAR, que respondía a preguntas sobre el análisis geológico de las rocas devueltas por las misiones lunares Apolo.

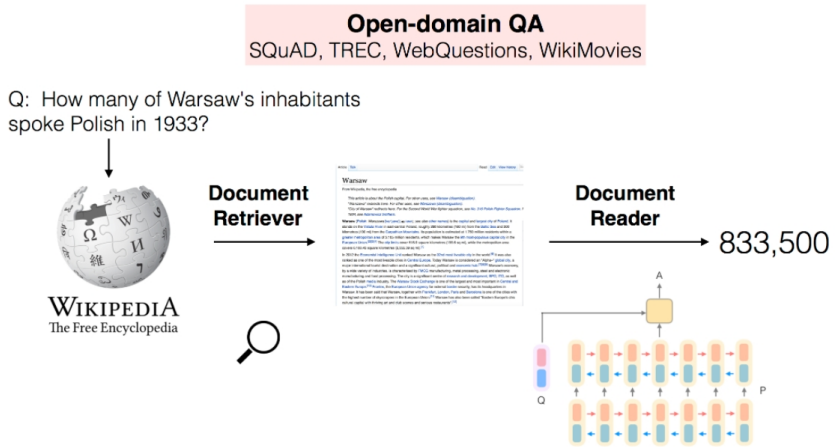
En los años siguientes se desarrollaron otros sistemas de respuesta a preguntas en dominios específicos que demostraron ser también muy eficaces. La característica común de todos estos sistemas era contar con una base de datos o sistema de conocimiento creado manualmente por expertos del dominio elegido.

En los años setenta, se desarrollaron sistemas expertos de respuesta a preguntas que utilizaban bases de conocimiento enfocadas a dominios específicos. Estos sistemas se parecían mucho a los sistemas modernos de respuesta a preguntas, excepto en su arquitectura interna. Los sistemas expertos se basan en gran medida en bases de conocimiento construidas y organizadas por expertos, mientras que muchos sistemas modernos de respuesta a preguntas se basan en el procesamiento estadístico de un gran corpus de texto en lenguaje natural no estructurado.

En los años setenta y ochenta se desarrollaron nuevos enfoques basados realmente en lingüística computacional que condujeron al desarrollo de proyectos de comprensión de textos y respuesta a preguntas muy ambiciosos. Un ejemplo de este tipo de sistema fue el Consultor Unix (*Unix Consultant*, UC), desarrollado en la Universidad de Berkeley a finales de los ochenta, que respondía a preguntas relacionadas con el sistema operativo Unix a partir de una amplia base de conocimientos elaborada manualmente sobre ese sistema operativo.

En 2011, Watson, un sistema informático de respuesta a preguntas desarrollado por IBM, compitió en el concurso de preguntas y respuestas “Jeopardy!” en EE. UU., ganando por un amplio margen.

Actualmente, los enfoques basados en *deep learning* han supuesto, como en otras tareas del procesamiento del lenguaje natural, un gran impulso en este tipo de sistemas. Por ejemplo, en 2017 Facebook Research liberó su sistema DrQA, capaz de responder a preguntas de dominio abierto utilizando Wikipedia como fuente de conocimiento, basado en aprendizaje profundo.



DrQA
Fuente de la imagen: GitHub.com

2. Proceso de respuesta a preguntas

2.1. Fundamentos

Típicamente, dada una pregunta del usuario, un sistema de respuesta a preguntas busca primero en su base de conocimientos (en su colección de textos, por ejemplo, todos los artículos de Wikipedia) para encontrar aquellos textos o fragmentos concretos donde puede estar contenida la respuesta, y luego procesa y analiza dichos textos para detectar y extraer la respuesta concreta a la consulta realizada.

Así, por ejemplo, para la consulta:

¿Cuál es la capital de Eslovenia?

El sistema encontraría, entre otros, la página: https://es.wikipedia.org/wiki/Eslovenia

En concreto, en esa página encontraría el siguiente fragmento (párrafo), que contiene las palabras “capital” y “Eslovenia”:

Eslovenia, oficialmente República de Eslovenia (en esloveno: Republika Slovenija, antigua Carantania) es uno de los veintisiete estados soberanos que forman la Unión Europea. Limita con Italia al oeste; con el mar Adriático, al suroeste; con Croacia al sur y al este; con Hungría, al noreste; y con Austria, al norte. Tiene una población de 2 080 908 habitantes a fecha del 1 de enero de 2019. La capital y ciudad más poblada es Liubliana.

Por último, procesando este fragmento, encontraría que la frase más relevante es:

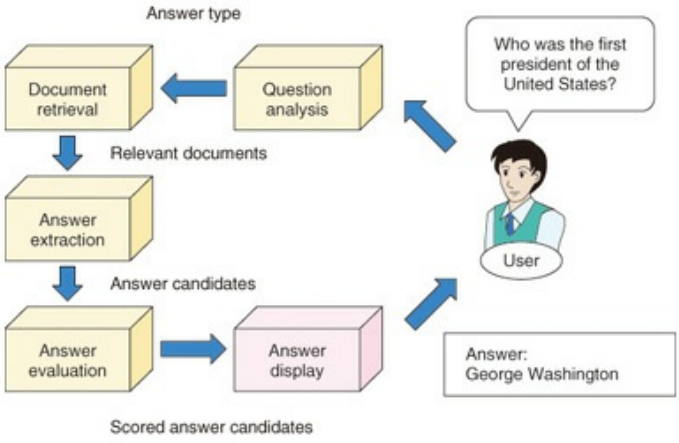
La capital y ciudad más poblada es Liubliana.

Finalmente, el sistema extraería la respuesta, utilizando la técnica en que esté basado internamente su motor (por ejemplo, con un enfoque basado en reglas), a las preguntas de tipo “cuál” (*what* o *which*) que se encuentran en estructuras sintácticas:

{question} VERBO_SER {answer}

Así, la respuesta sería “Liubliana”.

El proceso detallado del sistema se describe, a continuación, en los siguientes apartados.



Proceso de respuesta a preguntas
Fuente de la imagen: Ntt-review.jp

Análisis de la pregunta
El sistema determina el tipo de pregunta que está haciendo el usuario.
Recuperación de contextos
El sistema busca un conjunto de documentos con las palabras clave adecuadas que contengan la respuesta pretendida.
Extracción de la respuesta
El sistema procesa y analiza dichos documentos para detectar y extraer la respuesta concreta a la consulta realizada.

2.2. Análisis de la pregunta

El sistema toma como entrada una pregunta en lenguaje natural, por ejemplo, “¿Qué día nació Pablo Picasso?”. A continuación, la frase se transforma en una consulta a través de su forma lógica, para interpretar el tipo de pregunta.

Habitualmente los sistemas de respuesta a preguntas suelen incluir un módulo clasificador de preguntas que determina el tipo de pregunta y el tipo de respuesta que se espera recibir. Determinar el tipo de pregunta es una tarea crucial, ya que todo el proceso de extracción de respuestas depende de encontrar el tipo de pregunta correcto y, por tanto, el tipo de respuesta esperado.

Para identificar el tipo de pregunta, el primer paso es una extracción de palabras clave. En algunos casos, hay pistas claras que indican directamente el tipo de pregunta, como, por ejemplo: "Quién", "Dónde", “Cuándo” o "Cuántos", que indican al sistema que las respuestas deben ser del tipo "Persona", "Lugar", “Fecha” o "Número", respectivamente. En otros casos no hay marcas claras del tipo de pregunta (por ejemplo: “Cuál”, “Qué”, “Cómo”), con lo que hay que recurrir a otras palabras de la pregunta como anclaje. A veces se utiliza para ello un diccionario que incluye la serie de palabras que ayudan a determinar el contexto de la pregunta.

Otras técnicas que se pueden emplear para determinar el tipo de pregunta son el análisis morfológico *(part-of-speech)* y sintáctico, por ejemplo, considerando el sujeto, el verbo principal y los complementos de la frase de la pregunta.

Típicamente, los sistemas de respuesta a preguntas han abordado preguntas de definición y terminología, preguntas sobre hechos concretos (fechas, autores, lugares), listas de elementos, preguntas de explicación tipo "Cómo" y "Por qué", etc.

De forma más especulativa, también se ha investigado en preguntas multilingües (donde el idioma de la pregunta y el de la base de conocimientos no es el mismo), y preguntas sobre el contenido de elementos de audio, imágenes o vídeos (por ejemplo, "¿Qué político aparece saludando en la foto?").

2.3. Recuperación de contextos

Una vez identificado el tipo de pregunta, se utiliza un sistema de recuperación de información para encontrar un conjunto de documentos que contengan las palabras clave adecuadas.

Se denomina contexto o fragmento candidato a los documentos o partes de ellos (párrafos, frases e incluso fragmentos de frases) que el sistema ha identificado como relevantes para la pregunta, es decir, en los que confía que contengan la información buscada.

Es evidente que la respuesta a preguntas depende en gran medida de un buen corpus de búsqueda, ya que sin documentos que contengan la respuesta, poco puede hacer un sistema de respuesta a preguntas. Por tanto, es lógico que una colección de mayor tamaño mejore el rendimiento de la respuesta a las preguntas. Además, el hecho de haber redundancia en los datos (la misma información expresada de maneras diferentes en distintos contextos y documentos) permite aumentar la confianza del sistema en la respuesta si aparece varias veces, mejorando la robustez del sistema.

2.4. Extracción de la respuesta

El último paso consiste en procesar y analizar dichos contextos para detectar y extraer la respuesta concreta a la consulta realizada.

En algunos sistemas se utilizan técnicas de procesamiento del lenguaje natural para segmentar los textos, analizarlos morfosintácticamente y determinar si las palabras clave se ubican en los lugares pertinentes del árbol sintáctico (en el sujeto, en el objeto directo, etc.).

En algunos casos, como en las preguntas de tipo "Quién" o "Dónde", se puede utilizar también un sistema de reconocimiento de entidades con nombre para encontrar los nombres de "Persona" y "Lugar" pertinentes en los documentos recuperados. Solo se seleccionan los párrafos relevantes para la clasificación.

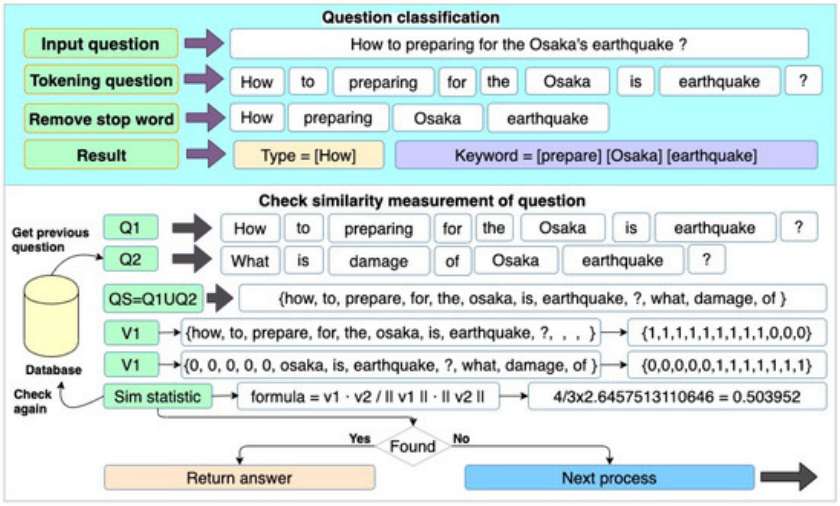
Otra técnica habitual es emplear un modelo de espacio vectorial para clasificar las respuestas candidatas. Tras comprobar si la respuesta es del tipo correcto determinado en la etapa de análisis del tipo de pregunta, se asigna una puntuación a cada una de las respuestas candidatas según la distancia entre su representación como vector y el vector de la consulta. Cuando más similares sean, previsiblemente más relacionadas estarán entre sí.

Algunos sistemas devuelven las frases que contienen la respuesta. Otros sistemas intentan extraer la respuesta concreta de la frase mediante reglas heurísticas y/o análisis sintáctico.

Hoy por hoy, estas tecnologías aún no permiten dar respuesta satisfactoria a preguntas complejas donde se incorpore un razonamiento elaborado, estilo "¿Cómo se llama el rey de España anterior a Felipe II?".

2.5. Ejemplo

La siguiente figura ilustra mediante un ejemplo el proceso completo de respuesta a preguntas:



Respuesta a preguntas
Fuente de la imagen: Mdpi.com

El usuario realiza una pregunta al sistema (la consulta es agramatical en inglés, por el uso del gerundio “preparing” en vez del infinitivo “prepare”):

> How to preparing for the Osaka’s earthquake?

El primer paso es el análisis de la pregunta.

El sistema procesa la consulta, segmentándola en *tokens* (obsérvese que el genitivo sajón “‘s” se ha segmentado incorrectamente como el verbo “to be”):

> How, to, preparing, for, the, Osaka, is, earthquake, ?

A continuación, se eliminan las palabras de parada:

> How, preparing, Osaka, earthquake

Por último, se lleva a cabo una clasificación del tipo de pregunta. La palabra interrogativa “How” indica que es una pregunta del tipo “How” (“Cómo”). Además, el sistema lematiza las palabras clave restantes, quedando:

> prepare, Osaka, earthquake

A continuación, un módulo de recuperación de información recupera todos los documentos que sean relevantes para la consulta con las palabras clave anteriores, por ejemplo, el documento:

What is the damage of Osaka earthquake?

Para cada uno de ellos, el sistema calcula la relevancia empleando la fórmula del coseno, o producto escalar del ángulo que forman sus vectores.

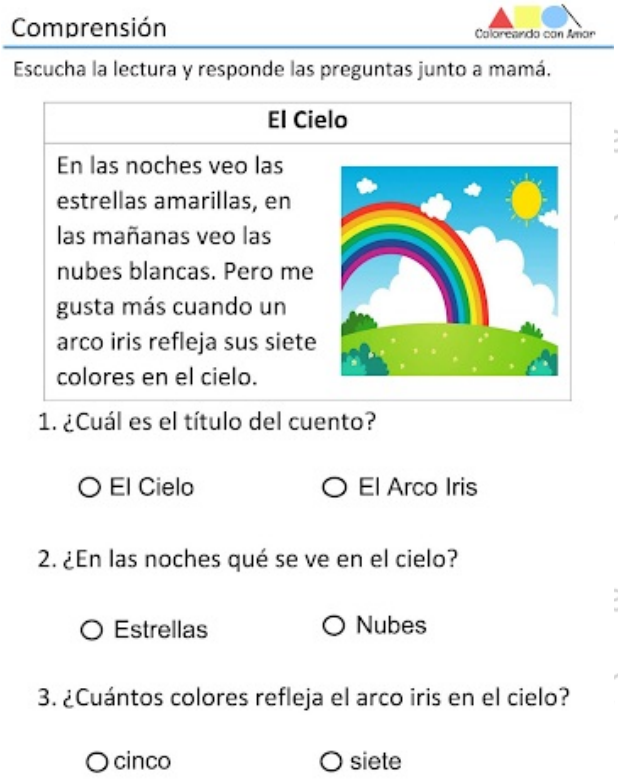
Si el tipo de pregunta de la consulta coincide con el documento (en este caso no, ya que es “what” y no “how”), se rechaza el documento y se comprueba con otro. En caso afirmativo, el sistema extrae la respuesta a partir de técnicas de extracción basadas en técnicas de procesamiento del lenguaje natural y la devuelve al usuario.

3. *Extractive question answering*

Se denomina *extractive question answering* (EQA, respuesta extractiva a preguntas) a la tarea de, dada una pregunta, extraer la respuesta de un texto.

Dado un texto con una cierta información, el objetivo del sistema es encontrar la respuesta concreta incluida en ese texto a una pregunta realizada por el usuario. Formalmente, la entrada es la tripla {texto, pregunta} y la salida esperada es la respuesta a dicha pregunta.

A veces también se denomina la tarea como comprensión lectora *(reading comprehension)*, ya que su objetivo es el mismo que el de la tarea del mismo nombre en el colegio. La figura siguiente muestra un típico ejercicio de comprensión lectora para niños de primaria, donde el objetivo es que los niños aprendan a interpretar el texto, demostrando que son capaces respondiendo a las preguntas formuladas, cuyas respuestas están contenidos en dicho texto.



Comprensión

Coloreando con Amor

Escucha la lectura y responde las preguntas junto a mamá.

El Cielo

En las noches veo las estrellas amarillas, en las mañanas veo las nubes blancas. Pero me gusta más cuando un arco iris refleja sus siete colores en el cielo.

1. ¿Cuál es el título del cuento?

○ El Cielo ○ El Arco Iris

2. ¿En las noches qué se ve en el cielo?

○ Estrellas ○ Nubes

3. ¿Cuántos colores refleja el arco iris en el cielo?

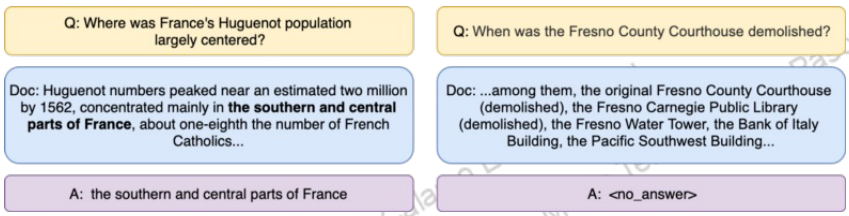
○ cinco ○ siete

Ejercicio de comprensión lectora para niños de 5 años
Fuente de la imagen: Materialeseducativosmaestras.com

Es decir, esta tarea se enfoca directamente a la última fase descrita en el proceso anterior, obviando el proceso de recuperación de contextos (que se podría realizar con un sistema externo, si fuera necesario).

Se trata de un problema bastante difícil para las máquinas. Por un lado, porque las máquinas tienen que aprender primero la estructura y el significado del lenguaje (a diferencia de un ser humano, que ya está familiarizado con la forma en que las palabras encajan en una frase y es capaz de "entender" el verdadero significado que hay detrás); una máquina necesita que se le enseñe de alguna manera. Y, por otro lado, porque puede que no resulte obvio dónde buscar en un texto la respuesta a una pregunta, ya que el lenguaje es muy flexible y la secuencia de palabras que se busca puede no aparecer literalmente (palabra por palabra) en el texto.

El corpus llamado **SQuAD** (Stanford Question Answering Dataset) se ha convertido en la referencia para entrenar estos sistemas, ya que contiene una colección de más de cien mil pares de preguntas y respuestas anotadas manualmente extraídas de artículos de Wikipedia.

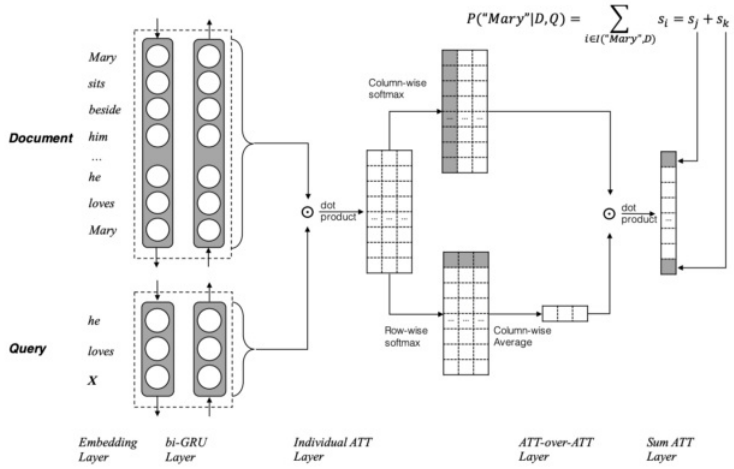


Ejemplos de SQuAD

Los enfoques más exitosos suelen utilizar redes neuronales profundas de extremo a extremo, es decir, desde el análisis del texto a la generación de la respuesta, ya que de esta forma captan una idea completa de las relaciones entre los contextos de las preguntas y las respuestas. Ningún modelo basado en técnicas de procesamiento del lenguaje natural, como el análisis sintáctico, el análisis de dependencias o los marcos semánticos, es capaz de alcanzar una precisión comparable.

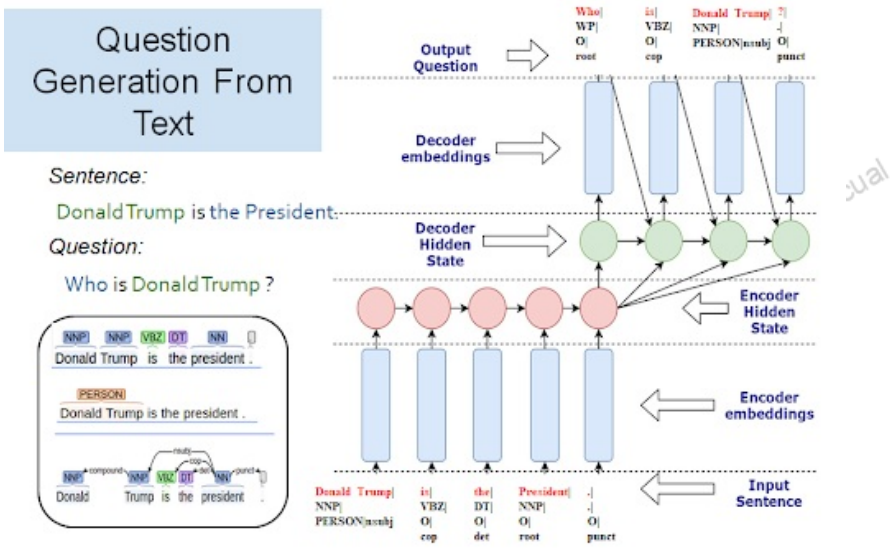
Como referencia, el rendimiento obtenido en la ejecución de la tarea por humanos es de un 82% de coincidencias exactas *(exact-match ratio)* y un 91% de Medida-F.

Así, la manera habitual de abordar esta tarea es hacerlo de forma similar a como se hace la traducción automática con *deep learning*, utilizando una arquitectura de red basada en el modelo *encoder-decoder* (o secuencia-a-secuencia), que convierte el texto de entrada y la consulta en la respuesta esperada. Otros algoritmos utilizados se basan en LSTM (Bi-LSTM, Match-LSTM) o GRU, con representación del significado mediante *embeddings* y capas mejoradas de atención.



Arquitectura para EQA (arquitectura "Attention over Attention")
Fuente de la imagen: Towardsdatascience.com

La tarea opuesta es la generación de preguntas a partir de texto. La siguiente figura muestra un ejemplo donde, a partir del texto "Donald Trump es el presidente de los Estados Unidos...", se puede generar la pregunta "¿Quién es Donald Trump?". El enfoque para abordar esa tarea es aplicar técnicas de *deep learning*, de nuevo, con un modelo estilo secuencia-a-secuencia, similar al que se aplica en traducción automática o la extracción de resúmenes, donde la secuencia de entrada es el texto y la secuencia de salida esperada es una pregunta que se podría responder con ese texto.



Generación de preguntas a partir de texto
Fuente de la imagen: Rnd.iitb.ac.in

4. Tarea del lingüista

El papel del lingüista en los proyectos relacionados con la respuesta a preguntas es parecido al caso de la recuperación de información, y típicamente consiste en:

1

La generación de conjuntos de evaluación del sistema, en este caso, listas de preguntas y las respuestas esperadas.

2

Colaborar en la definición del funcionamiento del sistema, por ejemplo, los algoritmos a emplear para procesar el texto, el desarrollo de reglas de extracción de la respuesta, el desarrollo del sistema de clasificación del tipo de pregunta (con técnicas, por ejemplo, basadas en clasificación con modelos de reglas), etc.

3

Llevar a cabo las evaluaciones del sistema utilizando las métricas definidas para el proyecto, analizar los resultados obtenidos y definir la estrategia de mejora del sistema.

RESUMEN

La tarea de **respuesta a preguntas** (del inglés *question answering*, QA) se ocupa de construir sistemas que respondan automáticamente a las preguntas planteadas por los seres humanos en lenguaje natural. A diferencia de un sistema de recuperación de información, que devolvería la lista de documentos donde estaría la respuesta a la consulta realizada, un sistema de respuesta a preguntas **devuelve directamente la respuesta solicitada**.

El proceso habitual de los sistemas de respuesta a preguntas es, dada una pregunta del usuario, **buscar primero en su colección de textos** para encontrar aquellos textos o fragmentos concretos donde puede estar contenida la respuesta, y luego **procesar y analizar dichos textos** para detectar y extraer la respuesta concreta a la consulta realizada, con diferentes técnicas basadas en lingüística computacional, recuperación de información y representación del conocimiento para encontrar respuestas.

La tarea de *extractive question answering* (respuesta extractiva a preguntas), también denominada **tarea de comprensión lectora** *(reading comprehension)*, se salta el paso de búsqueda de fragmentos y consiste en extraer la respuesta a una pregunta a partir de un determinado texto que se proporciona también como entrada.

Recursos

Enlaces de Interés

- **DrQA**
  https://github.com/facebookresearch/DrQA
- **Eslovenia**
  https://es.wikipedia.org/wiki/Eslovenia
- **Question Answering Technology**
  https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa4.html
- **Integrated Question-Answering System for Natural Disaster Domains Based on Social Media Messages Posted at the Time of Disaster**
  https://www.mdpi.com/2078-2489/11/9/456
- **Comprensión Lectora para 5 años - Materialeseducativosmaestras.com**
  http://www.materialeseducativosmaestras.com/2021/05/comprension-lectora-para-5-anos.html
- **Investigating the Machine Reading Comprehension Problem with Deep Learning**
  https://towardsdatascience.com/investigating-the-machine-reading-comprehension-problem-with-deep-learning-af850dbec4c0
- **Automating reading comprehension generating question and answer pair**
  https://rnd.iitb.ac.in/research-glimpse/automating-reading-comprehension-generating-question-and-answer-pair